# siRNA Designs to The Crucial Proteins of SARS Coronavirus*

ZHANG Yong[1)**], XU Jing-Yi[2)**], DENG Wei[1)], ZHANG Nan[1)], CAI Lun[2)], ZHAO Yi[2)], BU Dong-Bo[2)], CHEN Run-Sheng[1,2,3)***]

([1)] *Bioinformatics Laboratory, Institute of Biophysics, The Chinese Academy of Sciences, Beijing* 100101, China;
[2)] *Bioinformatics Research Group, Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, The Chinese Academy of Sciences, Beijing* 100080, China;
[3)] *Beijing Genomics Institute, The Chinese Academy of Sciences, Beijing* 101300, China)

**Abstract** RNA interference (RNAi) is the process of sequence-specific, post-transcriptional gene silencing in animals and plants, initiated by double-stranded RNA (dsRNA) that is homologous in sequence to the silenced gene. In mammal cells, small interfering RNA (siRNA) duplexes can induce RNAi potently, which may provide a new approach to the therapeutics of certain diseases. Focusing on the five genes which coding five crucial proteins of SARS coronavirus(SARS-CoV) respectively, 348 siRNA candidate targets were obtained following bioinformatic methods. In theory, potent siRNA duplexes specifically suppress expression of their corresponding SARS-CoV target gene, while have no influence on the normal expression of human gene. It would lay a foundation for the further experimental researches on the siRNA-like drug design for the SARS-CoV.

**Key words** SARS coronavirus (SARS-CoV), RNA interference (RNAi), small interfering RNA (siRNA), bioinformatics

Severe acute respiratory syndrome (SARS) has been becoming a world-wide threat against people's health. Up to May 17th, 2003, the disease has rapidly spread to 32 countries and regions. With the multilateral efforts, it has been confirmed that the pathogen of SARS is a kind of previously unknown coronavirus, now named SARS coronavirus (SARS-CoV). At present, no effective therapy is available. Scientists all over the world are exerting themselves to explore kinds of possible approaches to antagonize SARS[1,2].

RNA interference (RNAi), a research hotspot in recent years, is the process of sequence-specific, post-transcriptional gene silencing in animals and plants, and scientists have clearly known its mechanism on the molecular level[3]. However, the length of the transfected double-stranded RNA (dsRNA), if exceeding 30 nt, will result in violent antivirus reactions in most mammal cells, i. e. unspecifically suppressing expression of certain gene[4], which is opposite to the functions of dsRNA in other lower species. Further experiments indicate that transient transfected siRNA duplexes can avoid the antivirus reactions by using small interfering RNA (siRNA) duplexes described above[4,5] and the most effective siRNAs turn out to be 21 nt dsRNAs with 2 nt 3′ overhangs which need not match with the target sequence. The siRNA triggered RNAi may provide a new approach to the therapeutics of certain diseases. Recent experimental results demonstrated the utility of RNAi to modulate the HIV-1 (human immunodeficiency virus) and HCV (Hepatitis C virus) replication cycle in cultivated human cells[6,7]. Noticeably, SARS-CoV is positive-stranded RNA virus and intermediate RNA is necessary in replication cycle, which shares the same features to the HIV and HCV. This similarity indicates the attempt is significant, to inhibit SARS-CoV duplication with certain siRNA duplexes triggered RNAi as HIV and HCV. The accomplishment of complete genome sequence of SARS-CoV[8~10] makes the design of siRNA aiming at SARS-CoV possible. Focusing on the five genes which coding five crucial proteins of SARS-CoV, we design 348 siRNA candidate targets following bioinformatic methods, which lays a foundation for the further experimental researches.

## 1 Data and methods

The template sequence of SARS-CoV we used was downloaded from GenBank (ACCESSION: NC_004718; May 23, 2003 updated). In the annotated 13 protein ORFs, genes coding 5 crucial proteins cover 91.2% length of complete genome, and the five proteins are respectively spike protein (S), membrane protein

* This work was supported by grants from The National Knowledge Innovation Program of The Chinese Academy of Sciences (KSCX2-2-07) and The National "863" High-tech Program (2002AA231031).
** These authors contributed equally to this work.
*** Corresponding author.
Tel: 86-10-64888543, E-mail: crs@sun5.ibp.ac.cn
Received: May 29, 2003 Accepted: May 30, 2003

(M), envelope protein (E), nucleocapsid protein (N) and RNA polymerase. Thereinto, RNA polymerase is a complex protein approximately consisting of 7 000 amino acid residues. After autocatalysis, it could hydrolyze into a series of proteins, including the viral main proteinase ($3CL^{pro}$). Here, the siRNA candidate targets we designed only focus on the genes coding the five crucial proteins above.

