BMC Research Notes

TECHNICAL NOTE Open Access

Mini-clusters with mean probabilities for identifying effective siRNAs

Jia Xingang[1,2*], Zuhong Lu[2*] and Qiuhong Han[3]

Abstract

Background: The distinction between the effective siRNAs and the ineffective ones is in high demand for gene knockout technology. To design effective siRNAs, many approaches have been proposed. Those approaches attempt to classify the siRNAs into effective and ineffective classes but they are difficult to decide the boundary between these two classes.

Findings: Here, we try to split effective and ineffective siRNAs into many smaller subclasses by RMP-MiC(the relative mean probabilities of siRNAs with the mini-clusters algorithm). The relative mean probabilities of siRNAs are the modified arithmetic mean value of three probabilities, which come from three Markov chain of effective siRNAs. The mini-clusters algorithm is a modified version of micro-cluster algorithm.

Conclusions: When the RMP-MiC was applied to the experimental siRNAs, the result shows that all effective siRNAs can be identified correctly, and no more than 9% ineffective siRNAs are misidentified as effective ones. We observed that the efficiency of those misidentified ineffective siRNAs exceed 70%, which is very closed to the used efficiency threshold. From the analysis of the siRNAs data, we suggest that the mini-clusters algorithm with relative mean probabilities can provide new insights to the applications for distinguishing effective siRNAs from ineffective ones.

Findings

RNA interference (RNAi) is a cellular process for sequence specific destruction of mRNA [1]. The broad mechanistic details for the pathway have been largely characterized. Long double-stranded RNAs duplex or hairpin precursors are cleaved into small interfering RNAs (siRNAs) by the ribonuclease III enzyme Dicer. The typical siRNAs have a 19-nucleotide paired region followed by a 2-nucleotide 3' overhang [2]. The siRNAs are used to initiate RNAi [3-6]. Therefore, the distinguishing the effective siRNAs from the ineffective ones is in high demand for gene knockout technology. In order to design effective siRNAs, many computational approaches have been proposed [7-20]. Some approaches focus on finding the common features of effective siRNAs, though they initially and intuitively provide guidelines for siRNAs design, are far from satisfied due to low sensitivity and specificity [8,18]. The other approaches are motivated by statistical learning theory, attempt to classify the siRNAs into effective and ineffective classes. Although those two-class classifiers provide a promising way to screen potentially effective siRNAs, it is difficult to decide the boundary between the two classes.

Here, we use the set of effective siRNAs to estimate distributions of three Markov chains, where the order of three Markov chain are 1, 2 and 3, respectively. Each siRNA obtain three probabilities from the distributions of three Markov chains. Based on three probabilities of siRNAs, we introduce a robust feature of siRNAs, the relative mean probabilities, which is the modified arithmetic mean value of these three probabilities. It should be noticed that the siRNAs with similar relative mean probabilities have same efficacy(effective/ineffective) usually, most relative mean probabilities of effective siRNAs exceed most ineffective ones. However, there is no clear boundary between these two classes, so we give up the attempt of dichotomy. We try to split these two classes into many smaller effective or ineffective subclasses, respectively. Thus, we distinguish effective siRNAs from the ineffective ones by a mini-clusters algorithm, which adopted from [21](see Materials and methods). By RMP-MiC(the relative mean probabilities with the mini-clusters), all effective siRNAs

*Correspondence: hanqh15@163.com; 1010344@seu.edu.cn
[1]Department of Mathematics, Southeast University, Nanjing 210096, PR China
[2]State Key Laboratory of Bioelectronics, School of Biological Science and Medical Engineering, Southeast University, Nanjing 210096, PR China
Full list of author information is available at the end of the article

can be identified correctly, and no more than than 9% ineffective siRNAs are misidentified as effective siRNAs. We observed that the efficiency of those misidentified ineffective siRNAs exceed 70%, which is very closed to the used efficiency threshold.

