BIOINFORMATICS ORIGINAL PAPER

Vol. 21 no. 2 2005, pages 144–151
doi:10.1093/bioinformatics/bth498

Filtering of Ineffective siRNAs and Improved siRNA Design Tool

S. M. Yiu[1,*], Prudence W. H. Wong[3], T. W. Lam[1], Y. C. Mui[1], H. F. Kung[2], Marie Lin[2] and Y. T. Cheung[2]

[1]Department of Computer Science and [2]Institute of Molecular Biology, University of Hong Kong, Hong Kong and [3]Department of Computer Science, University of Liverpool, Liverpool L69 7ZF, UK

Received on May 6, 2004; revised on August 7, 2004; accepted on August 19, 2004
Advance Access publication August 27, 2004

ABSTRACT

Motivation: Short interfering RNAs (siRNAs) can be used to suppress gene expression and possess many potential applications in therapy, but how to design an effective siRNA is still not clear. Based on the MPI (Max-Planck-Institute) basic principles, a number of siRNA design tools have been developed recently. The set of candidates reported by these tools is usually large and often contains ineffective siRNAs. In view of this, we initiate the study of filtering ineffective siRNAs.

Results: The contribution of this paper is 2-fold. First, we propose a fair scheme to compare existing design tools based on real data in the literature. Second, we attempt to improve the MPI principles and existing tools by an algorithm that can filter ineffective siRNAs. The algorithm is based on some new observations on the secondary structure, which we have verified by AI techniques (decision trees and support vector machines). We have tested our algorithm together with the MPI principles and the existing tools. The results show that our filtering algorithm is effective.

Availability: The siRNA design software tool can be found in the website http://www.cs.hku.hk/∼sirna/

Contact: smyiu@cs.hku.hk

INTRODUCTION

Short interfering RNAs (siRNAs), of length about 21, can be used to suppress gene expression (Fire *et al*., 1998; Elbashir *et al*., 2001a,b; Caplen *et al*., 2001) and possess many potential applications in therapy, for example, it is believed that siRNAs can be used to suppress the HIV-1 replication in human cell lines (Jacque *et al*., 2002). Different genes require different siRNAs to suppress the expression. An siRNA is, in fact, a DNA sequence that is formed by a substring of the mRNA of the target gene. However, not every substring of the target mRNA can form an effective siRNA (Holen *et al*., 2002). A typical mRNA can have a length of thousands. The number of potential candidates for siRNAs is therefore huge. To verify whether a given siRNA is effective, one must go through the laboratory experiments. These experiments are both time-consuming and expensive. Yet how to design an effective siRNA (i.e. to select the right substring from the mRNA for the construction of the siRNA) is still not clear.

As the first attempt to solve the problem, Tuschl *et al*. (2003, http://www.rockefeller.edu/labheads/tuschl/sirna.html) provided a set of guidelines, commonly known as the MPI (Max-Planck-Institute) principles, on how to design effective siRNAs. These principles try to capture some properties that an effective siRNA should possess, for example, the GC-content[1] of an siRNA should be between 30 and 70%. However, there are two issues in these principles. The properties given in the principles are not exclusive for effective siRNAs. In fact, among 19 ineffective siRNAs that have been reported in the literature, 6 of them also follow the MPI principles. Another issue is that the principles are rather primitive and not selective, the number of candidates that follow the principles is usually large. We have tested the MPI principles using 52 mRNAs with an average length of 2338. The average number of candidates reported is 327.

In the past three years, several siRNA design tools have been developed by refining and extending the MPI principles. However, in general, the set of candidates reported by most of these tools is still large (hundreds) and often contains some ineffective siRNAs (Table 1).

Our contributions The contribution of this paper is 2-fold.

(1) *A comparison scheme*: Despite the fact that quite a number of design tools have been developed, there is no study on comparing these tools. In fact, it is not trivial to compare these tools directly as the number of candidate siRNAs reported by these tools vary a lot. It is desirable to have a fair scheme to evaluate these tools. In this paper, we propose a fair scheme to compare these

*To whom correspondence should be addressed.

[1]GC-content is the percentage of the nucleotides G and C on the length-21 siRNA.

Table 1. Number of ineffective siRNAs filtered by our algorithm

Design tools	No. of relevant cases	No. of ineffective siRNAs reported before filtering	No. of ineffective siRNAs reported after filtering
Ambion_AA	4	3	1
OptiRNAi_AA	4	0	0
WI_AA (default)	4	0	0
Dharmacon_NA (default)	8	0	0
Emboss_NA	8	7	2
JackLin_NA	8	3	1
MPI principles	8	5	0
Dharmacon_NN	19	0	0
Qiagen_NN	19	2	0
WI_NN	19	12	6

The tools are grouped by the starting nucleotides of the siRNAs reported.

tools based on the published siRNAs. The idea is to make use of a random selector that will randomly pick the candidates from the target mRNA. The number of candidates to be chosen by the random selector depends on the output size of the tool in concern. Based on the published siRNAs, we calculate some indices showing how much the choice of the tool is better than a random choice. The random selector actually acts as a reference (control) for comparison. Our aim is to filter ineffective siRNAs, so the focus of our comparison is mainly on published ineffective siRNAs, the comparison for effective siRNAs is used as a reference.