SARS-CoV is positive-stranded RNA virus, which has the same nucleotide sequences to the mRNA corresponding to certain proteins. Based on the progress of recent researches[4, 11] and the successful experience recommended by Ambion Corp's researchers, we design siRNA candidate targets in SARS-CoV genome, according to the conditions as following: a. 21 nt in length; b. Preferred 5′ end structure is "AA"; c. G/C content within 30% and 50%; d. As a result of BLAST[12] alignment with the whole human genome sequences (download from GenBank; Aug 1, 2002 updated), exactly matched bases is not over 16 nt.

Based on the targets satisfied with the four conditions above, one can synthesize potent siRNA duplexes in effect.

## 2 Results and discussion

The eventually numbers of eligible siRNA candidate targets are listed respectively according to certain corresponding proteins (Table 1).

**Table 1 Numbers of eligible siRNA candidate targets we selected based on theoretical design**

| Corresponding protein[1)] | Number of candidate targets |
|---|---|
| RNA polymerase | 264 |
| S | 46 |
| E | 3 |
| M | 6 |
| N | 29 |

[1)] Corresponding proteins: proteins coded by the gene where target sequences position.

The sequences of candidate targets aimed at genes coding the four proteins except RNA polymerase are shown in Table 2. Because of the large number of candidate targets related to RNA polymerase, we provide the sequence file online (http://www.bioinfo.org.cn/SARS/RNAi/) or any one who interested in it could contact with the authors directly.

**Table 2 Sequences of siRNA candidate targets focusing on gene coding four structural protein of SARS-CoV**

| Corresponding proteins[1)] | Sequences of siRNA candidate target (5′-3′) | Position of 5′ end[2)] |
|---|---|---|
| S | AAGCTCCTAATTACACTCAAC | 21 568 |
| S | AATCATACGTTTGGCAACCCT | 21 708 |
| S | AACAACAAGTCACAGTCGGTG | 21 813 |
| S | AACAAGTCACAGTCGGTGATT | 21 816 |
| S | AAGTCACAGTCGGTGATTATT | 21 819 |
| S | AATGTTGTTATACGAGCATGT | 21 855 |
| S | AATTGCACTTTCGAGTACATA | 21 963 |
| S | AAGGGCTATCAACCTATAGAT | 22 083 |
| S | AACCTTTAAGTGCTATGGCGT | 22 577 |
| S | AAGACAAATAGCGCCAGGACA | 22 673 |
| S | AAATAGCGCCAGGACAAACTG | 22 678 |
| S | AAGAACCAGTGTGTCAATTTT | 23 052 |
| S | AATGGACTCACTGGTACTGGT | 23 079 |
| S | AACAATTTGGCCGTGATGTTT | 23 137 |
| S | AATTTGGCCGTGATGTTTCTG | 23 140 |
| S | AAGTGTAATTACACCTGGAAC | 23 234 |
| S | AAGCAGGCTGTCTTATAGGAG | 23 386 |
| S | AATAACACCATTGCTATACCT | 23 562 |
| S | AACACCATTGCTATACCTACT | 23 565 |
| S | AAAACCTCCGTAGATTGTAAT | 23 634 |
| S | AACCTCCGTAGATTGTAATAT | 23 636 |
| S | AATATGTACATCTGCGGAGAT | 23 652 |
| S | AACTAAATCGTGCACTCTCAG | 23 722 |
| S | AAATCGTGCACTCTCAGGTAT | 23 726 |
| S | AAGTGTTCGCTCAAGTCAAAC | 23 776 |
| S | AAATATTACCTGACCCTCTAA | 23 848 |
| S | AAGCAATATGGCGAATGCCTA | 23 940 |
| S | AATATGGCGAATGCCTAGGTG | 23 944 |
| S | AACCAAAAACAAATCGCCAAC | 24 192 |
| S | AAAAACAAATCGCCAACCAAT | 24 196 |
| S | AACCAATTTAACAAGGCGATT | 24 210 |
| S | AACAAGGCGATTAGTCAAATT | 24 219 |
| S | AAGGCGATTAGTCAAATTCAA | 24 222 |
| S | AACATCAACTGCATTGGGCAA | 24 257 |
| S | AAGACGTTGTTAACCAGAATG | 24 283 |
| S | AATGCTCAAGCATTAAACACA | 24 300 |
| S | AAATGATATCCTTTCGCGACT | 24 368 |
| S | AACAACTAATCAGGGCTGCTG | 24 466 |
| S | AAAGGGCTACCACCTTATGTC | 24 569 |
| S | AAATTGTGATGTCGTTATTGG | 24 809 |
| S | AACACAGTTTATGATCCTCTG | 24 840 |
| S | AACGCTTCTGTCGTCAACATT | 24 954 |
| S | AAAAAGAAATTGACCGCCTCA | 24 976 |
| S | AAAGAAATTGACCGCCTCAAT | 24 978 |
| S | AAATTGACCGCCTCAATGAGG | 24 982 |
| S | AAATGGCCTTGGTATGTTTGG | 25 068 |
| E | AAGAAACAGGTACGTTAATAG | 26 136 |
| E | AAAACCAACGGTTTACGTCTA | 26 272 |
| E | AACCAACGGTTTACGTCTACT | 26 274 |
| M | AACGGTACTATTACCGTTGAG | 26 407 |
| M | AACTCCTGGAACAATGGAACC | 26 438 |
| M | AATGGAACCTAGTAATAGGTT | 26 450 |
| M | AACCTAGTAATAGGTTTCCTA | 26 455 |

