RFRCD-siRNA: Improved design of siRNAs by random forest regression model coupled with database searching

Peng Jiang, Haonan Wu, Yao Da, Fei Sang, Jiawei Wei, Xiao Sun, Zuhong Lu*

State Key Laboratory of Bioelectronics, Department of Biological Science and Medical Engineering, Southeast University, Nanjing 210096, PR China

Abstract

Although the observations concerning the factors which influence the siRNA efficacy give clues to the mechanism of RNAi, the quantitative prediction of the siRNA efficacy is still a challenge task. In this paper, we introduced a novel non-linear regression method: random forest regression (RFR), to quantitatively estimate siRNAs efficacy values. Compared with an alternative machine learning regression algorithm, support vector machine regression (SVR) and four other score-based algorithms [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; 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; 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; M. Amarzguioui, H. Prydz, An algorithm for selection of functional siRNA sequences, Biochem. Biophys. Res. Commun. 316 (2004) 1050–1058] our RFR model achieved the best performance of all. A web-server, RFRCD-siRNA (http://www.bioinf.seu.edu.cn/siRNA/index.htm), has been developed. RFRCD-siRNA consists of two modules: a siRNA-centric database and a RFR prediction system. RFRCD-siRNA works as follows: (1) Instead of directly predicting the gene silencing activity of siRNAs, the service takes these siRNAs as queries to search against the siRNA-centric database. The matched sequences with the exceeding the user defined functionality value threshold are kept. (2) The mismatched sequences are then processed into the RFR prediction system for further analysis.

1. Introduction

RNA interference (RNAi) is a cellular process for sequence-specific destruction of mRNA [1]. Long double stranded RNA duplex or hairpin precursors are cleaved into small interfering RNAs (siRNAs) by the ribonuclease III enzyme Dicer. The siRNAs are 21–23 nucleotides (nt) in length with 2 nt 3’ overhangs [2]. Guided by RNA induced silencing complex (RISC), siRNA binds to its complementary target mRNA and induces its degradation [3].

Because the introduction of long dsRNA into mammalian cells frequently induces a fatal interferon response, RNAi-based method for silencing mammalian genes is thought to be more promising [4]. However, this method is far beyond...
readily usable for the large-scale gene silencing since only a limited fraction of siRNAs appear capable of producing highly effective RNAi [5,6]. Therefore, the distinguishing of the effective siRNAs from the ineffective ones is in high demand for gene knockout technology. In order to design effective siRNAs, Reynolds et al. formulated eight empirical rules based on the experimental evidence obtained from the systematic screening of sequence dependence of siRNA functionality [7]. Those rules include: 30–52% G/C content, presence of nucleotide ‘A’ or ‘U’ at position 15–19 (sense strand), absence of internal repeats, an nucleotide ‘A’ base at position 19 (sense strand), an nucleotide ‘A’ base at position 3 (sense strand), a ‘U’ base at position 10 (sense strand), a base other than ‘G’ or ‘C’ at position 19 (sense strand), a base other than ‘G’ at position 13 (sense strand). To improve siRNA design, many other computational approaches have been proposed [8,9]. Those approaches are classified into two groups: the score-based algorithms (or designing rules) [7,10–16] and the machine learning classification algorithms [8,17–19]. The first group approaches, which focus on finding the common features of effective siRNAs, though they initially and intuitively provide guidelines for siRNAs design, are far beyond satisfied due to low sensitivity and specificity [8,18]. The second group approaches, which are motivated by statistical learning theory, attempt to classify the siRNA into effective or ineffective class. Though those two-class classifiers provide a promising way to screen potentially effective siRNAs, it is difficult to decide the boundary between the two classes. Different algorithms used different thresholds to define the effective and ineffective siRNAs [8,17,18].

Although the observations concerning the factors which influence the siRNA efficacy give clues to the mechanism of RNAi and the machine learning classifier opens a new approach to screen effective siRNAs, the quantitative prediction of the siRNA gene silencing activity is still a challenge task.

