FEBS Letters 579 (2005) 2878–2882 FEBS 29581

Prediction of siRNA functionality using generalized string kernel and support vector machine

Reiji Teramoto[1], Mikio Aoki[1], Toru Kimura, Masaharu Kanaoka*

Genomic Science Laboratories, Sumitomo Pharmaceuticals Co., Ltd. 3-1-98 Kasugade-Naka, Konohana-ku, Osaka 554-0022, Japan

Received 14 February 2005; revised 18 April 2005; accepted 19 April 2005

Available online 29 April 2005

Edited by Robert B. Russell

Abstract Small interfering RNAs (siRNAs) are becoming widely used for sequence-specific gene silencing in mammalian cells, but designing an effective siRNA is still a challenging task. In this study, we developed an algorithm for predicting siRNA functionality by using generalized string kernel (GSK) combined with support vector machine (SVM). With GSK, siRNA sequences were represented as vectors in a multi-dimensional feature space according to the numbers of subsequences in each siRNA, and subsequently classified with SVM into effective or ineffective siRNAs. We applied this algorithm to published siRNAs, and could classify effective and ineffective siRNAs with 90.6%, 86.2% accuracy, respectively.

Keywords: Small interfering RNA; RNA interference; String kernel; Support vector machine; Functional genomics; Leave-one-out cross-validation

1. Introduction

RNA interference (RNAi) is a process of double-stranded (ds) RNA-dependent, post-transcriptional gene silencing [1–4]. dsRNA introduced into cells is digested by Dicer to yield small interfering RNAs (siRNAs) of 21–23 nucleotides (nt) in length [5,6]. These siRNAs are incorporated into a multi-component nuclease complex, RNA-induced silencing complex (RISC), which is responsible for the destruction of cognate mRNAs [6,7].

siRNA-based method for silencing mammalian genes is thought to be more promising than that of the long dsRNA [8], because introduction of long dsRNA into mammalian cells frequently induces a fatal interferon response [9]. siRNA-based RNAi, however, is not readily usable for the mammalian gene silencing, since only a limited fraction of siRNAs are capable of performing highly effective RNAi in mammalian cells [10,11].

Tuschl and co-workers formulated empirical rules for designing functional siRNA, based on the experimental evidence obtained from the systematic screening of sequence dependence of siRNA functionality [12,13]. Briefly, these rules include: (1) siRNA duplexes should be composed of 21 nt sense and antisense strands, paired so as to have a 2 nt 3′ overhang at each end, (2) a target sequence should be selected from a region of the given mRNA sequence beginning 50–100 nt downstream of the start codon, and (3) the target sequence should be 23 nt composed of a motif AA(N19)TT or NA(N21) (N, any nucleotide) with approximately 50% G/C-content (30–70% G/C content also works in some cases). Although these empirical rules provide a basis for designing siRNAs, predicting the knockdown efficacy of siRNAs remains to be improved. Recently, Reynolds et al. [14] and Ui-Tei et al. [15] reported guidelines for rational siRNA design based on position-dependent characteristics associated with siRNA functionality.

In this study, we developed an algorithm for predicting siRNA functionality by using generalized string kernel (GSK) combined with support vector machine (SVM) [16] to extract sequence feature and to discriminate functionality, respectively. Application of the algorithm to published data sets demonstrated that the method could distinguish effective and ineffective siRNAs with high accuracy.

2. Materials and methods

2.1. Data sets

From Khvorova's large data sets containing sequence and function of siRNAs [17], a subset of 94 siRNAs targeting the firefly luciferase and human cyclophilin B genes, and belonging to two functional classes, effective and ineffective, were used in this study. Out of the 94 chosen siRNAs, the effective class contained 53 siRNAs with 90% or more gene silencing activity and the ineffective class contained 41 siRNAs with less than 50% gene silencing activity.