Continued

| Corresponding proteins[1)] | Sequences of siRNA candidate target (5′-3′) | Position of 5′ end[2)] |
|---|---|---|
| M | AATGGCTTGTATTGTAGGCTT | 26 643 |
| M | AATTGTGACCAGACCGCTCAT | 26 775 |
| N | AATGGACCCCAATCAAACCAA | 28 129 |
| N | AATCAAACCAACGTAGTGCCC | 28 139 |
| N | AACCAGAATGGAGGACGCAAT | 28 201 |
| N | AATAATACTGCGTCTTGGTTC | 28 261 |
| N | AACACCAATAGTGGTCCAGAT | 28 345 |
| N | AAATTGGCTACTACCGAAGAG | 28 370 |
| N | AACAAAGAAGGCATCGTATGG | 28 498 |
| N | AAAGAAGGCATCGTATGGGTT | 28 501 |
| N | AAGGCATCGTATGGGTTGCAA | 28 505 |
| N | AAATTCTCCTGCTCGAATGGC | 28 734 |
| N | AACAAGGCCAAACTGTCACTA | 28 844 |
| N | AAGGCCAAACTGTCACTAAGA | 28 847 |
| N | AAAAAGCCTCGCCAAAAACGT | 28 888 |
| N | AAAGCCTCGCCAAAAACGTAC | 28 890 |
| N | AAACGTACTGCCACAAAACAG | 28 903 |
| N | AAAACAGTACAACGTCACTCA | 28 917 |
| N | AACAGTACAACGTCACTCAAG | 28 919 |
| N | AACGTCACTCAAGCATTTGGG | 28 927 |
| N | AAACCCAAGGAAATTTCGGGG | 28 964 |
| N | AAGGAAATTTCGGGGACCAAG | 28 970 |
| N | AAATTTCGGGGACCAAGACCT | 28 974 |
| N | AAGACCTAATCAGACAAGGAA | 28 988 |
| N | AAACATTGGCCGCAAATTGCA | 29 017 |
| N | AATGTCACGCATTGGCATGGA | 29 070 |
| N | AAGTCACACCTTCGGGAACAT | 29 090 |
| N | AACGTCATACTGCTGAACAAG | 29 167 |
| N | AACAAGCACATTGACGCATAC | 29 182 |
| N | AAGCACATTGACGCATACAAA | 29 185 |
| N | AAAGACTGATGAAGCTCAGCC | 29 244 |

[1)] Corresponding proteins: proteins coded by the gene where target sequences position; [2)] Position of 5′ end: the 5′ end position of siRNA candidate targets sequence relative to SARS-CoV genome.

There are only 19 nt of the RNAi duplexes sequence complementary to its corresponding target. The group of sequences including siRNA candidate target, sense and antisense strands of the siRNA duplexes corresponding to the target is illustrated as following (orientated from 5′ to 3′; "tt" at the 3′ end are unmatched nucleotides).

Target: AAGCTCCTAATTACACTCAAC
Sense strand: GCUCCUAAUUACACUCAACtt
Antisense strand: GUUGAGUGUAAUUAGGAGCtt

Based on our theoretical design, the selected potent siRNA duplexes according to the candidate target can specifically suppress expression of corresponding SARS-CoV target gene. At the same time, the fourth condition described in **Data and methods** guarantees that the selected potent siRNA duplexes have no influence on the normal expression of gene in human cells because the sequence specificity of siRNA is very stringent, i. e. single base pair mismatches between the siRNA and its target mRNA dramatically reduced silencing[4,11].

