微生物学报 *Acta Microbiologica Sinica*
49(11):1445 – 1453;4 November 2009
ISSN 0001 – 6209;CN 11 – 1995/Q
http://journals.im.ac.cn/actamicrocn

极端嗜盐古菌中 CRISPR 结构的生物信息学分析

张帆[1],张兵[1],向华[2],胡松年[1*]
([1] 中国科学院北京基因组研究所,北京 100029)
([2] 中国科学院微生物研究所微生物资源前期开发国家重点实验室,北京 100101)

摘要:【目的】利用生物信息学方法了解目前拥有全基因组序列的极端嗜盐古菌中 CRISPR 结构的特征。【方法】通过比对,保守性分析,GC 含量分析,RNA 结构预测等方法对已有全基因组序列的嗜盐古菌基因组进行研究。【结果】在 5 株嗜盐古菌基因组中发现 CRISPR 结构,在 leader 序列内得到具有回文性质的保守 motif。发现在大 CRISPR 结构内 repeat 序列具有很强的保守性。同时根据第四位碱基的不同,repeat 序列可形成两类不同的 RNA 二级结构。【结论】leader 序列中回文结构的发现对其可能为蛋白结合位点的假设提供了进一步的理论依据。Repeat 序列 RNA 二级结构的形成提示其可能介导外源 DNA 或 RNA 与 CAS 编码蛋白的相互作用。

关键词:CRISPR;leader;repeat;嗜盐古菌;motif;palindromic
中图分类号:Q933 **文献标识码**:A **文章编号**:0001-6209(2009)11-1445-09

规律成簇的间隔短回文重复(Clustered regularly interspaced short palindromic repeats,CRISPR)是一种广泛分布于细菌与古菌中高度多样性的遗传结构[1]。该结构以连续相同的重复序列(repeat)为特征,repeat 序列一般长 24 至 47 个碱基,在每两个 repeat 序列间含有高度特异性的插入短片段(spacer),在两端的侧翼区域中经常出现 CRISPR – associated(*cas*)基因[2-5]。在一个 CRISPR 结构中所有 repeat 序列之间几乎完全一致且具有回文性质[1,6]。有文献报道在 repeat 序列的回文结构中发现互补的碱基突变而认为该回文结构与 RNA 二级结构的形成有关[6]。Repeat 序列间的 spacer 序列是整个 CRISPR 结构中最多样化的部分,已有研究表明这些片段来自外源遗传组分,且只有极少数是通过复制产生[4]。目前认为 repeat 序列和 spacer 序列均通过转录为 RNA 二级结构而发挥作用,在转录产生的 RNA 二级结构中,repeat 序列形成茎部分而 spacer 序列形成环部分[6,7]。

根据 Daniel H. Haft et al. 等的研究,目前共有 45 个 *cas* 基因家族,这 45 个基因家族又大致分为 8 个亚型,每个亚型由该亚型特异的基因组成[2]。除此之外有 6 个基因(*cas*1-6),因发现与大多数不同的亚型基因均关联,被认为是 *cas* 基因中的核心基因[2]。在这六个核心基因中又以 *cas*1 为 CRISPR 结构的标志基因[2]。但目前对于所有 *cas* 基因的功能和作用机制并不清楚,只通过个别发现的 motif 猜测其可能拥有的功能。

通常在 CRISPR 结构第一个 repeat 序列的 5′上游区域存在一段最长可达 550 bp 的 AT 富含区,称为 leader 序列,在不同物种间不保守[1,8]。有文献报道发现新插入的 spacer 序列总是加在 leader 序列与相邻的末端 repeat 序列之间,并且 leader 序列总是直接位于末端 repeat 序列的上游,提示 leader 序列可能在新 spacer 序列增加过程中作为识别位点,或作为 CRISPR 结构转录的启动子(promoter)存在[9]。

在 CRISPR 的研究过程中,因发现该结构中的

* 通信作者。Tel/Fax:+ 86-10-82995362;E-mail:husn@big.ac.cn
作者简介:张帆(1983 –),男,上海市人,硕士研究生,主要研究方向为基因组学。E-mail:fanz07@yahoo.com.cn
收稿日期:2009-04-28;修回日期:2009-05-21

spacer 序列与外源基因组如噬菌体或质粒中的某些片段高度相似,而设想 CRISPR 结构的功能与抵御外源入侵有关[8]。这一假设在最近通过不同的研究均得到了实验上的证实,表明 CRISPR 结构能够赋予细胞抵御噬菌体侵染的能力[9, 10]。现在普遍认为 CRISPR 结构是作为原核生物中一种抗噬菌体机制而存在,且这种抗性的发挥类似于真核生物中 RNA 干扰作用[3]。

