针对 H1N1 病毒的多特征 siRNA 设计

刘元宁[1],常亚萍[1],李 誌[2],张 浩[1],田明尧[3]

(1. 吉林大学 计算机科学与技术学院,长春 130012;2. 长春理工大学 应用技术学院,长春 130022; 3. 中国人民解放军军事科学院 生物研究所,长春 130062)

摘 要:针对甲型流感病毒 H1N1 基因,从 RNAi 的角度出发,采用多特征融合的方法,进行 siRNA 预测。对 2009 年的 46 株病毒序列的 PA 片段进行分析,从经过序列分析所获得的众多靶系列中,采用结构分析手段对靶序列进行筛选,获得较易干扰的靶序列及设计出相应的 siRNA。研究发现,2009 年爆发的 H1N1 病毒,序列保守性高,靶序列一致性高,结构保守性高。该方法可以有效选择可能的靶序列,并在此基础上进一步筛选,以获得少量较易干扰的靶序列,该方法为复杂序列 siRNA 的设计提供了新思路,对 siRNA 的优化设计有指导意义,有助于利用 RNAi 进行 H1N1 治疗的后续研究。

关键词:生物信息学;RNAi;siRNA;二级结构

中图分类号:TP31 **文献标志码**:A **文章编号**:1671-5497(2010)03-0776-06

siRNA design for H1N1 based on multi-characters

LIU Yuan-ning[1], CHANG Ya-ping[1], LI Zhi[2], ZHANG Hao[1], TIAN Ming-yao[3]

(1. *School of Computer Science and Technology, Jilin University, Changchun* 130012, *China*; 2. *School of Applied Technology, Changchun University of Science and Technology, Changchun* 130022, *China*; 3. *Institute of Biological Research, Academy of Military Sciences of Chinese People's Liberation Army, Changchun* 130062, *China*)

Abstract: Aiming at H1N1 gene and starting from the point of RNAi, siRNA prediction is conducted by means of multi-characters analyses, such as sequence and structure. First 46 viral sequence's PA fragments in year 2009 are analyzed. Then the siRNA target genes with strong ability of interference are selected on the basis of the secondary structure of siRNA target sequence. Our research reveals that the outbreaking H1N1 virus in 2009 is characterized by steady heredity, high sequence conservativeness, uniform siRNA target gene sequence, and high conservative structure. This method can be employed to choose the possible siRNA and obtain less but more valuable siRNA target genes by further sifting. It provides a new idea for the design of complex sequence siRNA, and it is of instructional significance for optimal design of siRNA. The research is helpful to the study of H1N1 treatment by RNAi.

Key words: bioinformatics; RNA interfering; small interfering RNA; secondary structure

收稿日期:2009-09-29.

基金项目:国家自然科学基金项目(60673099, 60873146, 60971089).

作者简介:刘元宁(1962-),男,教授,博士生导师. 研究方向:生物信息学. E-mail:lyn@jlu.edu.cn

通信作者:张浩(1971-),男,副教授,博士. 研究方向:生物信息学. E-mail:lyn@jlu.edu.cn

随着人类基因组学、转录组学、功能组学、RNA 组学和生物信息学等学科的飞速发展,RNA 干扰(RNA interfering,RNAi)技术的应用更加广泛和深入[1-4]。目前 siRNA 的设计遵循一些基本的经验,但根据这些经验针对一条基因往往得出成百上千条 siRNA,而且许多知名的 siRNA 设计网站给出的候选 siRNA 也难以统一,这种现状迫使研究者致力于研究如何高效快速筛选有效的 siRNA。

本文采用序列特征与结构特征相结合的方法,对 H1N1 病毒的 PA 片段进行 siRNA。首先对 PA 片断进行序列特征分析获得可能的靶序列,然后按候选靶序列的结构进行筛选,获得较易干扰的靶序列,并设计出相应的 siRNA。目前针对 H1N1 病毒 PA 片段的 siRNA 设计未见报道。