We have evaluated seven existing tools and the MPI principles. The tools include Ambion (Ambion, 2003, http://www.ambion.com/techlib/misc/siRNA_finder.html), Dharmacon (Dharmacon, 2003, http://design.dharmacon.com/rnadesign/default.aspx?SID=691011983), Emboss (Williams, 2002, http://www.hgmp.mrc.ac.uk/Software/EMBOSS/Apps/sirna.html), Jack Lin (Lin, 2002, http://www.sinc.sunysb.edu/Stu/shilin/rnai.html), Whitehead Institute siRNA selection program (WI) (Yuan *et al*., 2004, http://jura.wi.mit.edu/siRNAext/), Qiagen (Qiagen, 2003, http://python.penguindreams. net/ Xeragon_Order_ Entry/jsp/Index.jsp), and OptiRNAi (Cui *et al*., 2003, http://bioit.dbi.udel.edu/rnai). The result shows that in general, most of these tools still output quite a number of ineffective siRNAs and have a similar (if not worse) performance as the random selector. For effective siRNAs, Jack Lin seems to be the best based on the published data.

(2) *A filtering algorithm:* Basically, most of these tools still try to identify a set of properties for selecting effective siRNAs. In this paper, we initiate the study of the properties that an ineffective siRNA would possess, which enables one to filter out the candidates that are unlikely to be an effective siRNA. We develop a filtering algorithm to improve the MPI principles and existing tools. The algorithm is based on some new observations on the secondary structure, which we have verified by AI techniques (decision trees and support vector machines). We have evaluated our filtering algorithm by applying it to the existing tools and the MPI principles. The results show that our filtering algorithm is effective. The number of ineffective siRNAs reported can be reduced by up to 100% while the number of effective siRNAs reported is only reduced by an average of 15%.

Remarks We have exploited about 100 siRNAs in our experiments. This already includes all ineffective and most effective siRNAs for human genes that are published in the literature.

Organization of the paper The rest of this paper is organized as follows. We first present the scheme for comparing existing siRNA design tools and the comparison result of seven existing tools. Then, we present the filtering algorithm for filtering ineffective siRNAs and discuss the experimental results of applying our filtering algorithm on the seven existing tools. This is followed by a discussion on how we find the filtering rule. Finally, we provide a summary and conclusion of our work.

THE COMPARISON SCHEME

Idea and results

In this section, we compare the performance of existing siRNA design tools and the MPI principles using real data in the literature. From the literature, there are 70 effective siRNAs and 19 ineffective siRNAs for human genes. (The references for the real data are provided in the Appendix section.) We compare the following tools: Ambion, Dharmacon, Emboss, Jack Lin, WI, Qiagen and OptiRNAi. Note that if a tool has options to restrict the selected siRNAs to have AA, NA or NN as the starting two nucleotides, we tried the default and the NN options (where 'N' stands for any nucleotide).

For a given mRNA, the number of candidate siRNAs reported by the tools can vary a lot. It is not trivial how one can compare these tools directly. We propose to use a random selector that randomly picks candidates from the target mRNA as a reference for comparison. To handle the issue of different output sizes for the tools, we make sure that the number of candidates to be selected by the random selector would be the same as the number of candidates reported by the tool in concern. In addition, if the tool only reports siRNAs starting with AA, the random selector will only select siRNAs starting with AA.

We then compare the two sets of candidates against the known siRNAs. For ineffective siRNAs, intuitively, if the tool reports less such siRNAs than the random selector, the choice of the tool is better than a random choice. We calculate the percentages of known ineffective siRNAs that have been reported by the tool and the random selector. The difference in these percentages, the net percentage, is used as an index to show

Table 2. The net percentages of various siRNA design tools against ineffective siRNAs

Design tools	Against ineffective siRNAs		
	Actual %	Expected %	Net %
Ambion_AA	75	70	5
OptiRNAi_AA	0	6	−6
WI_AA (default)	0	5	−5
Dharmacon_NA (default)	0	3	−3
Emboss_NA	88	66	22
JackLin_NA	38	25	13
MPI principles	63	32	31
Dharmacon_NN	0	3	−3
Qiagen_NN	11	3	8
WI_NN	63	60	3

Table 3. The net percentages of various siRNA design tools against effective siRNAs