Methods

Estimating distributions of siRNAs

The siRNAs can be represented as an 19-tuple of vector. $x_i = (x_{i1}, x_{i2}, \cdots, x_{i19})$ is the i-th siRNA where x_{ij} represents its j-th nucleotide. Effective siRNAs are used to estimate Q_h, where Q_h is distribution of a h-order Markov chain, h equals 1, 2 and 3, respectively. $Q_h(i)$ is probability of the i-th siRNA in Q_h. We use $Q_h(i)(h = 1, 2, 3)$ to construct $Q_4(i)$, where

$$Q_4(i) = \frac{Q_1(i) + Q_2(i) + Q3(i)}{\lceil Q_1(i) \rceil + \lceil Q_2(i) \rceil + \lceil Q_3(i) \rceil},$$

$$Q_h(i) = Pr(x_{i1} \cdots x_{ih}) \prod_{s=h+1}^{19} q_h(ij),$$

$$q_h(ij) = Pr(x_{is} | x_{i(j-h)}, \cdots, x_{i(j-1)}).$$

If $Q_h(i)$ exceed zero, $\lceil Q_h(i) \rceil$ is 1, otherwise $\lceil Q_h(i) \rceil$ is zero. $Q_4(i)$ name as relative mean probabilities of x_i. It can be noticed that the siRNAs with similar relative mean probabilities have alike efficiency usually.

Mini-clusters algorithm

Based on the relative mean probabilities of siRNAs, we distinguish effective siRNAs from the ineffective ones by a mini-clusters algorithm, which adopted from [21], a commonly used micro-cluster algorithm. It is sketched as below.

Define the distance between i-th and j-th siRNAs as

$$d_h(i, j) = \sqrt{(Q_h(i) - Q_h(j))^2}.$$

We put the closest two elements in a cluster. In subsequent steps, we examine the two closest elements not already in a cluster. If either or both of these are closer to some element within a cluster, we put each element in the cluster to which it is closest, otherwise, we form a new cluster. Repeat this step until all siRNAs have been put into a mini-cluster.

For the siRNAs in testing set, we consider that their efficiency are unknown. In the process of testing the sensitivity and specificity, a mini-cluster is considered as effective if it has an effective siRNAs, and be considered as ineffective if all siRNAs are ineffective, otherwise its efficacy is uncertain. We denote effective, ineffective and uncertain mini-clusters as

$$A_1, A_2, \cdots, A_u; \quad A_{u+1}, A_{u+2}, \cdots, A_a; \quad B_1, B_2, \cdots, B_b;$$

respectively. Define the distance of A_i and B_j as

$$d_h(A_i, B_j) = \min_{u \in A_i, v \in B_j} d_h^s(u, v).$$

If

$$d_h(A_{i1}, B_j) = \min_{i=1,2\cdots,s} d_h(A_i, B_j),$$

the efficacy of B_j is regarded as that of A_{i1}. In other words, each uncertain mini-cluster is merged into the nearest determined ones.

Availability

Testing the performance of mini-clusters

To test the performance of RMP-MiC, it was firstly applied to a simulation data. The sequences of simulation data set belong to two groups X and Y, each of them contains 5 nucleotides. In order to simplify the problem, we assume the nucleotides are generated from different 1-order Markov chain, that is, the relative mean probabilities of sequences equal the probabilities of their 1-order Markov chain. For X, the probabilities of U base and C base at position 1 are 0.75 and 0.25, conditional probabilities of position 2 are

$$Pr(A|U) = 0.75, \quad Pr(U|U) = 0.25, \quad Pr(G|C) = 1$$

and others are zero. At 3-5 position, we assume that all conditional probabilities are 0.25. For each sequence of Y, we assume that 'U' base at position 1 and 'A' base at position 5 or 'C' base at position 1 and 'G' base at position 5 can not appear at the same time, nucleotides are random at other positions. An illustrative example within the simulation data is shown in Table 1, which consists of 17 sequences. These 17 sequences belong to two groups X and Y. The two groups are of size 10 and 7, respectively. The relative mean probabilities of these 17 sequences are shown in Table 1. For comparison, we also applied K-mean with Euclidean to cluster all sequences into 2 cluster, where the distance between two sequences are Euclidean distance of their mean probabilities. The clustering results by two methods are shown in Table 1.