The existing knowledge of siRNA has not guaranteed the efficiency of all the theoretically designed siRNA candidate targets, so further selections in experimental manner are indispensable. To avoid the impotence of siRNA targets arose by mutations in virus genome, some experiments are designed focusing on gene silence of certain mRNA in host cells. For example, researchers successfully prevent HIV′s invasion by suppressing expression of CD4 in host cells[13]. With the increasing further knowledge of nosogenesis of SARS-CoV, RNAi induced by siRNA, which focuses on certain host gene that would be necessary for the replication cycle of SARS-CoV, may be a possible alternative to SARS therapy.

## References


1 Kontoyiannis D P, Pasqualini R, Arap W. Aminopeptidase N inhibitors and SARS. Lancet, 2003, **361** (9368): 1558
2 Anand K, Ziebuhr J, Wadhwani P, *et al*. Coronavirus main proteinase ($3CL^{pro}$) structure: Basis for design of anti-SARS drugs. http://www.sciencemag.org/cgi/rapidpdf/1085658vl, 2003
3 Hammond S M, Caudy A A, Hannon G J. Post-transcriptional gene silencing by double-stranded RNA. Nat Rev Genet, 2001, **2** (2): 110~119
4 Elbashir S M, Harborth J, Lendeckel W, *et al*. Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. Nature, 2001, **411** (6836): 494~498
5 Caplen N J, Parrish S, Imani F, *et al*. Specific inhibition of gene expression by small double-stranded RNAs in invertebrate and vertebrate systems. Proc Natl Acad Sci USA, 2001, **98** (17): 9742~9747
6 Jacque J M, Triques K, Stevenson M. Modulation of HIV-1 replication by RNA interference. Nature, 2002, **418** (6896): 435~438
7 Wilson J A, Jayasena S, Khvorova A, *et al*. RNA interference blocks gene expression and RNA synthesis from hepatitis C replicons propagated in human liver cells. Proc Natl Acad Sci USA, 2003, **100** (5): 2783~2788
8 Marra M A, Jones S J, Astell C R, *et al*. The Genome sequence of the SARS-associated coronavirus. Science, 2003, **300** (5624): 1399~1404
9 Rota P A, Oberste M S, Monroe S S, *et al*. Characterization of a novel coronavirus associated with severe acute respiratory syndrome. Science, 2003, **300** (5624): 1394~1399
10 Qin E, Zhu Q Y, Yu M, *et al*. A complete sequence and comparative analysis of a SARS-associated virus (Isolate BJ01). Chinese Science Bulletin, 2003, **48** (10): 941~948
11 Elbashir S M, Martinez J, Patkaniowska A, *et al*. Functional anatomy of siRNAs for mediating efficient RNAi in Drosophila melanogaster embryo lysate. Embo J, 2001, **20** (23): 6877~6888
12 Altschul S F, Madden T L, Schaffer A A, *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res, 1997, **25** (17): 3389~3402
13 Novina C D, Murray M F, Dykxhoorn D M, *et al*. siRNA-directed inhibition of HIV-1 infection. Nat Med, 2002, **8** (7): 681~686

# 针对 SARS 冠状病毒重要蛋白的 siRNA 设计*

张 勇[1)**] 徐静怡[2)**] 邓 巍[1)] 张 楠[1)] 蔡 伦[2)] 赵 义[2)] 卜东波[2)] 陈润生[1,2,3)***]

([1)]中国科学院生物物理研究所生物信息学实验室，北京 100101;

[2)]中国科学院计算技术研究所智能信息处理重点实验室生物信息学研究组，北京 100080;

[3)]中国科学院基因组研究所，北京 101300)

**摘要** RNA 干涉 (RNA interference, RNAi) 是一种特异性地导致转录后基因沉默的现象，在哺乳动物细胞中小分子干扰 RNA 双链体 (small interfering RNA duplexes, siRNA duplexes) 可以有效地诱导 RNAi 现象，为一些疾病的治疗开辟了新的途径. 针对 SARS 冠状病毒 (SARS coronavirus, SARS-CoV) 中编码 5 个主要蛋白质的基因，用生物信息学的方法设计了 348 条候选 siRNA 靶标. 在理论上，相应的 siRNA 双链体能特异地抑制 SARS-CoV 靶基因的表达，同时不会影响人体细胞基因的正常表达，这为进一步 siRNA 类药物的实验研究提供了理论基础.

**关键词** SARS 冠状病毒，RNA 干涉，siRNA，生物信息学

**学科分类号** Q61

---

* 中国科学院知识创新工程项目 (KSCX2-2-07) 和国家“863”计划基金 (2002AA231031) 资助.

** 对本文有同等贡献.

*** 通讯联系人. Tel: 010-64888543, E-mail: crs@sun5.ibp.ac.cn

收稿日期: 2003-05-29, 接受日期: 2003-05-30