1　RNAi 的机制

RNA 干扰是一种有效的基因沉默过程,宿主细胞中导入与内源性 mRNA 编码区同源的双链 RNA(dsRNA)时,可诱发同源 mRNA 发生降解而导致基因表达沉默。RNAi 技术在多个领域被广泛应用,包括抗病毒治疗、功能基因组学、抗肿瘤治疗、研究药物作用的靶位、遗传学、干细胞生物学以及信号传递等方面。RNAi 的作用过程可以分为启动阶段和效应阶段[5-6]。启动阶段指 RNase Ⅲ将由外源导入或病毒感染等方式引入的 dsRNA 被 Dicer 酶特异识别,以一种 ATP 依赖的方式加工成 21-23ntsiRNA,且每条链的 3′端都带有 2 个悬垂的碱基(本文是 dTdT)。效应阶段指 siRNA 双链解开,其中反义 siRNA 和相应的蛋白结合并形成 RISC。siRNA 与 RISC 结合,并通过驾驭 RISC 到相应的 mRNA 位点,随即 RISC 执行 RNA 干扰的效应功能,酶切降解 mRNA,使转录的基因表达终止。siRNA 是严格按碱基配对的法则与目标 mRNA 结合的,对互补序列特异性要求相当高。RNA 干扰的特异性不仅是 siRNA 应用与基因研究时能够准确剖析具体基因功能的前提条件,也是使其用于治疗疾病时药效强大、不良反应小的保障。

2　siRNA 的设计

传统的 siRNA 设计都是分析序列特征,本文通过对 siRNA 序列特征和靶序列特征的分析发现,影响 siRNA 活性的因素不仅包括序列特征(长度、GC 含量、关键位的碱基构成),还包括结构特征(茎区特征、环区特征)。单一特征并不能完全体现 siRNA 活性变化,只有把众多特征融合在一起才能完全体现 siRNA 的特点。只有考虑多特征融合的方法进行 siRNA 设计,才能更准确地寻找对基因有影响的 siRNA。

2.1　siRNA 的特点

长度约在 21 nt 左右、依赖 Dicer 酶加工是 Dicer 的产物。siRNA 是人工体外合成的,通过转染进入人体内,是 RNA 干涉的中间产物。siRNA 是双链 RNA,可作用于 mRNA 的任何部位,也能够导致靶标基因的降解,即为转录水平后调控。

RNAi 作用的成功与否关键在于 siRNA 序列的结构。许多因素会影响 siRNA 的活性,主要包括 siRNA 的热力学稳定性、siRNA 的序列特征以及靶 mRNA 的结构特征等。

2.2　siRNA 的热力学稳定性

RISC 负责结合 siRNA,并使之解旋。一般来说,G≡C 碱基对比 A＝U 碱基对具有更高的热力学稳定性[6]。

2.3　siRNA 的序列特征

根据甲型流感病毒 H1N1 的 PA 序列特点和 siRNA 的特征,本文设计的 siRNA 的长度为 21 nt。siRNA 是严格按碱基配对的法则与目标 mRNA 结合,对互补序列特异性要求相当高。很多研究都证实与互补 mRNA 相差一个碱基序列的 siRNA,抑制目标基因表达的效能大大下降,相差两三个碱基序列的 siRNA 的作用完全丧失。

2.4　靶 mRNA 的结构分析

RNA 结构的保守性大于序列的保守性,mRNA 的二级结构在基因表达调控方面也扮演着重要的角色。在转录时,某些位置的 mRNA 上二级结构的形成参与了对转录终止的控制,在翻译水平上,mRNA 通过自身折叠来调节核糖体在其上的翻译速度,由此在新生肽链共翻译折叠时,间接影响了蛋白质二级结构单元以及蛋白质最终构象的形成,最终导致生物特性的变化。

从结构的角度分析,siRNA 的序列对原有结构的改变越大,它所起的作用也越明显。即 siRNA 序列对原有的结构关系破坏越大,干扰能力越强,对二级结构,可以将其分为环区、茎区、茎环区、多分支环区,经过分析,siRNA,对茎区的影响较大,对多分枝环的影响较小。对 mRNA 目标