2.2. Feature map for siRNAs

GSK is based on mismatch string kernel (MSK) as well as on the spectrum kernel [18,19]. The (k,m)-mismatch string kernel ((k,m)-MSK) maps feature space generated by shared occurrences of fixed k-length subsequences differing by at most m mismatches [18]. For a sequence x of a given length, we define the feature vectors for all the k-mer as $\Phi_{(k,m)}(x)$. The (k,m)-MSK $K_{(k,m)}(x,y)$ is the inner product in feature space of feature vectors:

$$K_{(k,m)}(x,y) = \langle \Phi_{(k,m)}(x), \Phi_{(k,m)}(y) \rangle.$$

In the case of $m = 0$, as was used in this study, $K_{(k,0)}(x,y)$ is k-spectrum kernel. For normalization, we introduced $K_{(k,m)}(x,y)$ as:

$$K_{(k,m)}(x,y) \leftarrow \frac{K_{(k,m)}(x,y)}{\sqrt{K_{(k,m)}(x,x)}\sqrt{K_{(k,m)}(y,y)}}.$$

GSK is a sum of all the (k_i,m_i)-mismatch kernels. The $(k_1,m_1,\ldots,k_s,m_s)$-GSK $K_{(k_1,m_1,\ldots,k_s,m_s)}(x,y)$ is defined as:

*Corresponding author. Fax: +81 6 6466 5491.
E-mail address: kanaoka@sumitomopharm.co.jp (M. Kanaoka).

[1] These authors are equally contributed to this work.

Abbreviations: siRNA, small interfering RNA; GSK, generalized string kernel; SVM, support vector machine; LOOCV, leave-one-out cross-validation; nt, nucleotide

0014-5793/$30.00
doi:10.1016/j.febslet.2005.04.045

$$K_{(k_1,m_1,\ldots,k_s,m_s)}(x,y) = \sum_i \langle \Phi_{k_i,m_i}(y), \Phi_{k_i,m_i}(y) \rangle = \sum_i K_{(k_i,m_i)}(x,y).$$

GSK also satisfies Mercer's Theorem, because MSK is Mercer Kernel [20]. This means that GSK assures to afford the global optimal solution by SVM.

2.3. SVM implementation

The core algorithm of SVM in this study was derived from LIBSVM [16,21] (www.csie.ntu.edu.tw/~cjlin/libsvm/). In the SVM procedure, linear kernel and soft margin were included in the algorithm.

3. Results

3.1. Feature extraction from siRNA sequence

The basis of our approach is to describe siRNA sequences as vectors in a multi-dimensional feature space reflecting the numbers of 1, 2 and 3-mer subsequences in each siRNA. We then subjected the feature vectors representing training sequences to a supervised machine learning algorithm, SVM.

To extract the feature from siRNA sequence, we employed GSK for a test data set of siRNAs published by Khvorova et al. [17], representing 53 effective and 41 ineffective siRNAs (Fig. 1). With GSK of k-mer subsequences, where k-mer is 1-mer (1-GSK), 2-mer (2-GSK), or 3-mer only (3-GSK), or with GSK of all the 1–3-mer subsequences ((1,2,3)-GSK), we could classify the test data sets with 55.3%, 80.9%, 87.2%, and 86.2% accuracy, respectively (Table 1). These results indicated that discriminative performance was higher with 3-GSK, and (1,2,3)-GSK than with 1-GSK or 2-GSK. Table 2 shows a list of top 20 of the SVM weight vectors for (1,2,3)-GSK. The absolute value of the SVM weight vector for each subsequence

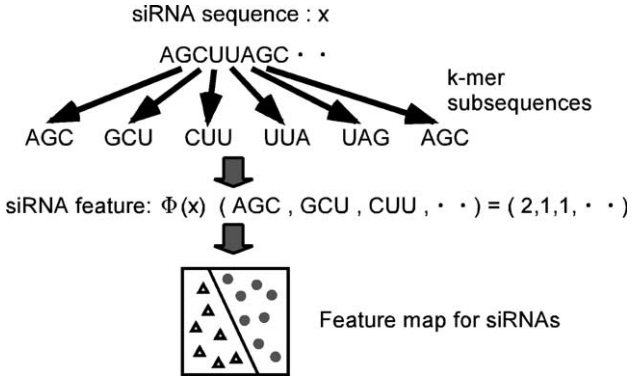

Fig. 1. Illustration of GSK. GSK maps feature space indexed by all possible subsequences of siRNAs of fixed length k. In this study, the feature of the siRNA sequence was extracted by counting the numbers of 1 to 3-mer subsequences of siRNAs.

