Deposited research article

Prediction for Target Sites of Small Interfering RNA Duplexes in SARS Coronavirus

Fengmin Ji[1] and Liaofu Luo[23]

Addresses: [1]Institute for Biological Science and Technology, Faculty of Sciences, North Jiaotong University, Beijing, 100044 China. [2]Laboratory of Theoretical Biophysics, Faculty of Sciences and Technology, Inner Mongolia University, Hohhot, 010021 China. [3]Center for Theoretical Biology, Peking University, Beijing, 100871 China.

Correspondence: Liaofu Luo. E-mail: lfluo@mail.imu.edu.cn

Posted: 16 January 2004

Genome ***Biology*** 2004, **5:**P6

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2004/5/2/P6

Received: 14 January 2004

This is the first version of this article to be made available publicly.

Prediction for Target Sites of Small Interfering RNA Duplexes in SARS Coronavirus

Fengmin Ji[1] Liaofu Luo[2, 3 *]

[1] Institute for Biological Science and Technology, Faculty of Sciences, North Jiaotong University, Beijing, 100044 China;

[2] Laboratory of Theoretical Biophysics, Faculty of Sciences and Technology, Inner Mongolia University, Hohhot, 010021 China;

[3] Center for Theoretical Biology, Peking University, Beijing, 100871 China

Key words

SARS Coronavirus, Small Interfering RNA Duplex, Target Site, Anti-virus drug design

Running head

Target Sites of siRNA in SARS-CoV

* corresponding author, to whom correspondence should be addressed.

Mailing address: Liaofu Luo

Laboratory of Theoretical Biophysics, Faculty of Sciences and Technology, Inner Mongolia University, Hohhot, 010021 China

Email address: lfluo@mail.imu.edu.cn Fax: 0086-471-4951761

Abstract

RNA interference is used for SARS-related pharmaceutical research and development. Following bioinformatic method twenty seven 21~25 base-long sequence segments in SARS-CoV genome are predicted as the optimal target sites of small interfering RNA duplexes.

SARS(severe acute respiratory syndrome) was first identified in Guangdong Province,China and rapidly spread to many regions in China and around the world. It caused death and disaster to thousands of human beings. However, the active drug in treating SARS has not been found yet. The genome sequences determined by several groups[1]-[3] show that it is a variant of coronaviruses, belonging to single-stranded plus sense RNA viruses. The genome is about 30 kb in length, and its several encoded proteins have been separated and purified. This provides a sound basis for SARS related pharmaceutical research and development . The use of double-stranded RNA (dsRNA) to manipulate gene expression (RNA interference or RNAi) has been proved highly effective, at least 10 times more effective than either using sense or antisense RNAs alone[4]. The RNAi triggered by dsRNA is a phenomenon of homology-dependent gene silencing[5][6][7]. It was found that the small interfering RNA (siRNA, 21-25 nt long) plays an important role in RNAi-related gene silencing pathways[8]. Progress has also been made in anti-HIV and anti-HCV drug design by applying the method of RNA interference[9]-[10]. To design anti-SARS-CoV drug, one strategy is to search for siRNAs which specifically interfere the gene expression and block the genome replication of SARS-associated coronavirus. In this note we shall make theoretical prediction on the possible target sites of siRNAs in the virus genome.

Upon infection of an appropriate host cell, the viral envelope is fused with cell membrane and the viral plus sense RNA enters the host cell. Then the 5' most ORF of the viral genome is translated into several nonstructural proteins including an RNA-dependent RNA polymerase and an ATPase helicase. These proteins in turn are responsible for replicating the viral genome as well as generating nested transcripts that are used in the synthesis of the viral proteins. The transcriptionally active, subgenomic-size minus strands are also discovered [2]. The siRNA-mediated RNA interference has strong specificity, and may play certain roles in affecting the process of virus expression and proliferation.

RNA secondary structure is composed of double-stranded region of stacked base pairs (stem) and single-stranded region (loop). RNA structure predictions comprise base-paired and non-base-paired regions in various types of loop and junction arrangements (including hairpin loop, bulge loop, interior loop and junctions or multi-loops[11]). Only 21~25 nt long (or more) non-base-paired regions can be served as the target sites of siRNA. They are called free segments. The long non-base-paired region containing one or several short stems (total length of stems 1~3 base pairs) is also considered in our statistics. The latter is called quasi-free segments.