区域进行二级结构预测,排除二级结构复杂的siRNA的靶序列。因为靶序列的二级结构越复杂,siRNA沉默效率越低。

3 siRNA设计的算法

结合以往的siRNA设计原则和PA基因的特点,本文给出siRNA设计的一般步骤如下:

(1)数据来源:在NCBI的GenBank(http://www.ncbi.nlm.nih.gov/Genbank/index.html)数据库检索H1N1甲型流感病毒中具有全氨基酸序列和编码mRNA核苷酸序列的PA(截止到2009年7月),得到46个样本(见表1)。

表1 病毒的地区及编号

Table 1 Region and number of virus

地区	名称	编号	名称	编号
美国	A/NewYork/3007/2009(H1N1)	CY040047	A/New York/3413/2009(H1N1)	CY041763
	A/NewYork/3008/2009(H1N1)	CY040004	A/New York/3463/2009(H1N1)	CY041795
	A/NewYork/3012/2009(H1N1)	CY040012	A/New York/3468/2009(H1N1)	CY041803
	A/NewYork/3049/2009(H1N1)	CY040028	A/New York/3532/2009(H1N1)	CY041779
	A/NewYork/3099/2009(H1N1)	CY040036	A/Texas/05/2009(H1N1)	FJ966970
	A/NewYork/3166/2009(H1N1)	CY041055	A/Ohio/07/2009(H1N1)	FJ984400
	A/NewYork/3337/2009(H1N1)	CY041787	A/Texas/04/2009(H1N1)	FJ969524
	A/New York/3348/2009(H1N1)	CY041835	A/Texas/09/2009(H1N1)	GQ117029
	A/New York/3351/2009(H1N1)	CY041819	A/California/04/2009(H1N1)	FJ969515
	A/New York/3352/2009(H1N1)	CY041827	A/California/04/2009(H1N1)	FJ966081
	A/New York/3354/2009(H1N1)	CY041771	A/California/07/2009(H1N1)	FJ969529
	A/New York/3408/2009(H1N1)	CY041811	A/California/08/2009(H1N1)	FJ984368
德国	A/Bayern/62/2009(H1N1)	GQ365657	A/Bayern/63/2009(H1N1)	GQ166210
俄罗斯	A/Moscow/IIV02/2009(H1N1)	GQ247729	A/Moscow/IIV03/2009(H1N1)	GQ330650
加拿大	A/Toronto/3141/2009(H1N1)	GQ373261		
墨西哥	A/Mexico/4108/2009(H1N1)	GQ149686	A/Mexico/InDRE4482/2009(H1N1)	GQ149676
意大利	A/Italy/85/2009(H1N1)	GQ351289		
日本	A/Kobe/1/2009(H1N1)	GQ222036	A/Utsunomiya/1/2009(H1N1)	GQ334360
	A/Narita/1/2009(H1N1)	GQ169305	A/Utsunomiya/2/2009(H1N1)	GQ365460
	A/Hyogo/1/2009(H1N1)	GQ222034	A/Amagasaki/1/2009(H1N1)	GQ222032
	A/Osaka/2/2009(H1N1)	GQ222038	A/Amagasaki/2/2009(H1N1)	GQ222033
	A/Sapporo/1/2009(H1N1)	GQ365451	A/Fukuoka-C/2/2009(H1N1)	GQ334335
	A/Shizuoka/759/2009(H1N1)	GQ334351	A/Fukuoka-C/3/2009(H1N1)	GQ334343
中国	A/Sichuan/1/2009(H1N1	GQ166226	A/GuangzhouSB/01/2009(H1N1	GQ223443

(2)对所有序列进行同源性分析,以获得相应的保守区。

(3)从基因编码序列开始,避开5′端或3′端的UTRs,搜索AA(N19)TT序列。

(4)根据siRNA的靶序列的碱基组成进一步筛定:选择G/C含量在40%~50%左右的mRNA区域,避免出现连续3 nt以上的G或C,因为多聚G或C系列能产生堆积形成类聚物而影响siRNA沉默机制。