Table 1
Comparison of discriminative performance among 1-, 2-, 3-, and (1,2,3)-GSK/SVM for the test data set representing 53 effective and 41 ineffective siRNAs

Kernel	TP	TN	FP	FN	Accuracy
1-GSK	37	15	26	16	55.3% (52/94)
2-GSK	44	32	9	9	80.9% (76/94)
3-GSK	49	33	8	4	87.2% (82/94)
1,2,3-GSK	48	33	8	5	86.2% (81/94)

TP: true positive, TN: true negative, FP: false positive, FN: false negative.

Table 2
Top 20 of the SVM weight vectors for (1,2,3)-GSK

Rank	Subsequence	Weight
1	CAC	0.599
2	GGA	0.374
3	AUA	0.368
4	UGC	0.338
5	CAA	0.334
6	AGC	0.317
7	CAU	0.301
8	GCC	0.300
9	UGA	0.283
10	UG	0.276
11	AAG	0.274
12	CUG	0.268
13	CUC	0.265
14	GAG	0.253
15	GA	0.240
16	GCA	0.231
17	GU	0.230
18	UCC	0.228
19	CCA	0.224
20	CUU	0.198

The weight is represented as absolute value along with the subsequence.

represents its importance on classification. Although 17 out of the top 20 SVM weight vectors were derived from 3-mer subsequences, the others (10th,15th, and 17th) were from 2-mer subsequences. The weight vectors derived from 1-mer subsequences C, A, G, and U were 0.087, 0.055, 0.030, and 0.027, respectively. These results indicated that the sequence feature derived from either 1-mer, or 2-mer still has considerable contribution to the discriminative performance. Therefore we used (1,2,3)-GSK for further analysis.

Fig. 2A shows distribution of the GSK/SVM scores for the 94 siRNAs. As shown, 90.6% of the effective and 80.5% of the ineffective siRNAs had positive and negative scores, respectively. In Fig. 2B, the graph shows the cumulative frequencies of the effective siRNAs arranged in order of the GSK/SVM scores against those of the ineffective siRNAs. All of the first 36 siRNAs and the last 24 siRNAs were classified as effective and ineffective, respectively.

Fig. 3 shows examples of the siRNA sequences along with the GSK/SVM scores. As shown, except only one case, GSK/SVM could distinguish the effective siRNAs from the ineffective ones despite they have overlapping nearly identical sequences. These results suggested that the feature of siRNA sequence extracted by the GSK could properly represent siRNA functionality.

3.2. Leave-one-out cross-validation (LOOCV) of the GSK/SVM algorithm

In order to validate the GSK/SVM algorithm for prediction of siRNA functionality, we performed a leave-one-out cross-validation (LOOCV). Fig. 4A shows distribution of LOOCV GSK/SVM scores for the 94 siRNAs. As shown, 75.5% of the effective and 68.3% of the ineffective siRNAs had positive and negative scores, respectively. The overall accuracy was 72.3% (=68/94), of which 40 were true positives, 28 true negatives, 13 false positives, and 13 false negatives. Fig. 4B shows the cumulative frequencies of the effective siRNAs arranged in order of the LOOCV GSK/SVM scores against those of the ineffective siRNAs. Among the first 10 siRNAs, 9 were effective, and 9 of the last 10 were ineffective. When these