In this paper, we introduced a novel non-linear regression method: random forest regression (RFR), to quantitatively estimate siRNA gene silencing activity value. Compared with an alternative machine learning regression algorithm: support vector machine regression (SVR), which was previously applied to the peptide binding affinity prediction [20], the development age of embryo prediction [21], etc., our RFR model outperformed it on many aspects. We further compared our prediction model with four other siRNA design rules [Reynolds et al. [7], Ui-Tei et al. [12], Hsieh et al. [22], Amarzguioui and Prydz [10]] in identifying highly effective siRNAs, our model outperformed them.

We develop a web-server RFRCDB-siRNA (http://www.bioinf.seu.edu.cn/siRNA/index.htm) to facilitate siRNA design. RFRCDB-siRNA consists of two modules: a siRNA-centric database which deposits 3589 experimental validated siRNAs from 9 publications [6,7,10–12,22–25] and a RFR prediction system. RFRCDB-siRNA works as follows: (1) Instead of directly predicting the siRNAs gene silencing activity, the service takes these siRNAs as queries to search against the siRNA-centric database. The matched sequences with exceeding the user defined functionality value threshold are kept. (2) The mismatched sequences are then processed into the RFR prediction system for further analysis.

2. Materials and methods

2.1. Datasets

We collect 3589 experimental validated siRNAs from 9 publications [6,7,10–12,22–25]. The Huesken’s dataset [23] which consisted of 2431 siRNAs was used to construct and optimize the random forest regression model. The non-redundancy Satron’s dataset [26] which consisted of 573 siRNAs were used as an independent dataset to evaluate the RFR model. All the datasets can be downloaded from http://www.bioinf.seu.edu.cn/siRNA/Supplementary/index.htm.

2.2. Methods

2.2.1. Random forest regression

Random forest, which was first proposed by Breiman [27], is an ensemble of B trees \( \{ T_1(X), \ldots, T_B(X) \} \), where \( X = (x_1, \ldots, x_p) \) is a p-dimension vector of siRNA features. The ensemble produces B outputs \( \{ \hat{Y}_1 = T_1(X), \ldots, \hat{Y}_B = T_B(X) \} \) where \( \hat{Y}_b, b = 1, \ldots, B \), is the prediction value for a siRNA sequence by the bth tree. Outputs of all trees are aggregated to produce one final prediction, \( \hat{Y} \). For regression problems, \( \hat{Y} \) is the average value of the individual tree predictions.

Given data on a set of \( n \) siRNA sequences for training, \( D = \{(X_1, Y_1), \ldots, (X_n, Y_n)\} \), where \( X_i, i = 1, \ldots, n \), is a vector of features and \( Y_i \) is experimental validated efficacy value, the training procedure are as follows [28]:

1. From the training data of \( n \) siRNA sequences, draw a bootstrap sample (i.e., randomly sample, with replacement, \( n \) siRNA sequences).
2. For each bootstrap sample, grow a tree with the following modification: at each node, choose the best split among a randomly selected subset of \( m_{try} \) (rather than all) features. Here \( m_{try} \) is essentially the only tuning parameter in the algorithm. The tree is grown to the maximum size (i.e., until no further splits are possible) and not pruned back.
3. Repeat the above steps until (a sufficiently large number) \( B \) such trees are grown.

The prediction performance of the RFR algorithm is assessed by a type of cross-validation in parallel with the training step by using the so-called Out-Of-Bag (OOB) samples. Specifically, in the process of training, each tree is grown using a particular bootstrap sample. Since bootstrapping is sampling with replacement from the training data, some of the sequences will be “left out” of the sample, while some others will be repeated in the sample. The “left out” sequences constitute the out-of-bag (OOB) sample. On the average, each tree is grown using about \( 1 - e^{-1} \approx 2/3 \) of the training sequences, leaving \( e^{-1} \approx 1/3 \) as OOB. Because OOB sequences have not been used in the tree construction, one can use them to estimate the prediction performance [28]. The random forest regression algorithm was implemented by the randomForest
We used the root mean square error (RMSE) of OOB estimation to determine the best features among many and ignore (often irrelevant) others. On the basis of previous studies on siRNA design rules [32], we performed the best. So we used the SVR with RBF kernel, as kernels (linear, RBF, 2,3-order polynomial) and the RBF kernel by the e1071 (version 1.5–12) R package [31]. We used different mulated by Vapnik [30]. The SVR algorithm was implemented by the e1071 R package [29]. The number of trees to grow was set to 1000. We used the root mean square error (RMSE) of OOB estimation to determine the best \( m \text{try} \) value.