(5)分析得到的靶序列的二级结构,选出较易干扰的靶序列设计出相应的siRNA,加上3′端悬垂结构,dTdT。

4 实验分析

用序列分析软件对H1N1的PA基因的序列进行保守性分析[7-10],结果见图1。

由图1可知,序列的保守性非常高,不同的序列之间也存在差异。同时也做了一下二级结构的分析,由于同源性比较高,所以只列出以下2个,见图2,发现在二级结构上也存在高度保守的区域。

按照siRNA设计的步骤,根据上述要求设定参数,对这46个H1N1的PA片段进行分析,共获得了9段靶序列,每个基因都有这9段靶序列,但其位置存在差异,一共发现9种不同的位置分布,这种现象是由它们的序列差异性决定的,结果见表2。通过实验发现9段靶序列均在保守区,由此可见,参数设定对PA片断有效。

根据上面获得的靶序列信息,共有9条可能的靶序列,对于生物实验而言其可供选择的序列过多,这样就需要对候选靶序列进行再次筛选,以获得较易干扰的靶序列。本文根据结构特征进行

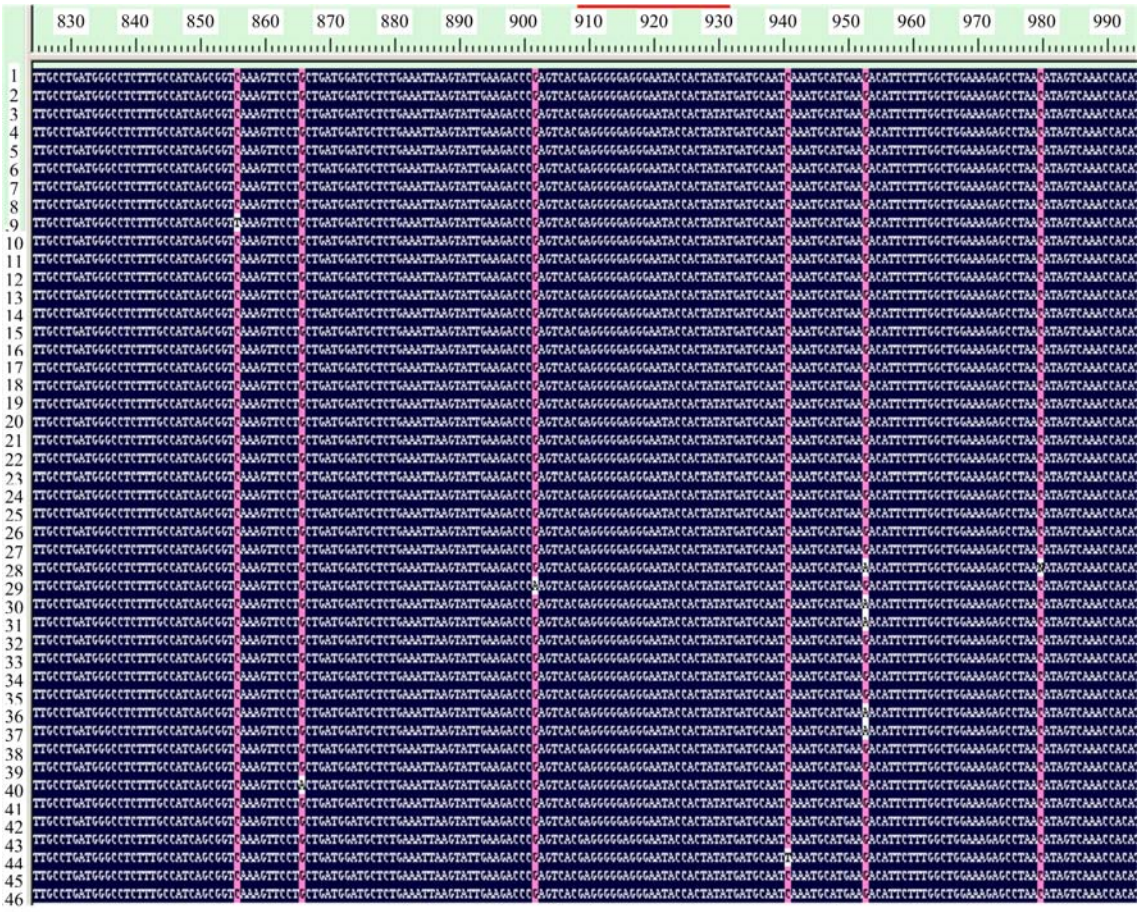

图 1　序列保守性分析

Fig. 1　Conservative analysis of sequence

(a) ABayern622009(H1N1)GQ365657 的二级结构

(b) AMoscow032009(H1N1)GQ375283 的二级结构

图 2　PA 基因结构比较

Fig. 2　Comparison of PA genetic structure

分析,siRNA 的序列对原有结构的改变越大,所起的作用也越明显。根据上文提到的二级结构的判断标准选择出更易干扰的靶序列,结构分析见表 3,筛选出上面的 7、4、5 的靶序列最有可能出现干扰。由此,获得靶序列对应的 siRNA 序列,见表 4。

表 2 靶序列序号差异分析

Table 2 Variance analysis of target sequence

编号	序列号	序列	编号	序列号	序列	编号	序列号	序列
CY040047	1814-1836	AGACATGACCAAGGAATTC		1832-1854	AGACATGACCAAGGAATTC		1827-1849	AGACATGACCAAGGAATTC
CY040028	1815-1837	GACATGACCAAGGAATTCT		1831-1853	GACATGACCAAGGAATTCT		1828-1850	GACATGACCAAGGAATTCT
	339-361	GAGAACCGGTTCATTGAAA		356-378	GAGAACCGGTTCATTGAAA		352-374	GAGAACCGGTTCATTGAAA
	206-228	TGCACTATTGAAGCACCGA		223-245	TGCACTATTGAAGCACCGA		219-241	TGCACTATTGAAGCACCGA
	991-1013	ATCCCAATTACCTCATGGC	GQ223443	1008-1030	ATCCCAATTACCTCATGGC	GQ373261	1004-1026	ATCCCAATTACCTCATGGC
	1538-1560	TGATACTGATGTGGTGAAC		1555-1577	TGATACTGATGTGGTGAAC		1551-1573	TGATACTGATGTGGTGAAC
	2108-2130	TGCATCTTGGTTCAACTCC		2125-2147	TGCATCTTGGTTCAACTCC		2121-2143	TGCATCTTGGTTCAACTCC
	1491-1513	ACAAACCTGTATGGGTTCA		1508-1530	ACAAACCTGTATGGGTTCA		1504-1526	ACAAACCTGTATGGGTTCA
	1076-1098	CATGAAGAGAACAAGCCAA		1093-1115	CATGAAGAGAACAAGCCAA		1089-1111	CATGAAGAGAACAAGCCAA
CY040004	1816-1838	AGACATGACCAAGGAATTC		1818-1840	AGACATGACCAAGGAATTC		1802-1824	AGACATGACCAAGGAATTC