Design tools	Against effective siRNAs		
	Actual %	Expected %	Net %
Ambion_AA	88	72	16
OptiRNAi_AA	48	9	39
WI_AA (default)	39	3	36
Dharmacon_NA (default)	5	3	3
Emboss_NA	93	71	22
JackLin_NA	62	20	42
MPI principles	86	52	34
Dharmacon_NN	7	5	2
Qiagen_NN	43	3	40
WI_NN	83	64	19

how much the choice of the tool is better than a random choice. Note that in calculating the percentages, if the tool only reports siRNAs starting with AA, we only consider the known siRNAs starting with AA. In fact, we do not actually run a random selector. We compute the expected percentage of ineffective siRNAs reported by the random selector. The detailed calculation will be discussed below. The net percentage is then defined as the actual percentage of ineffective siRNAs reported by the tool minus the expected percentage of the random selector. Obviously, a good siRNA tool should have a negative net percentage against ineffective siRNAs.

Table 2 shows the net percentages of various design tools against ineffective siRNAs. We see that many tools have positive net percentages; in other words, these tools report more ineffective siRNAs than the random selector. So, their choices of candidates are no better than the random choices with respect to ineffective siRNAs.

Computing the expected percentage for random selector

Now, we discuss the details of how to compute the expected percentage of the random selector. Consider a design tool T that reports siRNAs starting with AA. The other two cases for NA and NN are similar. Suppose M is the input mRNA. Let S_M be the set of ineffective siRNAs starting with AA that are reported in the literature and $\sigma_M = |S_M|$. Let n_M be the number of length-21 substrings of M that start with AA. Let k_M be the size of output of T for M. The random selector will select k_M siRNAs from the n_M candidates randomly. Let X_M denote the number of siRNAs reported by the random selector that are in S_M. Then the expected value of X_M can be computed as follows:

$$E(X_M) = \sum_{1 \le i \le \sigma_M} i \cdot \Pr(X_M = i) = \sum_{1 \le i \le \sigma_M} i \cdot \frac{\binom{\sigma_M}{i}\binom{n_M - \sigma_M}{k_M - i}}{\binom{n_M}{k_M}},$$

where $\binom{n}{r}$ denotes the number of combinations of choosing r items from n items. The expected number of ineffective siRNAs reported by the random selector equals to $\sum_M [E(X_M)]$, and the expected percentage equals to $\sum_M [E(X_M)]$ divided by the number of ineffective siRNAs in the literature that start with AA.

We have also performed the comparison of the tools against effective siRNAs. In this case, the actual percentage, the expected percentage and the net percentage are defined on known effective siRNAs. A good tool should have a positive net percentage. Table 3 shows the net percentages of various design tools against effective siRNAs. All the tools have positive net percentages, meaning that they report more effective siRNAs than the random selector. Their choices are better than random choices. In particular, Jack Lin seems to be the best based on the published data.

To conclude, the existing tools perform well in selecting the effective siRNAs but are not good for filtering out the ineffective ones. In the next section, we show how to enhance these tools by a filtering algorithm that filters potential ineffective siRNAs.

THE FILTERING ALGORITHM AND ITS PERFORMANCE

Performance of the filtering algorithm

Based on the discussion in the previous section, we see that both the MPI principles and most design tools report a certain number of ineffective siRNAs. In view of this, we attempt to improve the MPI principles and existing tools by an algorithm that can filter ineffective siRNAs. The target of the filtering algorithm is to reduce the number of ineffective siRNAs reported, and more importantly, reduce the net percentage against ineffective siRNAs.

We have applied the filtering algorithm on the output of the design tools to filter potential ineffective candidates. We observe that the output size is reduced by about ∼23% on

Table 4. Comparison of the net percentages against ineffective siRNAs before and after applying the filtering algorithm

Design tools	Net percentage against ineffective siRNAs		
	Before filtering	After filtering	Change
Ambion_AA	5	−33	−38
OptiRNAi_AA	−6	−5	+1
WI_AA (default)	−5	−4	+1
Dharmacon_NA (default)	−3	−2	+1
Emboss_NA	22	−26	−48
JackLin_NA	13	−9	−22
MPI principles	31	−22	−53
Dharmacon_NN	−3	−2	+1
Qiagen_NN	8	−2	−10
WI_NN	3	−14	−17

Table 5. Comparison of the net percentages against effective siRNAs before and after applying the filtering algorithm

Design tools	Net percentage against effective siRNAs		
	Before filtering	After filtering	Change
Ambion_AA	16	10	−6
OptiRNAi_AA	39	36	−3
WI_AA (default)	36	26	−10
Dharmacon_NA (default)	2	3	+1
Emboss_NA	22	19	−3
JackLin_NA	42	34	−8
MPI principles	34	30	−4
Dharmacon_NN	2	3	+1
Qiagen_NN	40	36	−4
WI_NN	19	16	−3

average. We have also shown in Table 1 that the number of ineffective siRNAs decreases by a significant amount. Regarding the net percentage against ineffective siRNAs, Table 4 shows that the percentages decrease drastically for most of the tools (up to 53% for the MPI principles). In particular, the net percentages of six of them become negative, implying that the corresponding tools now report fewer ineffective siRNAs than the random selector. This shows that our filtering algorithm is effective. Note that the expected percentage of the random selector is based on the reduced size of the output after filtering.