### 2.2.2. Support vector machine regression

The basic idea of support vector machine regression was formulated by Vapnik [30]. The SVR algorithm was implemented by the e1071 (version 1.5–12) R package [31]. We used different kernels (linear, RBF, 2,3-order polynomial) and the RBF kernel performed the best. So we used the SVR with RBF kernel, as an alternative machine learning method, to compare with the RFR algorithm. The parameters \( C \) and \( \gamma \) of RBF kernel were optimized by the standard grid search.

### 2.2.3. Feature selection

On the basis of previous studies on siRNA design rules [32], we selected 15 attributes which were highly correlated with the siRNA efficacy. The features are shown in Table 1. The stability profile for each two neighboring base pairs in the siRNA sense-antisense was calculated according to the nearest-neighbor method described by Xia et al. [33]. The free energy of local target RNA secondary structure values, which were discussed in Section 4, were calculated by RNAstructure 3.71 [34].

### 2.2.4. Feature relative importance estimating

Decision tree is known for its ability to select “important” features among many and ignore (often irrelevant) others. In addition, decision tree gives an explicit model describing the relationship between features and predictions, thus easing model interpretation. Random forest, as an ensemble of trees, inherits the ability to select “important” features. However, it does not produce an explicit model. Instead, the relationship between features and activity of interest is hidden inside a “black box”. Nonetheless, a measure of how much each feature contributes to the prediction performance of random forest can be calculated in the course of training. When a feature that contributes to prediction performance is “noised up” (e.g., replaced with random noise), the performance of prediction is noticeably degrade. On the other hand, if a feature is irrelevant, “noising” it up should have little effect on the performance. Thus, we can estimate the relative importance of features according to the following procedure [28]. As each tree is grown, it makes predictions on the OOB data for that tree. At the same time, each feature in the OOB data is randomly permuted, one at a time, and each such modified dataset is also predicted by a tree. At the end of the model training process, RMSEs are calculated based on the OOB prediction as well as the OOB predictions with each features permuted. Let \( M \) be the RMSE based on the OOB prediction and \( M_j \) the RMSE based on the OOB prediction with the \( j \)th feature permuted. Then the measure of importance for the \( j \)th feature is simply \( M_j - M \).

#### 2.2.5. Assessment of model performance

The root mean square errors, root pseudo R-squared (\( q \)) and Pearson correlation coefficient (\( r \)) for assessment of the regression model are defined as:

\[
\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\text{efficacy}_i - \text{efficacy}_i^*)^2}
\]

\[
q = \sqrt{1 - \frac{\text{MSE}}{\text{Var}(\text{efficacy}^*)}}
\]

\[
r = \frac{\sum_{i=1}^{n} (\text{efficacy} - \text{efficacy}^*)(\text{efficacy}^* - \text{efficacy}^*)}{\sqrt{\sum_{i=1}^{n} (\text{efficacy} - \text{efficacy}^*)^2} \sqrt{\sum_{i=1}^{n} (\text{efficacy}^* - \text{efficacy}^*)^2}}
\]

where \( n \) is the number of siRNA sequences in the dataset; \( \text{efficacy} \) and \( \text{efficacy}^* \) are the predicted and experimentally validated efficacy values, respectively.

In order to compare the performance of RFR model with other algorithms in identifying highly efficacious siRNAs, we set a functionality value threshold of 75% (siRNAs with gene silencing ability greater than 75% are defined as effective sequences and the others are regarded as ineffective ones). Performance is evaluated in a threshold independent manner by carrying out a receiver operating characteristic (ROC) analysis [35,36]. The ROC curves are generated by plotting the sensitivity versus 1—specificity for various prediction thresholds. The area under the ROC curve (AUC) provides a single measure of overall prediction accuracy. Values of AUC above 0.7 indicate excellent prediction accuracy.
while values between 0.5 and 0.7 indicate poor accuracy [37].