CY040012	1815-1837	GACATGACCAAGGAATTCT		1817-1839	GACATGACCAAGGAATTCT		1803-1825	GACATGACCAAGGAATTCT
CY040036	340-362	GAGAACCGGTTCATTGAAA		342-364	GAGAACCGGTTCATTGAAA		327-349	GAGAACCGGTTCATTGAAA
CY041787	207-229	TGCACTATTGAAGCACCGA		209-231	TGCACTATTGAAGCACCGA		194-216	TGCACTATTGAAGCACCGA
FJ966970	992-1014	ATCCCAATTACCTCATGGC	CY041763	994-1016	ATCCCAATTACCTCATGGC	GQ365657	979-1001	ATCCCAATTACCTCATGGC
CY041779	1539-1561	TGATACTGATGTGGTGAAC		1541-1563	TGATACTGATGTGGTGAAC		1526-1548	TGATACTGATGTGGTGAAC
GQ351289	2109-2131	TGCATCTTGGTTCAACTCC		2111-2133	TGCATCTTGGTTCAACTCC		2096-2118	TGCATCTTGGTTCAACTCC
CY041055	1492-1514	ACAAACCTGTATGGGTTCA		1494-1516	ACAAACCTGTATGGGTTCA		1479-1501	ACAAACCTGTATGGGTTCA
	1077-1099	CATGAAGAGAACAAGCCAA		1079-1101	CATGAAGAGAACAAGCCAA		1064-1086	CATGAAGAGAACAAGCCAA
	1820-1842	AGACATGACCAAGGAATTC		1825-1847	AGACATGACCAAGGAATTC	FJ966970、FJ984400、FJ969524	1813-1835	AGACATGACCAAGGAATTC
	1819-1841	GACATGACCAAGGAATTCT		1826-1848	GACATGACCAAGGAATTCT	GQ117029、FJ969515、FJ966081	1814-1836	GACATGACCAAGGAATTCT
	344-366	GAGAACCGGTTCATTGAAA		350-372	GAGAACCGGTTCATTGAAA	FJ969529、FJ984368、GQ166210	338-360	GAGAACCGGTTCATTGAAA
	211-233	TGCACTATTGAAGCACCGA		217-239	TGCACTATTGAAGCACCGA	GQ149686、GQ149676、GQ222036	205-227	TGCACTATTGAAGCACCGA
CY041835	996-1018	ATCCCAATTACCTCATGGC		1002-1024	ATCCCAATTACCTCATGGC	GQ169305、GQ222034、GQ222038	990-1012	ATCCCAATTACCTCATGGC
CY041819	1543-1565	TGATACTGATGTGGTGAAC	GQ247729	1549-1571	TGATACTGATGTGGTGAAC	GQ365451、GQ334351、GQ334360	1537-1559	TGATACTGATGTGGTGAAC
CY041827	2113-2135	TGCATCTTGGTTCAACTCC	GQ330650	2119-2141	TGCATCTTGGTTCAACTCC	GQ365460、GQ222032、GQ222033	2107-2129	TGCATCTTGGTTCAACTCC
CY041771	1496-1518	ACAAACCTGTATGGGTTCA		1502-1524	ACAAACCTGTATGGGTTCA	GQ334335、GQ334343、GQ166226	1490-1512	ACAAACCTGTATGGGTTCA
CY041811	1081-1103	CATGAAGAGAACAAGCCAA		1087-1109	CATGAAGAGAACAAGCCAA		1075-1097	CATGAAGAGAACAAGCCAA
CY041795								