In Table 1, RMP-MiC grouped these 17 sequences into 4 mini-clusters, sequences of each mini-clusters come from the same group. The Euclidean algorithm were clusters 7 sequences of cluster 1 incorrectly grouped in cluster 2. The reason may be that Euclidean distance takes the difference between data points directly, it may be overly sensitive to the magnitude of changes To further test these methods, we applied it to a larger data set containing 1,000 samples. Results were similar to those observed for the smaller data set(data not shown).

Table 1 List of simulation data and clustering results by two algorithm

								Results	
Group *X*	**Sequences**	P_1	P_2	P_3	P_4	P_5	$Q(4)$	**RMP-MiC**	**K-mean**
a1	UAAUC	0.75	0.75	0.25	0.25	0.25	0.0088	1	1
a2	UACCG	0.75	0.75	0.25	0.25	0.25	0.0088	1	1
a3	UAGAA	0.75	0.75	0.25	0.25	0.25	0.0088	1	1
a4	UUCCG	0.75	0.25	0.25	0.25	0.25	0.0029	2	2
a5	UUGAA	0.75	0.25	0.25	0.25	0.25	0.0029	2	2
a6	UUUGU	0.75	0.25	0.25	0.25	0.25	0.0029	2	2
a7	CGAUC	0.25	1	0.25	0.25	0.25	0.0039	3	2
a8	CGCCG	0.25	1	0.25	0.25	0.25	0.0039	3	2
a9	CGGAA	0.25	1	0.25	0.25	0.25	0.0039	3	2
a10	CGUGU	0.25	1	0.25	0.25	0.25	0.0039	3	2
Group *Y*									
b1	AACGA	0	0	0.25	0.25	0.25	0	4	2
b2	AUGGA	0	0	0.25	0.25	0.25	0	4	2
b3	UCAGC	0.75	0	0.25	0.25	0.25	0	4	2
b4	UGUUC	0.75	0	0.25	0.25	0.25	0	4	2
b5	UCCUG	0.75	0	0.25	0.25	0.25	0	4	2
b6	CCAAA	0.25	0	0.25	0.25	0.25	0	4	2
b7	CCUAC	0.25	0	0.25	0.25	0.25	0	4	2

P_1 is the probabilities of the leftmost nucleotides. $P_i (i = 2, 3, 4, 5)$ is conditional probabilities of the i-th position. $Q(4)$ is the the relative mean probabilities of sequences.

Identifying results of the experimental siRNAs

The data set can be downloaded from http://www.bioinf.seu.edu.cn/siRNA/Supplementary/index.htm. It collects 3589 experimental validated siRNAs from 9 publications [7,10-12,22-26]. The efficiency threshold of siRNA to be effective is 80%. According to this threshold, the data set has 582 effective siRNAs and 3007 ineffective siRNAs.

To validate the performance of $Q_4(i)$ with mini-clusters, we apply them to data set of experimental siRNAs, where $Q_4(i)$ are estimated by all effective siRNAs. The identifying results are summarized in Table 2. In fact, all effective siRNAs are correctly identified and only 264 ineffective siRNAs are misidentified into effective siRNAs by $Q_4(i)$ with mini-cluster. It should be noticed that when ineffective siRNAs are misidentified into effective siRNAs, its efficiency exceeds 70% mostly.

For comparison, we applied the $Q_h^s(i)(h = 1, 2, 3)$ with mini-clusters to the same data. The K-mean with Euclidean was also applied to cluster all sequences into 2 cluster, where the distance between two sequences are Euclidean distance of their $Q_4(i)$, the number of clusters is the same as the number of mini-clusters of $Q_4(i)$. The results are also summarized in Table 2. These results show that all effective siRNAs are correctly identified and 610, 534 and 100 ineffective siRNAs are misidentified with effective siRNAs by $Q_1(i)$, $Q_2(i)$ and $Q_3(i)$ with mini-cluster, respectively.