3. Results

3.1. Construction of the random forest regression model

For training a machine learning algorithm, a homogeneous and sufficiently large dataset is highly important. However, the functionality of a siRNA changes variously under different biological and experimental conditions. Fortunately, the published Huesken’s dataset from a high-throughput assay makes it possible to break the bottleneck of directly combining the siRNA sequences from different resources [23]. We used Huesken’s dataset as the training dataset in our RFR prediction model.

Previous studies indicated that many attributes, such as sequence features, energy features, RNA secondary features [6,24,26,38], influenced the functionality of siRNAs. Shabalina et al. refined those features into a more optimal subset, which consisted of 15 attributes (as shown in Table 1) [32]. Therefore, those features were selected in our method. We estimated the relative importance of their contributions to the prediction model (Table 1) (more details see methods). It indicated that the summarized position-dependent consensus, the standard free energy of two neighboring base pairs in the siRNA sense-antisense in position 1 and the standard free energy difference between position 1 and 18 were the top three features which highly correlated with siRNA functionality.

A 3-fold cross-validation was used to evaluate the performance of RFR model (ntree = 1000 and mtry = 10). We obtained a RMSE and r of 8.924 and 0.851, respectively. The root pseudo R-squared (q) was up to 0.851, suggesting that the regression model fitted well to the experimental data. To improve and optimize the RFR model, we implemented a step-wise outlier exclusion procedure. If at least one sequence in the dataset produced a residual value ≥15 units in the model (the residual is defined as the absolute value of difference between the predicted efficacy and the experimental validated efficacy value), then the sequence with the maximum residual value was excluded as an outlier, and a replacement model was constructed using the remaining sequences. This procedure was repeated until all the sequences in the dataset had residual values <15 units. Following the outlier exclusion, the RFR model determined and removed 277 outliers. The r and q were increased to 0.917 and 0.918, respectively.

3.2. Comparison with support vector machine regression model

The previous machine learning regression algorithm studies indicated that the SVR model significantly outperformed other regression models on many aspects [20]. Hence we used the SVR model, as an alternative algorithm, to compare with our RFR prediction model.

A 3-fold cross-validation of the SVR model (RBF kernel with parameters C and γ optimized to 200 and 0.001, respectively) on the same dataset resulted in a RMSE of 9.414 which was 5.50% higher than the RFR model (before outlier exclusion), indicating a more deviation of the prediction result from the observed data. Besides, both the r and the q values of the SVR model were lower than those of the RFR model. To improve and optimize the SVR model, a step-wise outlier exclusion procedure with the same residual cutoff was also implemented. The procedure determined and removed 293 outliers which were 16 more than those of the RFR model. The smaller number of outliers determined by the RFR method suggests that the RFR method has more “descriptive power” than the SVR method [20]. As shown in Table 2, the SVR model underperformed our RFR model after the outliers were removed.

3.3. RFR model outperforms other siRNA design rules in identifying highly efficacious siRNA sequences

In order to compare our RFR model with other in identifying highly efficacious siRNA sequences, we used another dataset (Satron’s dataset) and defined an efficacy threshold (75%) to assign each siRNA sequences into efficacy or inefficacy category. For the RFR model and the SVR model, rather than using a k-fold cross-validation, we used the cross-dataset testing (Huesken’s dataset for training and Satron’s dataset for testing).

We applied the ROC curves which depicted the relative trade-off between the sensitivity and the specificity by altering the misclassification cost to visualize and compare different methods. As shown in Fig. 1, the RFR model achieved the highest AUC value, indicating that our RFR model performed best in identifying highly efficacious siRNA sequences.

<table>
<thead>
<tr>
<th>Table 2 – The performance of our RFR model and the alternative SVR model in quantificational modeling the siRNA efficacy</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>RFR</strong></td>
</tr>
<tr>
<td><strong>With outlier</strong></td>
</tr>
<tr>
<td>RMSE</td>
</tr>
<tr>
<td>r</td>
</tr>
<tr>
<td>q</td>
</tr>
</tbody>
</table>

*a 277 outliers were excluded.