表 3 结构分析表

Table 3 Structural analysis

编号	序列号	靶序列	二级结构
1	1814-1836	AGACATGACCAAGGAATTC	.....(((((....)
2	1815-1837	GACATGACCAAGGAATTCT	.....(((((....))
3	339-361	GAGAACCGGTTCATTGAAA	...)))...)))))...))
4	206-228	TGCACTATTGAAGCACCGA	))))..)))))).....))))
5	991-1013	ATCCCAATTACCTCATGGC	))))......))))..))))
6	1538-1560	TGATACTGATGTGGTGAAC	..(((((......((....))
7	2108-2130	TGCATCTTGGTTCAACTCC	))))))))))))...))))
8	1491-1513	ACAAACCTGTATGGGTTCA	...(((((....)))))..
9	1076-1098	CATGAAGAGAACAAGCCAA	............((((..(

表 4 靶序列对应的 siRNA 序列

Table 4 siRNA sequences corresponding to target sequences

项目	序号		
	7	4	5
靶序列	TGCATCTTGGTTCAACTCC	TGCACTATTGAAGCACCGA	ATCCCAATTACCTCATGGC
siRNA	S 5′:UGCAUCUUGGUUCAACUCCdTdT AS 3′:TdTdACGUAGAACCAAGUUGAGG	S 5′:UGCACUAUUGAAGCACCGAdTdT AS 3′:TdTdACGUGAUAACUUCGUGGCU	S 5′:AUCCCAAUUACCUCAUGGCdTdT AS 3′:TdTdUAGGGUUAAUGGAGUACCG