For the net percentage against effective siRNAs, Table 5 shows that the percentages decrease after applying the filtering but by a smaller amount; precisely, the net percentage decreases by at most 10%.

Details of the filtering algorithm

In this section, we give the details of the filtering algorithm. Note that in the process of suppressing gene expression, the siRNA needs to approach the corresponding target site on the mRNA. One of the factors affecting the success is the accessibility of the mRNA near the target site. This motivates us to study the secondary structure of the mRNA in concern, i.e. the pairing of the bases of the mRNA.

Repelling loops and big repelling loops Our filtering algorithm is based on the secondary structure properties that we call repelling loops and big repelling loops. The idea is as follows. The pairing of the bases may introduce loops [e.g. internal loop and multi-branched loop (Waterman, 2000)]. See Figure 1 for an example of a multi-branched loop. Suppose that a target site is hidden between two very close branches of a loop (because of the repelling force on the other side of the loop). Then it may not be easy for the corresponding siRNA to access the target site, so the siRNA has a high chance to be ineffective. We call such a loop a repelling loop with respect to that siRNA and the two branches are called the enclosing

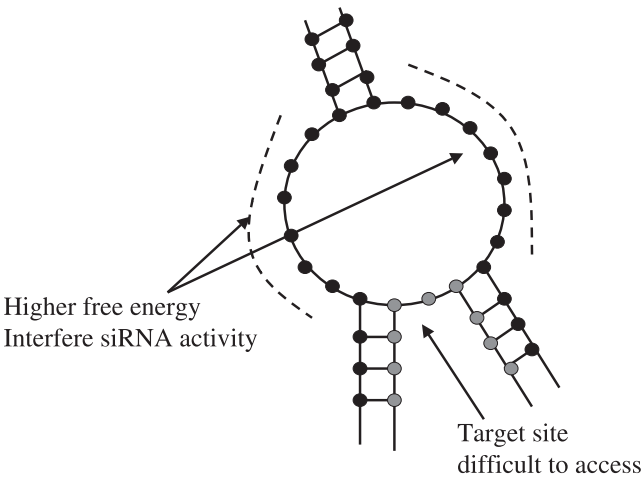

Fig. 1. A repelling loop of min_size-20 and max_degree-0.1.

branches. Figure 1 shows an example of a repelling loop and the corresponding target site. Furthermore, if the repelling loop is big, i.e. the number of unpaired bases is large, then the unpaired bases on the other sides of the loop have higher free energy, thus may interfere with the siRNA activity. We call such a loop a big repelling loop (with respect to that siRNA).

Now we give precise definitions for repelling loops and big repelling loops. For a given target site, consider the loops that are near to the target site and with at least two branches. A target site is near to a loop if it overlaps with the loop or is within a short distance from the loop, say 10 nt (we have tried other values and 10 seems to be a sensible choice and is used in all our experiments). Intuitively, for each loop, if the segment enclosed by the enclosing branches (with respect to the target site) is small relative to the total length of the loop, the target site is more difficult to be accessed. Therefore, we

measure the ratio between the length of the segment enclosed by the enclosing branches (with respect to the target site) and the total length of the loop. If this ratio is at most $r < 0.5$, we say that the loop is a repelling loop of max_degree-r. For example, the loop in Figure 1 is a repelling loop of max_degree-$(2/20)$.

Whether a repelling loop is considered to be big, we measure the length of the loop. If a repelling loop of max_degree-r has a length at least ℓ, we say that it is a repelling loop of min_size-ℓ and max_degree-r. Figure 1 shows a repelling loop of min_size-20 and max_degree-0.1. In particular, our filtering algorithm considers repelling loops of max_degree-0.25 and big repelling loops of min_size-15. The thresholds for repelling loops and big repelling loops are obtained by using AI techniques, which we will describe in the next section.