*b 293 outliers were excluded.
Fig. 1 – The performance of our RFR model, the alternative SVR model and other four score-based design rules in identifying highly efficacious siRNA on Satron’s dataset. For RFR and SVR methods, Huesken’s dataset is used for training models. The area under curve (AUC) is provided following the label of each predicting method.

3.4. Comparison with other machine learning methods

Huesken et al. [23] randomly selected 2431 siRNAs targeting 34 mRNA species, assayed through a high-throughput fluorescent reporter gene system, together with the potency prediction system BIOPREDsi, which was based on an artificial neural network model. Vert et al. proposed a LASSO regression model [39] for the same purpose [40]. Three sets of features (sparse: the presence or absence of each nucleotide at each of the positions; spectral: the number of occurrences of each nucleotide motif of length 1–3; Composite: concatenation of the Sparse and Spectral representations) were introduced.

In order to compare our RFR model with them, two independent datasets (the Reynolds dataset [7] of 240 siRNAs and the Vickers dataset [25] of 76 siRNAs) were used as the testing datasets. All the machine learning methods were trained on the same dataset (Huesken dataset [23]). The results, as shown in Table 3, indicated that the Pearson coefficients between predicted and true efficacy of our RFR prediction were higher than those of the other two methods on both datasets.

3.5. Testing the RFR model with another dataset

HuSiDa [41] is a gene-centric database which provides sequences of published functional siRNA molecules targeting human genes. To retrieve functional siRNAs sequences for a specific gene, the human siRNA database can be searched for UniGene cluster IDs, gene names, RefSeq accession numbers and target gene descriptions via a web interface (www.human-siRNA-database.net). P53 (Accession: NM_000546), APC (Accession: NM_000038) and RANBP2L1 (Accession: NM_005054) are three important tumor related genes. We take them as queries to search HuSiDa and retrieve three experimental validated siRNAs (one siRNA for a gene, respectively). Then we take these siRNAs to run in our RFR prediction system. The results show that all of the predicted efficacy values are greater than 60% (60.9, 74.1 and 78.2%, respectively), which is consensus with the experimental data. (The HuSiDa does not provide the exact siRNA efficacy value, but all of the siRNAs, which are recommended to the users, have at least 50% experimental validated efficacy values.)

3.6. Implementation

To facilitate the designing of siRNAs, a web-server, named RFRCD-siRNA, has been developed. RFRCD-siRNA consists of two independent modules: a siRNA-centric database and a RFR prediction system. As shown in Fig. 2, the working process is as follows:

1. The target mRNA is scanned by a 19 nt-sliding window with a step of 1 nt.
2. Rather than directly predicting siRNA efficacy, the siRNA candidates are used as queries to search against the siRNA-centric database. The matched sequences with exceeding the user defined efficacy threshold are kept.
3. The mismatched sequences are further predicted by the RFR prediction system. The sequences with exceeding the efficacy threshold are kept and others are discarded.
4. An output format example is shown in Fig. 3. The publication references are followed for those experimentally validated siRNAs.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>LASSO model</th>
<th>BIOPREDsi</th>
<th>RFR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sparse</td>
<td>Spectral</td>
<td>Composite</td>
</tr>
<tr>
<td>Reynolds (240)</td>
<td>0.54</td>
<td>0.49</td>
<td>0.55</td>
</tr>
<tr>
<td>Vickers (76)</td>
<td>0.58</td>
<td>0.54</td>
<td>0.49</td>
</tr>
</tbody>
</table>

All of the models are trained on the Huesken’s dataset [23]. The performance is measured as the Pearson coefficient between predicted and true efficacy. All of the correlations are significant (p-value <0.01).

a LASSO model with different features [40]. Sparse: the presence or absence of each nucleotide at each of the 19 positions; spectral: the number of occurrences of each nucleotide motif of length 1–3 for the 19 length siRNA sequence; composite: concatenation of the sparse and spectral representations.
Fig. 2 – Schematic representation of the working procedure of RFRCD-siRNA.

validated sequences. Considering that the misclassifying siRNAs with high efficacy rates as non-functionality are of much lesser consequence than the misclassifying siRNAs with low efficacy rates as functionality, four other siRNA design scores (Reynolds et al. [7], Ui-Tei et al. [12], Hsieh et al. [22], Amarzguioui and Prydz [10]) are also provided for the RFR prediction sequences. In this case, the conflicting results can be noted and evaluated by users.