By using program RNAstructure (version 3.7)[12] we folded the RNA sequence of viral genome in a window of 3000 nucleotides, and shifted the window 1500 nucleotides each step along the sequence, so that each site in virus genome has participated in different folds more than 10 times. We selected 21~25 base-long free and quasi-free segments as the candidates for target

sites of RNA interference when these segments frequently occurred in non-base-paired regions based on the above calculation. A given RNA sequence segment may have different configurations of secondary structure with lower free energy, some containing short stems (quasi-free) but some not (free). The total frequency of a segment occurring in non-base-paired region of different folds is called appearance rate. If each quasi-free case is multiplied by a reduced factor in numeration, namely, by 0.9 for 1 base pair, 0.8 for 2 base pair, and 0.7 for 3 base pairs (base pairs may be continuous in structure or disconnected) then the total number of folds is called reduced appearance rate.

The antisense oligonucleotide (AO) complementary to a specific sub-sequence of an RNA target has been extensively investigated. AO efficacy is affected by many factors. Apart from the binding energy between AO and RNA, which describes the AO accessibility to the RNA, the sequence motif is another important factor. The correlation of 9 sequence motifs with AO efficacy was deduced empirically in [13][14]. If the target sequence contains CCAC, TCCC, ACTC, GCCA and CTCT, then it will make a positive score. If the target sequence contains GGGG, ACTG, TAA and AAA, then it will make a negative score.

On the other hand, experiment shows that 2 nt 3' overhangs in siRNA duplex has played an important role in its stabilization [8]. That means AA in 5' end of the sequence segment is favorable for its target.

In SARS-CoV genome we have found several tens' long segments (length >20 nt) matching

with those of human beings. To guarantee the safety of the designed drug, we make alignment of free and quasi-free segments of high appearance rate with human genome and delete the matching ones (more than 18 exactly matching bases) in siRNA target candidates.

By the use of RNA sequence data of SARS-CoV, Isolate Tor2, twenty seven optimal 20~25 base-long siRNA targets are selected from 60000 candidates in both strands. They are listed in Table 1 and 2 for minus-strand and plus-strand respectively. Each segment is scored. The main term of score is the value of reduced appearance rate (column 5 of Table 1 and 2). The sum of AO efficacies (multiplied by 10) in a segment is also listed for reference (column 6). The enhancing factor of AA occurred in 5' end is indicated in column 7 by notation +. The results of multiple sequence alignment of 19 complete SARS coronavirus genome give the mutational sites between different strains [15]. The last term of score is related to mutational sites. Each point mutation in siRNA target sequence contributes -1 in score (column 8). Though the relative importance of these terms cannot be quantitatively estimated at present we expect that the main contribution to the score comes from the reduced appearance rate (column 5).

Generally, in the proliferation of plus-sense RNA viruses the concentration of plus-strand is much higher than that of minus-strand. For example, they may differ by 100 times in TMV (tobacco mosaic virus) [16]. If the concentration of minus-strand in SARS-CoV is lower，then the RNA interference targeted at virus minus-strand will be more effective . We suggest that the latter point should be checked by experiments immediately since it is important for designing an effective siRNA duplex.

The above approach is of broad interest to other anti-virus drug design.

Acknowledgement Authors are grateful to Dr. J.C. Luo for his sending SARS- related material. The work was supported by National Science Foundation of China, Project 90103030.

References

1. Rota, P. A., Oberste, M. S., Monroe, S. S. et al (2003) Characterization of a Novel Coronavirus Associated with Severe Acute Respiratory Syndrome, *Science* **300** (5624) 1394-1399.

2. Marra, M. A., Jones, S. J., Astell, C. R. et al. (2003) The Genome Sequence of the SARS-Associated Coronavirus. *Science* **300** (5624) 1399-1404.

3 Qin, E., Zhu, Q. Y., Yu, M. et al. (2003) A complete sequence and comparative analysis of a SARS-associated virus (Isolate BJ01). *Chinese Science Bulletin*. **48** (10), 941-948.

4. Fire, A, Xu, S, Montgomery, M.K., et al (1998) Potent and specific genetic interence by double-stranded RNA in Caenorhabditis elegans . *Nature* **391**(6669), 806-811.

5. Bernstein, E., Denli, A.M., Hannon, G.J. (2001) The rest is silence. *RNA* **7**, 1509-1521.

6. Plasterk, R.H.A. RNA splicing: The genome's immune system (2002). *Science* **296**, 1263-1265.

7. Zamore, P.D. (2002). Ancient pathways programmed by small RNAs. *Science* **296**, 1265-1269.

8. Elbashir, S.M., Harborth, J., Lendeckel, W. et al.(2001) Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. *Nature* **411** (6836), 494.

9. Jacque JM., Triques K., Stevenson M. (2002) Modulation of HIV-1 replication by RNA interference. *Nature* **418** (6896), 435.