The filtering algorithm Based on the concept of repelling loops and big repelling loops, we have devised a filtering algorithm to filter potential ineffective siRNAs. The details are as follows. Consider any mRNA and a set of candidate siRNAs. We first obtain the secondary structures of the mRNA from Zuker's MFOLD algorithm (Zuker, 2003). The MFOLD algorithm usually reports over 10 secondary structures, each with a free energy indicates the stability of the corresponding structure. We focus on the five structures that have the lowest free energy, i.e. the five most stable structures. To determine whether to filter a candidate siRNA, we count, for each such structure, the number of repelling loops of max_degree-0.25 and the number of big repelling loops of min_size-15 and max_degree-0.25 with respect to that candidate. The filtering condition is as follows:

> A candidate is filtered if out of the five most stable secondary structures, (1) three or more structures each contain at least one big repelling loop of min_size-15 and max_degree-0.25 and (2) three or more structures [can be the same set of structures in (1)] each contain at least two repelling loops of max_degree-0.25 with respect to the candidate.

The filtering algorithm checks the filtering condition for each candidate siRNA and report those that are not filtered.

FINDING THE FILTERING RULE BY AI TECHNIQUES

In this section, we discuss how the filtering rule is derived using decision tree learning (Quinlan, 1987, http://www2.cs.uregina.ca/~hamilton/courses/831/notes/ml/dtrees/c4.5/tutorial.html). In addition to the repelling loops, big repelling loops we mentioned in the previous section, we have also considered the following two factors that are related to our observations.

- *The number of branches in a repelling loop*: Intuitively, if the number of branches increases, branches will be closer together, so if a target site is enclosed by the branches, it may be difficult to access it.
- *The free bases in the target site*: Free base has higher free energy and may interfere siRNA activity. Hence, we also consider the number of free bases in the target site. To reflect the relative strength of a CG-bond and an AT-bond, we assign a weight of 2 to a free A or T base and a weight of 3 to a free C or G base. The total weight of the free bases will be used in the decision tree training.

Training process For repelling loops and big repelling loops, there are two parameters, r (the repelling loop threshold) and ℓ (the big loop threshold), to consider. We repeatedly train the decision tree by fixing the values of r in the range 0.05–0.45, incremented by 0.05 each time. For a particular value of r, we compute the following attributes for each siRNA. Recall that when considering the secondary structures, we use the five most stable structures reported by the MFOLD algorithm.

(1) The largest number α such that at least 3 out of the 5 most stable secondary structures each contain at least α repelling loops of max_degree-r. Among these structures, select three that have more repelling loops and we break the ties by selecting the one with the lower free energy (i.e. the more stable one).

(2) For each integer ℓ in the range [13, 17], the largest number β_ℓ such that at least three out of the five most stable secondary structures each contain at least β_ℓ big repelling loops of min_size-ℓ and max_degree-r. Note that we have also tried other values of ℓ, the threshold 15 is clearly much better than the rest, thus we limit our experiments to examine the values from 13 to 17.

(3) The average number of branches of all the repelling loops in the three structures selected in Step 1. Note that the two branches that enclose the target site are not counted.

(4) The average of the total weight of free bases in the three structures selected in Step 1.

We then train the decision trees by including different attributes as follows. We have designed six sets of experiments: the first five sets correspond to one of the values of the big loop threshold β_ℓ, and the last set correspond to all the big loop thresholds. For each experiment, we include the attributes α and the corresponding β('s), while the remaining two attributes may or may not be included. As a result, we have four combinations: including both the unpaired weight and the number of branches, including either one of them, and including none of them.

We train the decision tree using a subset of data from the literature, which contains 27 effective and 6 ineffective siRNAs