4. Discussion

Although the observations concerning the factors which influence the siRNA efficacy give clues to the mechanism of RNAi, the accurately prediction of the siRNA efficacy is still a challenge task. In our study, we proposed a novel machine learning algorithm: random forest regression to quantitatively model the siRNA efficacy. To improve and optimize the RFR model, a step-wise outlier exclusion procedure is implemented. Compared with SVR model, the RFR model outperformed it on several aspects. Compared with four other siRNA design rules (Reynolds et al. [7], Ui-Tei et al. [12], Hsieh et al. [22], Amarzguioui and Prydz [10]) in identifying highly efficacious siRNA sequences using ROC curves, our model achieved the highest AUC value. Compared with two machine learning methods (BIOPREDsi and LASSO regression model), our RFR model outperformed them on two independent datasets testing. A web-server, named RFRCD-siRNA, has been developed and it is available at the website: http://www.bioinf.seu.edu.cn/siRNA/index.htm. The most outstanding character of our web-server is that we incorporating both the siRNA database searching and the RFR prediction model to improve the siRNA design.

Several studies indicated that the local target mRNA stabilities were correlated with the efficacy of siRNAs [42,43]. However, a recent study, which used the same training dataset with our RFR model, showed that the local target mRNA stabilities were less stably correlated with the siRNA efficacy [32]. Moreover, Truss et al. claimed that the relevant information required for the efficient prediction of siRNAs efficacy was contained in its primary sequence and the mRNA secondary structure had probably been overrated in the past [41]. To further assess the importance of the local target mRNA stabilities in a siRNA efficacy prediction model, an independent dataset: Vickers dataset [25], which consists of 76 siRNAs, is used. A significant correlation between siRNA efficacy and the free energy of the local target mRNA is detected (Pearson correlation coefficient is 0.23, p-value <0.05). Then we combine the free energy of the local target mRNA feature in our prediction model. The results (3-fold cross-validation) indicate that the combination of the local target mRNA secondary structure has not improved our RFR prediction model and even declines the prediction performance (the root mean square errors of the RFR models with and without the free energy of the secondary structure of the target mRNA are 21.36 and 21.08, respectively). In order to analyze the reasons, we use the principal component analysis (PCA) to identify a series of new orthogonal axes accounting for the greatest variation among the 15 features (without the local target mRNA secondary structure feature). Then a linear regression analysis between the first two axes in PCA analysis of the 15 features and the free energy of the secondary structure of the target mRNA feature is implemented. The results, as shown in Table 4, indicate that the axis 2 of the major trend (15 features) is significantly correlated (p-value <0.05) with the local free energy of the secondary structure of the target mRNA feature. In other words, the 15 features which we use in the RFR prediction model correlate with the secondary structure of the target mRNA feature. The performance of a machine learning prediction model can be increased by orthologous features but decreased by analogous (or corre-
Fig. 3 – An output example of RFRCDB-siRNA. For the experimentally validated sequences, the references are given. For the prediction based sequences, four other siRNA design scores (Reynolds et al. [7], Ui-Tei et al. [12], Hsieh et al. [22], Amarzguioui and Prydz [10]) are followed.
lated) features. That can partly explain the reason why the combination of the local free energy of the local secondary structure of the target mRNA does not improve the prediction performance.

Poor prediction accuracy for siRNA functionality is an obstacle for application of the RNAi technology in practice. Our future work will include: collect more experimental validated data and information in database, incorporate and refine attributes to improve RFR prediction system. We aim to continue to expand RFRDDB-siRNA into an increasingly valuable web-service.

REFERENCES