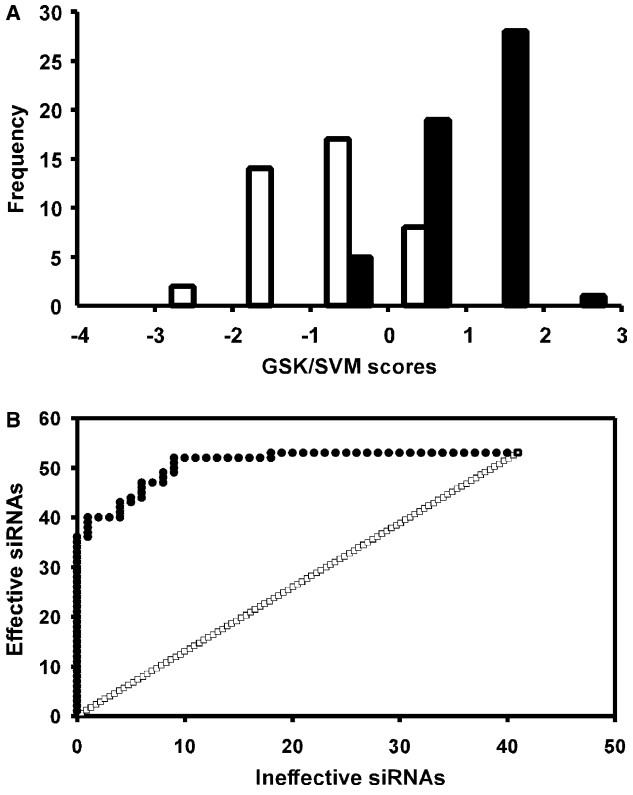

Fig. 2. Classification of the test data by GSK/SVM algorithm. (A) Distribution of GSK/SVM scores of the effective and ineffective siRNAs. Black bars and white bars show distribution of GSK/SVM scores for the effective and ineffective siRNAs, respectively. (B). Accumulation curve of the effective siRNAs arranged in order of GSK/SVM scores against those of the ineffective siRNAs. The graph plots the cumulative frequencies of the effective siRNAs (*y*-axis) arranged in order of GSK/SVM scores (closed circles) or by random selection (open squares) against those of the ineffective siRNAs (*x*-axis).

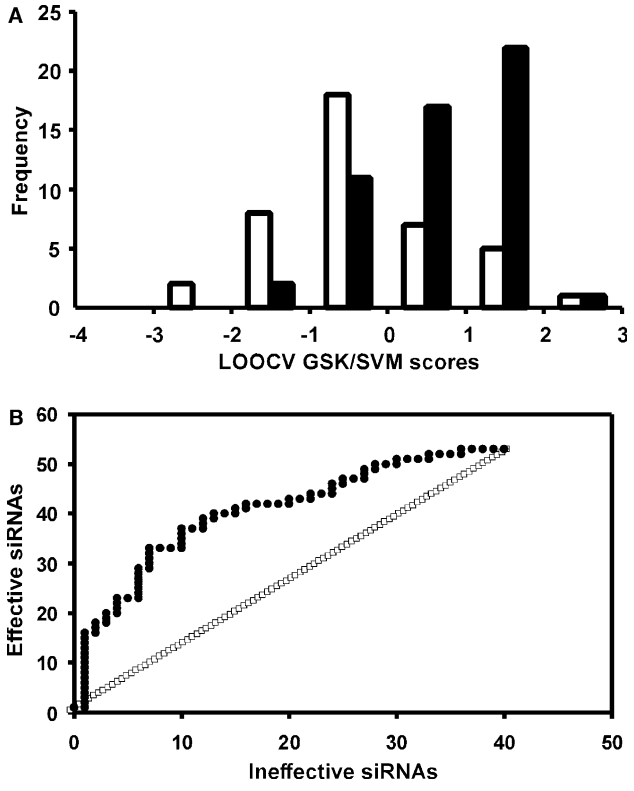

Fig. 4. LOOCV of the GSK/SVM algorithm. (A) Distribution of LOOCV GSK/SVM scores for the effective and ineffective siRNAs. Black bars and white bars show distribution of LOOCV GSK/SVM scores for the effective and ineffective siRNAs, respectively. (B) Accumulation curve of the effective siRNAs arranged in order of LOOCV GSK/SVM scores against those of the ineffective siRNAs. The graph plots the cumulative frequencies of the effective siRNAs arranged in order of LOOCV GSK/SVM scores (closed circles) or by random selection (open squares) against those of the ineffective siRNAs.

LOOCV GSK/SVM scores were plotted along with the model scores (Fig. 5), they showed a significant correlation with a correlation coefficient of 0.78.

Collectively, these results indicated that the GSK/SVM algorithm was effective in predicting siRNA functionality.

3.3. Validation of predictive performance of GSK/SVM algorithm against other genes

We evaluated predictive performance of GSK/SVM algorithm for 16 human SEAP gene siRNAs published by Khvorova et al. [17], and further compared its predictive performance with Reynolds' rational design algorithm [14] against 10 siRNAs for glyceraldehyde-3-phosphate dehydrogenase (GAPD) gene and 4 siRNAs for diazepam binding inhibitor (DBI) gene. Although our algorithm was trained with siRNAs that are clearly effective (>90%) or ineffective (<50%), a weak positive correlation between GSK/SVM score and gene silencing efficacy was observed in the whole range, and most of the siRNAs with positive GSK/SVM score practically exhibited 80% or more gene silencing, as shown in Fig. 6A and B. To compare between different platforms, 80% knockdown was used as a threshold to define effective siRNA accordingly. As for the