For comparison, The K-mean with Euclidean was also applied to cluster all sequences into 2 cluster, where the

Table 2 The identifying results of siRNAs by five different algorithms

Algorithm	Feature	Total	Sensitivity(%)	Specificity(%)
Mini-cluster	$Q_1(i)$	1192	1	48.83
Mini-cluster	$Q_2(i)$	1116	1	52.15
Mini-cluster	$Q_3(i)$	682	1	85.34
Mini-cluster	$Q_4(i)$	846	1	68.79
K-means	$Q_4(i)$	1588	1	36.65

The total number is the number of the identified effective siRNAs. Sensitivity, the number of effective siRNAs/582. Specificity, the number of effective siRNAs/total number of cluster members.

distance between two sequences are Euclidean distance of their $Q_4(i)$, the number of clusters is the same as the number of mini-clusters of $Q_4(i)$. The results are also summarized in Table 2. These result shows that all effective siRNAs are correctly identified but 1006 ineffective siRNAs are misidentified with effective siRNAs.

To test the sensitivity and specificity of $Q_4(i)$ with mini-clusters, 80% effective siRNAs are chosen as training data set. The siRNAs of training data set are used to estimate $Q_4(i)$. To assure each siRNA may be in test set, we construct 1,000 different training data set. The results show that only 13 effective siRNAs are incorrectly identified and 516 ineffective siRNAs are misidentified with effective siRNAs , where the number of the misidentified effective and ineffective siRNAs are the mean values acquired from averaging across each training set. The result shows that $Q_4(i)$ with mini-clusters is reliable for identifying effective siRNAs. However, when we use $Q_3(i)$ to substitute $Q_4(i)$, only 18% effective siRNAs of training data set can identify correctly. The reason may be that many $Q_3(i)$ of effective siRNAs of training data set become zero. It can result in which these effective siRNAs are misidentified to ineffective siRNAs. However, even if $Q_3(i)$ of these effective siRNAs are zero but their $Q_1(i)$ and $Q_2(i)$ may be very large, so their $Q_4(i)$ are also different with ineffective siRNAs. Thus, they may construct new mini-clusters or enter into effective mini-clusters.

Secondly, we randomly generate 1,0000 simulation siRNAs. A new data set of siRNAs are formed by these 1,0000 simulation siRNAs and 3587 experimental siRNAs. By $Q_4(i)$ with mini-clusters, these 1,3587 siRNAs are put into different mini-clusters, where 1587 simulation siRNAs are put into effective mini-clusters, $Q_4(i)$ are estimated by all effective experimental siRNAs. The efficiency of these 1587 simulation siRNAs are de novo validated by a web-server RFRCDB-siRNA [27], which is available at http://www. bioinf.seu.edu.cn/siRNA/index.htm. By the web-server, 1536 simulation siRNAs are identified as effective. The result shows that effective siRNAs should have specific features at some positions, and $Q_4(i)$ can incarnate these specific features.

Identifying results of the shRNAs

To systematically analyze the interplay between nucleotide composition, shRNA processing, and biologic activity, Christof Fellmann et al transduced the entire Sensor library into human HEK293T and chicken ERC cells, generated and quantified small RNA libraries designed to represent shRNA intermediates after major biogenesis steps, which contains 18,720 shRNAs [28]. The efficiency threshold of shRNA to be effective is that its score exceed 10. According to this threshold, the data set has 453 effective siRNAs and 18267 ineffective siRNAs. The data set of shRNAs can be downloaded from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3130540/?tool=pubmed.

To validate the performance of $Q_4(i)$ with mini-clusters to distinguish effective shRNAs, it is applied to data set of shRNAs, where $Q_4(i)$ are estimated by all effective shRNAs. The identifying results shows that all effective shRNAs are correctly identified and only 1446 ineffective shRNAs are misidentified into effective shRNAs by $Q_4(i)$ with mini-cluster. It should be noticed that when ineffective shRNAs are misidentified into effective shRNAs, their efficiency are very closed to the effective threshold.

Comparison to existing design algorithms

To compare our results to existing siRNA-based design tools, we obtained the top predictions for transcripts using three different algorithms [17-19] and compared them to the 50 highest scoring Sensor-derived shRNAs for gene. Strikingly, exceed 70% of scoring shRNAs were not identified in the top 50 predictions of any algorithm. While such false negatives, in principle, may have little practical significance, the majority of algorithm-predicted shRNAs did not score in the Sensor assay, closely resembling their low validation rate in empirical testing. Together, these results demonstrate that siRNA algorithms are poor at predicting potent shRNAs [29] and underscore the value of the Sensor approach.