5 结 论

采用生物信息学的方法对目前流行的甲型流感病毒 H1N1 亚型进行了分析，对可能影响 siRNA 活性的一级序列、二级结构因素做出了分析和评价。研究表明，2009 年爆发的 H1N1 病毒序列保守性高，通过序列分析所获得的序列一致，通用性高；通过对甲型流感病毒 H1N1 亚型基因二级结构的分析发现，基因中二级结构也存在“空间保守区域”的现象，为从二级结构角度考虑 siRNA 优化设计提出了新思路，采用多特征融合的方法（序列特征和结构特征），进行了 siRNA 设计，获得了多条共同的 siRNA，为生物实验提供了理论依据。综上，本研究为利用 RNAi 进行 H1N1 治疗的后续研发奠定了基础，为流感病毒基因的 siRNA 优化设计提供了重要依据。

参考文献：

[1] Elbashir S M, Harborth J, Weber K, et al. Analysis of gene function in somatic mammalian cells using small interfering RNAs[J]. Methods, 2002, 26(2): 199-213.

[2] Reynolds A, Leake D, Boese Q, et al. Rational siRNA design for RNA interference[J]. Natbiotec-hnol, 2004, 22(3): 326-330.

[3] Amarzguioui M, Prydz H. An algorithm for selection of functional siRNA sequences[J]. Biochem Bioph Res Co, 2004, 316(4): 1050-1058.

[4] Chalk A M, Wahlestedt C, Sonnhammer E L L. Improved and automated prediction of effective siRNA [J]. Biochem Bioph Res Co, 2004, 319(1): 264-274.

[5] 许德晖，黄辰，刘利英，等. 高效 siRNA 设计的研究进展[J]. 遗传，2006，28(11)：1457-1461.
Xu De-hui, Huang Chen, Liu Li-ying, et al. New progress of the highly efficient siRNA design[J]. Hereditas, 2006, 28(11): 1457-1461.

[6] 胡颖，叶枫，谢幸. RNA 干扰技术中 siRNA 设计原则的研究进展[J]. 国际遗传学，2007，30(6)：419-423.
Hu Ying, Ye Feng, Xie Xing. Advances in the design of RNAi technology[J]. International Journal of Genetics, 2007, 30(6): 419-423.

[7] 付洁. 靶向 HBV 的 siRNA 优化设计研究[D]. 长春：中国人民解放军军事医学科学院，2008.
Fu Jie. Optimization of siRNA design targeting against HBV[D]. Changchun: Academy of Military Medcial Sciences of China, 2008.

[8] 郭骁才，郭红霞. 甲型流感病毒 H5N1 的 siRNA 设计[J]. 应用与环境生物学报，2004，10(1)：133-138.
Guo Xiao-cai, Guo Hong-xia. SiRNA design to the H5N1 subtype of influenza a virus[J]. Chinese Journal of Applied and Environmental Biology 2004, 10(1): 133-138.

[9] 汪本助，李艳，张海燕，等. 靶向 EGFR 的 siRNA 有效抑制小鼠肺癌移植瘤的生长[J]. 中国科学技术大学学报，2009，39(7)：706-712.
Wang Ben-zhu, Li Yan, Zhang Hai-yan, et al. Inhibition of lung tumor growth in nude mice by siRNA targeting on EGFR[J]. Journal of University of Science and Technology of China, 2009, 39(7): 706-712.

[10] 王芳芳，马志强，王素华. 基于遗传算法的序列比对方法[J]. 吉林大学学报：信息科学版，2006，24(4)：423-429.
Wang Fang-fang, Ma Zhi-qiang, Wang Su-hua. Approach to sequence alignment based on genetic algorithm[J]. Journal of Jilin University (Information Science Edition), 2006, 24(4): 423-429.