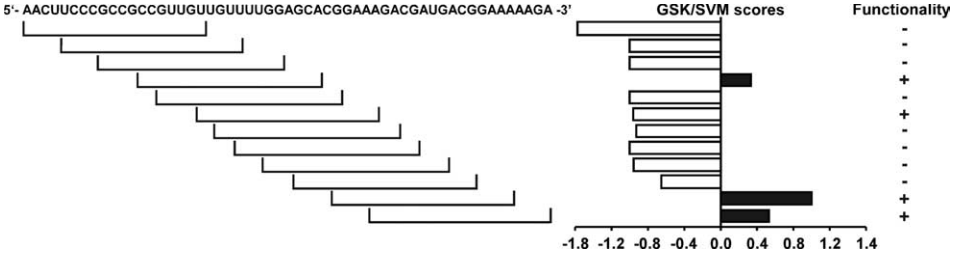

Fig. 3. Relationship between luciferase siRNA sequence and GSK/SVM scores. Brackets under the sequence indicate locations of target sequences of each siRNA. GSK/SVM scores for each siRNA were represented as bar graph. Black bar and white bar indicate positive and negative scores, respectively. Functionality; effective siRNA (+), ineffective siRNA (−).

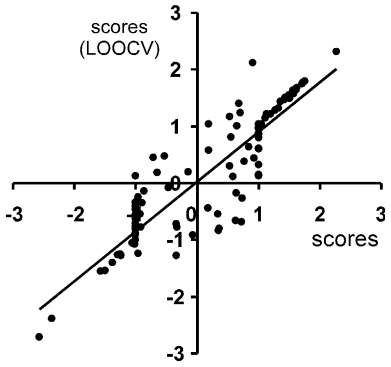

Fig. 5. Correlation between GSK/SVM scores and LOOCV GSK/SVM scores. LOOCV GSK/SVM score of each siRNA was plotted along with the GSK/SVM score (closed circles). The solid line in the plot represents linear regression ($y = 0.8772x + 0.0283$), showing a significant correlation with a correlation coefficient of 0.78.

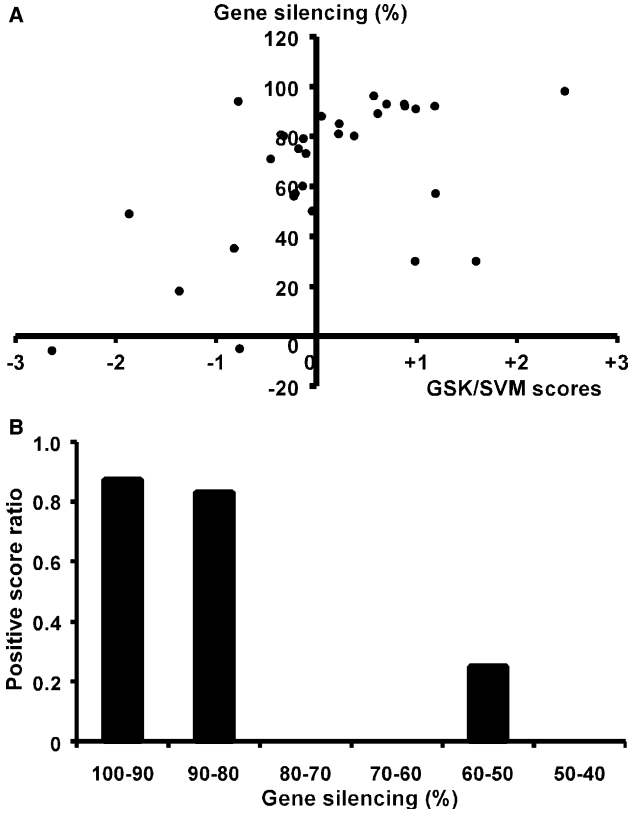

Fig. 6. Relationship between GSK/SVM score and gene silencing activity. (A) Correlation between GSK/SVM score and gene silencing activity. Gene silencing activity of each siRNA for SEAP, GAPD, and DBI was plotted along with the GSK/SVM score (closed circles). (B) Positive score ratio. Positive score ratio of different range of gene silencing was represented as bar graph.