Requirements

Since effective siRNAs have specific nucleotides at some position, it is reasonable to use relative mean probabilities as their feature indicator. However, effective siRNAs may have different relative mean probabilities, but the mini-clusters algorithm place siRNAs with similar relative mean probabilities in the same mini-clusters.

In fact, relative mean probabilities can be viewed as specific probabilities of siRNAs, so the absolute value of their logarithm can be regarded as entropies of siRNAs. Since siRNAs with similar relative mean probabilities are in the same mini-clusters, the deviance of efficiency of siRNAs can be regarded as the difference in their entropies.

Conclusions

From the analysis of the siRNAs data, we demonstrate that mini-clusters algorithm using $Q_4(i)$ are appropriate for analyzing siRNAs data. Its success indicates that an effective algorithms for analyzing biological data must be based on an understanding of the biological nature of the experimental data.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions
Conceived and designed the experiments: JX and ZL. Performed the experiments: QH. Analyzed the data: JX. Wrote the paper: JX. All authors read and approved the final manuscript.

Acknowledgements
The work is supported by National Natural Science Foundation of China (Project No. 30871393). Funding to pay Open access publication charges for this article was provided by the National Natural Science Foundation of China (No. 30871393).

Author details
[1]Department of Mathematics, Southeast University, Nanjing 210096, PR China. [2]State Key Laboratory of Bioelectronics, School of Biological Science and Medical Engineering, Southeast University, Nanjing 210096, PR China. [3]Department of Mathematics, Nanjing Forestry University, Nanjing 210037, PR China.