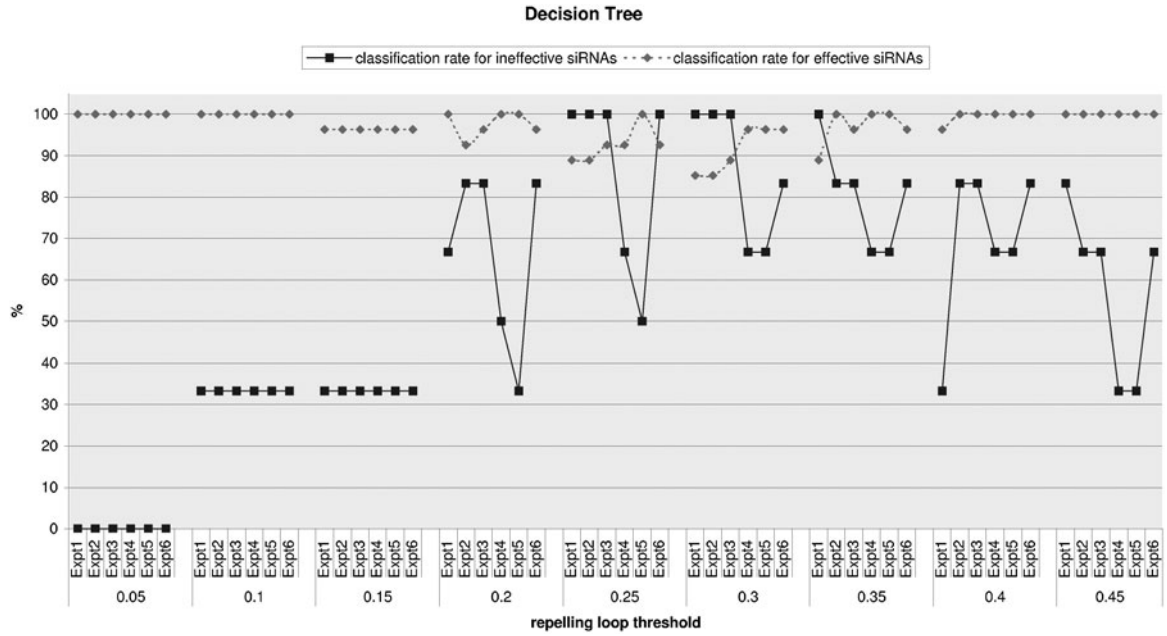

Fig. 2. The classification rates of the decision trees in various experiments.

that satisfy the following basic principles: (i) the position of the target site is at least 100 from the start codon of the corresponding mRNA, (ii) the GC-content of the siRNA is between 30 and 70% and (iii) the siRNA does not contain consecutive long runs of four or more equal nucleotides, e.g. GGGG.

Results Four decision trees are returned for each experiment. For each decision tree, we compute the rate of correct classification for ineffective and effective siRNAs. Since our target is to filter ineffective siRNAs, the ideal decision tree is one having 100% correct classification for ineffective siRNAs and a high correct classification rate for effective siRNAs. Figure 2 shows the classification rate of the decision trees for all the experiments. In each experiment, we only report the best decision tree, i.e. the one with the highest classification rate for ineffective siRNAs.

The results show that the best decision tree is the one for Experiments 3 (using 15 as the big loop threshold) and 6 (using all values in [13, 17] as the big loop thresholds) with repelling loop threshold $r = 0.25$. The classification rate for ineffective and effective siRNAs is 100 and 92.6%, respectively. In fact, when the repelling loop threshold is 0.25, all the eight decision trees for Experiments 3 and 6 are the same. This shows that 15 should be the appropriate big loop threshold and also even if we include the unpaired weight or the number of branches, the best decision tree only involves the attributes of the number of repelling loops of max_degree-0.25 and the number of big repelling loops of min_size-15 and max_degree-0.25.

We also make use of the support vector machine (Joachims, 1999) to see if the classification is consistent with that of the decision tree. We train the support vector machine learning module using the attributes α and β as in Experiments 3 and 6. The support vector machine obtained has a similar performance as the decision tree we obtained. Precisely, the classification rate for both ineffective and effective siRNAs are the same as that of the decision tree, and even further, the sets of siRNAs that are classified as ineffective and effective are the same for both the decision tree and the support vector machine. These results show that the attributes and the thresholds are selected appropriately.

As a remark, we have tried to use the rules to select siRNAs directly and compared its performance with the random selector. The results show that the rules perform better than the random selector. The output based on the rules show a net percentage of 23% for effective siRNAs and a net percentage of -15% for ineffective siRNAs.

CONCLUSION

In this paper, we have proposed a scheme to evaluate existing siRNA design tools based on the published effective and ineffective siRNAs. In the scheme, the output of each design tool is compared with a set of randomly selected siRNA candidates. The results show that existing tools are not good at filtering ineffective siRNAs. We also propose a filtering algorithm to filter potential ineffective siRNA candidates from the output of existing tools. The algorithm is based on two observations, namely repelling loops and big repelling loops, on secondary structures of the target mRNA. The rule for classifying potential ineffective siRNAs from other candidates is generated with the help of AI techniques, in particular, the decision tree and support vector machine. The filtering algorithm is shown to be effective.

The results of this paper provide evidence that the secondary structures should be considered for the design of siRNA. We are in the process of designing laboratory experiments to further verify our observations on secondary structures.

REFERENCES

Ambion (2003) Ambion siRNA Target Finder.

Caplen,N.J., Parrish,S., Imani,F., Fire,A. and Morgan,R.A. (2001) Specific inhibition of gene expression by small double-stranded RNAs in invertebrate and vertebrate systems. *Proc. Natl Acad. Sci. USA*, **98**, 9742–9747.

Cui,W., Ning,J., Naik,U.P. and Ducan,M.K. (2003) OptiRNAi, a web-based program to select siRNA sequences. In *Proceedings of the Computational Systems Bioinformatics Conference*. Stanford, CA, pp. 433–434.

Dharmacon (2003) Dharmacon siDESIGN Center.

Elbashir,S.M., Harborth,J., Lendeckel,W., Yalcin,A., Weber,K. and Tuschl,T. (2001a) Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. *Nature*, **411**, 494–498.