SEAP siRNAs, GSK/SVM algorithm could classify siRNA functionality with 75% accuracy (Table 3). As shown in Table 4, with GSK/SVM and the Reynolds' algorithm, the GAPD siRNAs were classified with 90%, and 80% accuracy, respectively. Interestingly, GSK/SVM correctly classified all of the four DBI siRNAs (100% accuracy), whereas the Reynolds' classified none of them correctly (0% accuracy).

Table 3
Evaluation of predictive performance of GSK/SVM algorithm for 16 human SEAP gene siRNAs

siRNA	Position	Functionality	GSK/SVM score	GSK/SVM prediction
Human SEAP (NM_001632)				
SP-68	136	+	0.22	T
SP-147	815	–	0.99	F
SP-155	223	–	−1.86	T
SP-206	206	–	−0.22	T
SP-309	377	+	0.62	T
SP-500	568	–	−0.77	T
SP-812	812	–	1.19	F
SP-923	923	+	−0.32	F
SP-1035	1103	+	0.23	T
SP-1070	1328	–	−0.14	T
SP-1260	1138	–	−1.37	T
SP-1113	1181	–	−2.63	T
SP-1117	1117	–	−0.17	T
SP-1271	1339	–	1.59	F
SP-1795	3′-UTR	+	0.57	T
SP-2217	3′-UTR	+	2.48	T
			Accuracy	75% (12/16)

Functionality: effective (+), ineffective (−). GSK/SVM prediction: True (T), False (F).

Table 4
Comparison of predictive performance of GSK/SVM with that of Reynolds' rational design algorithm over 10 siRNAs for GAPD gene and 4 siRNAs for DBI gene

siRNA	Functionality	GSK/SVM score	GSK/SVM prediction	Reynolds prediction
(A) Human GAPD (NM_002046)				
GAPD_343	+	0.05	T	T
GAPD_347	+	0.70	T	T
GAPD_389	+	0.38	T	T
GAPD_401	+	0.89	T	T
GAPD_407	+	1.18	T	T
GAPD_409	+	0.88	T	T
GAPD_417	–	−0.13	T	F
GAPD_419	–	−0.21	T	F
GAPD_421	+	−0.77	F	T
GAPD_479	+	0.99	T	T
		Accuracy	90% (9/10)	80% (8/10)
(B) Human DBI (NM_020548.2)				
DBI 254	–	−0.04	T	F
DBI 263	–	−0.82	T	F
DBI 280		−0.10	T	F
DBI 287	–	−0.45	T	F
		Accuracy	100% (4/4)	0% (0/4)

(A): GAPD siRNAs, (B): DBI siRNAs.

These results indicated that GSK/SVM algorithm could predict siRNA functionality better, at least in some cases, than the rational design.

4. Discussion

In this study, we showed that GSK/SVM algorithm could predict siRNA functionality with high accuracy, and suggested that frequencies of subsequences are sufficient to predict siRNA functionality through classification of test data set, LOOCV, and validation on other genes that were not included in the training data set. As shown in Fig. 3, except only one instance, GSK/SVM could distinguish effective siRNAs from

the ineffective ones despite overlapping nearly identical sequences. We cannot delineate precisely how GSK/SVM can distinguish the effective sequences from ineffective ones with similar sequences, but it might be because the feature vectors were somewhat different between the siRNAs shifted by at least two bases, and SVM learned these subtle differences and utilized for classification. In the case where the prediction failed, additional factors such as position-dependent information might have made the prediction more accurate.

One of the advantages of our algorithm is that, without a prior knowledge, we could determine contribution of each parameter to siRNA functionality in the course of training on SVM. We do not know the propriety of the assumption in the rational design algorithms [14,15] that all of the position-dependent information has equal contributions to siRNA functionality, since cross-validation study was missing in the reports. Unfortunately, we could not deduce simple sequence rules out of SVM weight vectors for prediction, since SVM learns relationships between siRNA functionality and subsequences implicitly. Another advantage is that it can be applied to siRNAs shorter or longer than 21-mer in length, since our GSK/SVM algorithm utilizes the subsequence-based feature map, and not the position-dependent sequence feature.

Poor prediction accuracy for siRNA functionality has been an obstacle for application of the RNAi technology in practice [10,11]. So far, all the attempts at prediction have been based on position-dependent sequence features derived from a small number of active siRNAs [14,15]. The method described here is essentially independent from position-dependent statistics and provides a novel approach to successful prediction. Incorporation of more siRNA data will refine the feature map and improve the reliability of prediction.