Received: 28 October 2011 Accepted: 7 August 2012
Published: 18 September 2012

References

1. Fire A, Xu S, Montgomery MK, Kostas SA, Driver SE, Mello CC: **Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans.** *Nature* 1998, **391:**806–811.
2. Zamore PD, Tuschl T, Sharp PA, Bartel DP: **RNAi: double-stranded RNA directs the ATP-dependent cleavage of mRNA at 21 to 23 nucleotide intervals.** *Cell* 2000, **101:**25–33.
3. Elbashir SM, Lendeckel W, Tuschl T: **RNA interference is mediated by 21- and 22-nucleotide RNAs.** *Genes* 2001, **Dev 15:**188–200.
4. Hamilton A, Voinnet O, Chappell L, Baulcombe D: **Two classes of short interfering RNA in RNA silencing.** *EMBO J* 2002, **21:**4671–4679.
5. Samuel-Abraham S, Leonard JN: **Staying on message: design principles for controlling nonspecific responses to siRNA.** *FEBS J* 2010, **277:**4828–4836.
6. Sioud M: **Deciphering the code of innate immunity recognition of siRNAs.** *Methods Mol Biol* 2009, **487:**41-59.
7. Reynolds A, Leake D, Boese Q, Scaringe S, Marshall WS, Khvorova A: **Rational siRNA design for RNA interference.** *Nat Biotechnol* 2004, **22:**326–330.
8. Jia P, Shi T, Cai Y, Li Y: **Demonstration of two novel methods for predicting functional siRNA efficiency.** *BMC Bioinformatics* 2006, **7:**271. (electronic resource).
9. Matveeva O, Nechipurenko Y, Rossi L, Moore B, Saetrom P, Ogurtsov AY, Atkins JF, Shabalina SA: **Comparison of approaches for rational siRNA design leading to a new efficient and transparent method.** *Nucleic Acids Res* 2007, **35:**e63.
10. Amarzguioui M, Prydz H: **An algorithm for selection of functional siRNA sequences.** *Biochem Biophys Res Commun* 2004, **316:**1050–1058.
11. Khvorova A, Reynolds A, Jayasena SD: **Functional siRNAs and miRNAs exhibit strand bias.** *Cell* 2003, **115:**209–216.
12. Ui-Tei K, Naito Y, Takahashi F, Haraguchi T, Ohki-Hamazaki H, Juni A, Ueda R, Saigo K: **Guidelines for the selection of highly effective siRNA sequences for mammalian and chick RNA interference.** *Nucleic Acids Res* 2004, **32:**936–948.
13. Takasaki S, Kotani S, Konagaya A: **An effective method for selecting siRNA target sequences in mammalian cells.** *Cell Cycle* 2004, **3:**790–795.
14. Holen T: **Efficient prediction of siRNAs with siRNArules 1.0: an open-source JAVA approach to siRNA algorithms.** *RNA* 2006, **12:**1620–1625.
15. Gong W, Ren Y, Xu Q, Wang Y, Lin D, Zhou H, Li T: **Integrated siRNA design based on surveying of features associated with high RNAi effectiveness.** *BMC Bioinformatics* 2006, **7:**516. (electronic resource).
16. Katoh T, Suzuki T: **Specific residues at every third position of siRNA shape its efficient RNAi activity.** *Nucleic Acids Res* 2007, **35:**e27.
17. Ge G, Wong GW, Luo B: **Prediction of siRNA knockdown efficiency using artificial neural network models.** *Biochem Biophys Res Commun* 2005, **336:**723–728.
18. Teramoto R, Aoki M, Kimura T, Kanaoka M: **Prediction of siRNA functionality using generalized string kernel and support vector machine.** *FEBS Lett* 2005, **579:**2878–2882.
19. Ladunga I: **More complete gene silencing by fewer siRNAs: transparent optimized design and biophysical signature.** *Nucleic Acids Res* 2007, **35:**433–440.
20. Matveeva OV, Kang Y, Spiridonov AN, Saetrom P, Nemtsov VA, Ogurtsov AY, Nechipurenko YD, Shabalina SA: **Optimization of duplex stability and terminal asymmetry for shRNA design.** *PLoS ONE* 2010, **5:**e10180.
21. Beliakov G, King M: **Density based fuzzy c-means clustering of non-convex patterns.** *Eur J Operational Res* 2006, **173**(3):717–728.
22. Hsieh AC, Bo R, Manola J, Vazquez F, Bare O, Khvorova A, Scaringe S, Sellers WR: **A library of siRNA duplexes targeting the phosphoinositide 3-kinase pathway: determinants of gene silencing for use in cell-based screens.** *Nucleic Acids Res* 2004, **32:**893–901.
23. Huesken D, Lange J, Mickanin C, Weiler J, Asselbergs F, Warner J, Meloon B, Engel S, Rosenberg A, Cohen D, *et al*: **Design of a genome-wide siRNA library using an artificial neural network, Nat.** *Biotechnol* 2005, **23:**995–1001.
24. Jagla B, Aulner N, Kelly PD, Song D, Volchuk A, Zatorski A, Shum D, Mayer T, De Angelis DA, Ouerfelli O, et al: **Sequence characteristics of functional siRNAs.** *RNA* 2005, **11:**864–872.
25. Vickers TA, Koo S, Bennett CF, Crooke ST, Dean NM, Baker BF: **Efficient reduction of target RNAs by small interfering RNA and RNase H-dependent antisense agents. A comparative analysis.** *J Biol Chem* 2003, **278:**7108–7118.
26. Harborth J, Elbashir SM, Vandenburgh K, Manninga H, Scaringe SA, Weber K, Tuschl T: **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* 2003, **13:**83–105.
27. Jiang P, Wu H, Da Y, Sang F, Wei J, Sun X, Lu Z: **RFRCDB-siRNA: Improved design of siRNAs by random forest regression model coupled with database searching.** *Comput Methods Programs Biomed* 2007, **87**(3):230–238.
28. Fellmann C, Zuber J, McJunkin1 K, Chang1 K, Malone1 CD, Dickins RA, Xu Q, Hengartner M, Elledge SJ, Hannon GJ, Lowe SW: **Functional identification of optimized RNAi triggers using a massively parallel Sensor assay.** *Mol Cell* 2011, **41**(6):733–746.
29. Bassik MC, Lebbink RJ, Churchman LS, Ingolia NT, Patena W, LeProust EM, Schuldiner M, Weissman JS, McManus MT: **Rapid creation and quantitative monitoring of high coverage shRNA libraries.** *Nat Methods* 2009, **6:**443–445.

doi:10.1186/1756-0500-5-512
Cite this article as: Xingang *et al.*: Mini-clusters with mean probabilities for identifying effective siRNAs. *BMC Research Notes* 2012 **5**:512.