Elbashir,S.M., Lendeckel,W. and Tuschl,T. (2001b) RNA interference is mediated by 21- and 22-nucleotide RNAs. *Genes Dev.*, **15**, 188–200.

Fire,A., Xu,S., Montgomery,M.K., Kostas,S.A., Driver,S.E. and Mello,C.C. (1998) Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature*, **391**, 806–811.

Holen,T., Amarzguioui,M., Wiiger,M.T., Babaie,E. and Prydz,H. (2002) Positional effects of short interfering RNAs targeting the human coagulation trigger Tissue Factor. *Nucleic Acids Res.*, **30**, 1757–1766.

Jacque,J.-M., Triques,K. and Stevenson,M. (2002) Modulation of HIV-1 replication by RNA interference. *Nature*, **418**, 435–438.

Joachims,T. (1999) Making large-scale SVM learning practical. In Scholkopf,B., Burges,C. and Smola,A. (eds), *Advances in Kernel Methods–Support Vector Learning*. MIT Press.

Lin,J. (2002) Jack Lin's siRNA Sequence Finder.

Qiagen (2003) The siRNA design tool Qiagen.

Quinlan,J. (1987) Decision Tree C4.5.

Tuschl,T., Elbashir,S., Harborth,J. and Weber,K. (2003) The siRNA user guide.

Waterman,M.S. (2000) *Introduction to Computational Biology—Maps, sequences and genomes*. Chapman & Hall/CRC.

Williams,G. (2002) The siRNA design tool EMBOSS.

Yuan,B., Latek,R., Hossbach,M., Tuschl,T. and Lewitter,F. (2004) siRNA Selection Server: an automated siRNA oligonucleotide prediction server. *Nucleic Acids Res.*, **32**, W130–W134.

Zuker,M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, **31**, 3406–3415.

APPENDIX

References for the published data

Ancellin,N., Colmont,C., Su,J., Li,Q., Mittereder,N., Chae,S.-S., Stefansson,S., Liau,G. and Hla,T. (2002) Extracellular export of sphingosine kinase-1 enzyme. Spingosine 1-phosphate generation and the induction of angiogenic vascular maturation. *J. Biol. Chem.*, **277**, 6667–6675.

Bai,X., Zhou,D., Brown,J.R., Crawford,B.E., Hennet,T. and Esko,J.E. (2001) Biosynthesis of the linkage region of glycosaminoglycans: cloning and activity of galactosyltransferase II, the sixth member of the beta 1,3-galactosyltransferase family (beta 3GalT6). *J. Biol. Chem.*, **276**, 48189–48195.

Bakker,J., Lin,X. and Nelson,W.G. (2002) Methyl-CpG binding domain protein 2 represses transcription from hypermethylated pi-class glutathione S-transferase gene promoters in hepatocellular carcinoma cells. *J. Biol. Chem.*, **277**, 22573–22580.

Boehm,M., Yoshimoto,T., Crook,M.F., Nallamshetty,S., True,A., Nabel,G.J. and Nabel,E.G. (2002) A growth factor-dependent nuclear kinase phosphorylates p27(Kip1) and regulates cell cycle progression. *EMBO J.*, **21**, 3390–3401.

Chevrier,V., Piel,M., Collomb,N., Saoudi,Y., Frank,R., Paintrand,M., Narumiya,S., Bornens,M. and Job,D. (2002) The Rho-associated protein kinase p160ROCK is required for centrosome positioning. *J. Cell Biol.*, **157**, 807–817.

Cortez,D., Guntuku,S., Qin,J. and Elledge,S.J. (2001) ATR and ATRIP: partners in checkpoint signaling. *Science*, **294**, 1713–1716.

Edbauer,D., Winkler,E., Haass,C. and Steiner,H. (2002) Presenilin and nicastrin regulate each other and determine amyloid beta-peptide production via complex formation. *Proc. Natl Acad. Sci. USA*, **99**, 8666–8671.

Eide,E.J., Vielhaber,E.L., Hinz,W.A. and Virshup,D.M. (2002) The circadian regulatory proteins BMAL1 and cryptochromes are substrates of casein kinase I epsilon. *J. Biol. Chem.*, **277**, 17248–17254.

Harborth,J., Elbashir,S.M., Bechert,K., Tuschl,T. and Weber,K. (2001) Identification of essential genes in cultured mammalian cells using small interfering RNAs. *J. Cell Sci.*, **114**, 4557–4565.

Heinonen,J.E., Smith,C.I. and Nore,B.F. (2002) Silencing of Bruton's tyrosine kinase (Btk) using short interfering RNA duplexes (siRNA). *FEBS Lett.*, **527**, 274–278.