Acknowledgments: We are thankful to the authors of LIBSVM who made it available through the Internet. Our source code and the supplementary materials are available upon request for academic and non-profit users.

References

[1] 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.
[2] McManus, M.T. and Sharp, P.A. (2002) Gene silencing in mammals by small interfering RNAs. Nat. Rev. Genet. 3, 737–747.
[3] Hammond, S.M., Caudy, A.A. and Hannon, G.J. (2002) Post-transcriptional gene silencing by double-stranded RNA. Nat. Rev. Genet. 2, 110–119.
[4] Hannon, G.J. (2002) RNA interference. Nature 418, 244–251.
[5] Bernstein, E., Caudy, A.A., Hammond, S.M. and Hannon, G.J. (2001) Role for a bidentate ribonuclease in the initiation step of RNA interference. Nature 409, 363–366.
[6] Hammond, S.M., Bernstein, E., Beach, D. and Hannon, G.J. (2000) An RNA-directed nuclease mediates post-transcriptional gene silencing in *Drosophila* cells. Nature 404, 293–296.
[7] Hammond, S.M., Boettcher, S., Caudy, A.A., Kobayashi, R. and Hannon, G.J. (2001) Argonaute2, a link between genetic and biochemical analyses of RNAi. Science 293, 1146–1150.
[8] Elbashir, S.M., Harborth, J., Lendeckel, W., Yalcin, A., Weber, K. and Tuschl, T. (2001) Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. Nature 411, 494–498.
[9] Stark, G.R., Kerr, I.M., Williams, B.R., Silverman, R.H. and Schreiber, R.D. (1998) How cells respond to interferons. Annu. Rev. Biochem. 67, 277.
[10] Holen, T., Amrzguioui, 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.
[11] Harborth, J., Elbashir, S.M., Vandenburgh, K., Manninga, H., Scaringe, S.A., Weber, K. and Tuschl, T. (2003) Sequence, chemical and structural variation of small interfering RNAs and short hairpin RNAs and the effect on mammalian gene silencing. Antisense Nucleic Acid Drug Dev. 13, 83–105.
[12] Elbashir, S.M., Lendeckel, W. and Tuschl, T. (2001) RNA interference is mediated by 21 and 22 nt RNAs. Genes Dev. 15, 188–200.
[13] Elbashir, S.M., Martinez, J., Patkaniowska, A., Lendeckel, W. and Tuschl, T. (2001) Functional anatomy of siRNAs for mediating efficient RNAi in *Drosophila melanogaster* embryo lysate. EMBO J. 20, 6877–6888.
[14] Reynolds, A., Leake, D., Boese, Q., Scaringe, S., Marshall, W.S. and Khvorova, A. (2004) Rational siRNA design for RNA interference. Nat. Biotechnol. 22, 326–330.
[15] Ui-Tei, K., Naito, Y., Takahashi, F., Haraguchi, T., Ohki-Hamazaki, H., Juni, A., Ueda, R. and Saigo, K. (2004) Guidelines for the selection of highly effective siRNA sequences for mammalian and chick RNA interference. Nucleic Acids Res. 32, 936–948.
[16] Vapnik, V.N. (1998) Statistical Learning Theory, Wiley, New York, USA.
[17] Khvorova, A., Reynolds, A. and Jayasena, S.D. (2003) Functional siRNAs and miRNAs exhibit strand bias. Cell 115, 209–216.
[18] Leslie, C.S., Eskin, E., Cohen, A., Weston, J. and Noble, W.S. (2004) Mismatch string kernels for discriminative protein classification. Bioinformatics 20, 467–476.
[19] Leslie, C.S., Eskin, E. and Noble, W.S. (2002) The spectrum kernel: a string kernel for SVM protein classification. Proc. Pac. Biocomput. Symp. 7, 1441–1448.
[20] Haussler, D. (1999) Convolution kernels on discrete structures. Technical Report, Department of Computer Science, University of California at Santa Cruz. Santa Cruz, CA, USA. Available from: <www.cse.ucsc.edu/~haussler/pubs.html/UCSC-CRL-99-10>.
[21] Chang, C.-C. and Lin, C.-J. (2003) LIBSVM: a library for support vector machines. Technical Report, Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan. Available from: <www.csie.ntu.edu.tw/~cjlin/papers/libsvm.pdf>.