10. Wilson J A, Jayasena S., Khvorova A.. et al.(2003) RNA interference blocks gene expression and RNA synthesis from hepatitis C replicons propagated in human liver cells. *Proc. Natl. Acad. Sci. USA* **100** (5), 2783.

11. Mount D.M.. *Bioinformatics*.(2001) Cold Spring Harbor Laboratory Press. Ch. 5.

12. Mathews D.H., Sabina J., Zuker M, Turner D.H. (1999) Expanded Sequence Dependence of Thermodynamic Parameters Improves Prediction of RNA Secondary Structure. *J Mol Biol*, **288**, 911-940.

13. Matveeva, O.V., Tsodkov A.D., Giddings, M. et al.(2000) Identification of sequence motifs in oligonucleotides whose presence is correlated with antisense activity. *Nucleic Acids Res*. **28**, 2862-2865.

14. Chalk, A.M, Sonnhammer E.L.L.(2002) Computational antisense oligo prediction with a neural network model. *Bioinformatics* **18**, 1567-1575.

15. Results of multiple sequence alignment of 19 complete SARS coronavirus genome, at http://antisars.cbi.pku.edu.cn/news/news.jsp?id=114.

16. Ishikawa M., Meshi, T. Ohno, T, Okada Y. et al. (1991) Specific Cessation of Minus Strand RNA Accumulation at an Early Stage of TMV Infection. *J.Virol*. **65**, 861-868.

Table 1 siRNA target sequence in minus-strand SARS-CoV

target sequence 5'-3'	position	length	appear. rate	reduced appear. rate	AO eff (×10)	AA 5' end	mut. site score
AAUUUCUUGAAUUACCGCGACUAC	1041-1064	24	8	8		+	-1
AAUUGAUCUAAGAGUAAAAAAU	4446-4467	22	9	6.3	-4.4	+	-1
UGUCAACACAAAGUAAUCACC	12902-12922	21	7	5.6	-1.2		0
CUCCCUUCGAAUUGUUAUAGU	17025-17045	21	13	10.4	2.6		0
AAGACAUCAAAAACAAAAGUG	20292-20312	21	7	4.9	-5.0	+	0
ACACCAUCUAAAGCUACACCC	21671-21691	21	11	8.8	-1.2		0
AAUUAGAUAAGAGUACACCAA	22788-22808	21	10	10		+	0
UAGCAUCACGACCACACACAC	24177-24197	21	12	12	1.7		0
CUAGUAUAAAAGAAGAAUCGG	25280-25300	21	11	10.6	-2.2		0
AAUUUUAAUUCCUUUAUACUU	25327-25347	21	10	8.6		+	0
CCUUCAAUAACUAAAUUUUCA	28430-28450	21	10	10	-1.4		0
AGCUACACAGAUUUUAAAGUU	29662-29682	21	8	7.8	-1.2		0

Table 2 siRNA target sequence in plus-strand SARS-CoV

target sequence 5'-3'	position	length	appear. rate	reduced appear. rate	AO eff. (×10)	AA 5' end	mut. site score
AAACAAUAAUAAAUUUUACUG	128-148	21	8	8	-4.2	+	0
UUGUUUCUGUUACCUUCUCUU	11107-11127	21	12	12	1.8		0
AAUCAUUAUUAAAGACUGUA?	13921-13941	21	14	9.8	-3.0	+	0
UACCCAGAUCCAUCAAGAAUAUU	15861-15883	23	10	10			0
UAUCUCACCUUAUAAUUCACA	17699-17719	21	9	8.1			0
AAUUGCCUUUCUUUUACUAUU	19291-19311	21	8	7.2		+	0
GACUACAAAAGAGAAGCCCCA	19809-19829	21	8	6.4	-2.0		0
AACCUUCUACCCAAAACUACAA	20567-20588	22	11	11	-2.0	+	0
UUUUCUUAUUAUUUCUUACUC	21499-21519	21	12	11.8	0.9		0
AUUAUUAACAAUUCUACUAAU	21837-21857	21	13	10.7			0
AAACAACUUAGCUCUAAUUUU	24327-24347	21	13	9.1	-3.0	+	0
AUUAACAACACAGUUUAUGAU	24834-24854	21	12	10.8			0
AAUAUGAGCAAUAUAUUAAAU	25051-25071	21	8	6.4	-1.2	+	0
AACGAACUAACUAUUAUUAUU	26347-26367	21	9	9		+	0
AACGAACAUGAAAAUUAUUCUCUU	27266-27289	24	10	10		+	0