Hewitt,E.W., Duncan,L., Mufti,D., Baker,J., Stevenson,P.G. and Lehner,P.J. (2002) Ubiquitylation of MHC class I by the K3 viral protein signals internalization and TSG101-dependent degradation. *EMBO J.*, **21**, 2418–2429.

Hirai,I. and Wang,H.-G. (2002) A role of the C-terminal region of human Rad9 (hRad9) in nuclear transport of the hRad9 checkpoint complex. *J. Biol. Chem.*, **277**, 25722–25727.

Holen,T., Amarzguioui,M., Wiiger,M.T., Babaie,E. and Prydz,H. (2002) Positional effects of short interfering RNAs targeting the human coagulation trigger Tissue Factor. *Nucleic Acids Res.*, **30**, 1757–1766.

Jiang,X., Kim,H.-E., Shu,H., Zhao,Y., Zhang,H., Kofron,J., Donnelly,J., Burns,D., Ng,S.C., Rosenberg,S. and Wang,X. (2003) Distinctive roles of PHAP proteins and prothymosin-alpha–in a death regulatory pathway. *Science*, **299**, 223–226.

Kisielow,M., Kleiner,S., Nagasawa,M., Faisal,A. and Nagamine,Y. (2002) Isoform-specific knockdown and expression of adaptor protein ShcA using small interfering RNA. *Biochem. J.*, **363**, 1–5.

Koepp,D.M., Schaefer,L.K., Ye,X., Keyomarsi,K., Chu,C., Harper,J.W. and Elledge,S.J. (2001) Phosphorylation-dependent ubiquitination of cyclin E by the SCFFbw7 ubiquitin ligase. *Science*, **294**, 173–177.

Lassus,P., Opitz-Araya,X. and Lazebnik,Y. (2002) Requirement for caspase-2 in stress-induced apoptosis before mitochondrial permeabilization. *Science*, **297**, 1352–1354.

Li,L., Mao,J., Sun,L., Liu,W. and Wu,D. (2002) Second cysteine-rich domain of Dickkopf-2 activates canonical Wnt signaling pathway via LRP-6 independently of dishevelled. *J. Biol. Chem.*, **277**, 5977–5981.

Liu,J., Yao,F., Wu,R., Morgan,M., Thorburn,A., Finely,R.L.,Jr and Chen,Y.O. (2002) Mediation of the DCC apoptotic signal by DIP13 alpha. *J. Biol. Chem.*, **277**, 26281–26285.

Liu,X. and Erikson,R.L. (2002) Activation of Cdc2/cyclin B and inhibition of centrosome amplification in cells depleted of Plk1 by siRNA. *Proc. Natl Acad. Sci. USA*, **99**, 8672–8676.

Mailand,N., Lukas,C., Kaiser,B.K., Jackson,P.K., Bartek,J. and Lukas,J. (2002) Deregulated human Cdc14A phosphatase disrupts centrosome separation and chromosome segregation. *Nat. Cell Biol.*, **4**, 317–322.

Martin-Lluesma,S., Stucke,V.M. and Nigg,E.A. (2002) Role of Hec1 in spindle checkpoint signaling and kinetochore recruitment of Mad1/Mad2. *Science*, **297**, 2267–2270.

Martins,L.M., Iaccarino,I., Tenev,T., Gschmeissner,S., Totty,N.F., Lemoine,N.R., Savopoulos,J., Gray,C.W., Creasy,C.L., Dingwall,C. and Downward,J. (2002) The serine protease Omi/HtrA2 regulates apoptosis by binding XIAP through a reaper-like motif. *J. Biol. Chem.*, **277**, 439–444.

Prasanth,S.G., Prasanth,K.V. and Stillman,B. (2002) Orc6 involved in DNA replication, chromosome segregation, and cytokinesis. *Science*, **297**, 1026–1031.

Shang,Y. and Brown,M. (2002) Molecular determinants for the tissue specificity of SERMs. *Science*, **295**, 2465–2468.

Surka,M.C., Tsang,C.W. and Trimble,W.S. (2002) The mammalian septin MSF localizes with microtubules and is required for completion of cytokinesis. *Mol. Biol. Cell*, **13**, 3532–3545.

Tsuneoka,M., Koda,Y., Soejima,M., Teye,K. and Kimura,H. (2002) A novel myc target gene, mina53, that is involved in cell proliferation. *J. Biol. Chem.*, **277**, 35450–35459.

Zhang,D., Li,F., Weidner,D., Mnjoyan,Z.H. and Fujise,K. (2002) Physical and functional interaction between myeloid cell leukemia 1 protein (MCL1) and Fortilin. The potential role of MCL1 as a fortilin chaperone. *J. Biol. Chem.*, **277**, 37430–37438.

Zou,L. and Elledge,S.J. (2003) Sensing DNA damage through ATRIP recognition of RPA–ssDNA complexes. *Science*, **300**, 1542–1548.