目前在细菌中关于 CRISPR 结构的研究主要集中在乳酸菌(lactic acid bacteria)和嗜热链球菌(*Streptococcus thermophilus*)等中,关于 CRISPR 结构功能的实验证明也主要是在这些菌株中完成的[9, 10]。而在嗜盐古菌中 CRISPR 结构的研究还相对较少,一是由于对 CRISPR 结构研究的限制,二是因为具有全基因组序列的嗜盐古菌数量的稀少。在美国国立生物技术信息中心(NCBI)的网站上现有 5 株具有全基因组序列的嗜盐古菌:*Haloarcula marismortui* ATCC 43049[11], *Halobacterium salinarum* R1[12], *Halobacterium sp*. NRC-1[13], *Haloquadratum walsbyi* DSM 16790[14] 和 *Natronomonas pharaonis* DSM 2160[15]。在加州大学圣克鲁兹分校(http://genome.ucsc.edu)的网站上有 *Haloferax volcanii* DS2[16] 的全基因组序列。另一株菌 *Haloferax mediterranii* AS2087 的测序项目已完成(未发表)。古菌是细胞生命的第三种形式。现在大部分的古菌(90%)中均发现了 CRISPR 结构的存在,而在细菌中这一比例大概为 40%[4]。这一现象提示 CRISPR 结构对于古菌的生存更为重要。同时,古菌 CRISPR 结构具有很多区别于细菌的特征,如古菌中 CRISPR 结构经常很巨大[8];*cas* 基因更加多样等[2]。因此对古菌中 CRISPR 结构多样性的研究将丰富其进化功能的新认识。目前实验观测到的 CRISPR 结构 RNA 转录现象就是在古菌 *Archaeoglobus fulgidus* 中完成[7]。

1 材料和方法

1.1 实验数据

H. mediterranii 全基因组序列由中国科学院微生物研究所向华实验室与北京基因组研究所胡松年实验室联合完成(未发表)。*H. marismortui*, *H. salinarum*, *Halobacterium sp.*, *H. walsbyi* 和 *N. pharaonis* 的全基因组序列从 NCBI(http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi)下载,*H. volcanii* 的全基因组序列从加州大学圣克鲁兹分校基因组生物信息中心(http://archaea.ucsc.edu/cgi-bin/hgGateway?db=haloVolc1)下载。

1.2 实验方法

H. marismortui, *H. salinarum*, *Halobacterium sp.*, *H. walsbyi* 和 *N. pharaonis* 的 CRISPR 信息(包括 repeat 序列、spacer 序列和侧翼序列)从 CRISPRdb 数据库下载[5]。*H. mediterranii* 与 *H. volcanii* 的 CRISPR 信息(包括 repeat 序列、spacer 序列和侧翼序列)通过 CRISPRFinder 软件[17] 获得。利用 Glimmer 软件[18] 对新测序的 *H. mediterranii* 进行开放阅读框(ORF)的预测,将预测产生的 ORF 与另六株菌的基因序列共同通过 Blast[19] 与 45 个 *cas* 基因家族比对寻找 *cas* 基因(E 值为 1e-05)。通过 NCBI Blast,缺省参数,与 NCBI 的 nr 数据库比对寻找 spacer 序列的外源相似片段。Leader 序列的多序列比对通过 Clustal W[20] 完成。RNA 二级结构的预测通过 RNAfold 完成[21]。

根据文献报道 CRISPR 结构通常不位于编码区内[4, 5],而无论是 CRISPRdb 数据库还是 CRISPRFinder 软件均涵盖了所有的可能性,因此在得到全部 CRISPR 信息后将进行人工校对。将位于编码区内的位点排除,同时对于 repeat 序列超过 48 bp的 CRISPR 结构也予以排除[4-5]。另外为便于叙述,将排除的位点称为非真位点(not real CRISPR locus),将拥有较少 repeat 序列(如 2 到 5 个)的 CRISPR 结构称为疑似位点。

2 结果和分析

2.1 7 株嗜盐古菌 CRISPR 概况

7 株古菌中除 *Halobacterium sp*. NRC-1 和 *H. salinarum* R1 外在其他 5 株菌中均发现了 CRISPR 结构的存在,这一结果与之前报道的未在 *Halobacterium sp*. 中发现 CRISPR 结构相一致[8]。在这 5 株菌种,总共发现有 24 个 CRISPR 位点(包括疑似),所有这些位点的信息总结于表 1。

在 *H. marismortui* 中,共发现 5 个 CRISPR 位点,这些位点全部位于质粒上,其中 2 个位于质粒 pNG300,3 个位于质粒 pNG400(表 1)。在这些位点中有 3 个位点拥有较多的 repeat 序列数(48,26,52),且这 3 个位点的 repeat 序列大小均为 30 bp。其中,位点 Hmari 5 的 repeat 序列数(52 个)是 24 个 CRISPR 中最多的。另外在质粒 pNG400 上发现有 8 个 *cas* 基因位于位点 Hmari 4 和 Hmari 5 之间(图 1-A)。在 *H. marismortui* 的 5 个 CRISPR 位点中,Hmari 3 和 Hmari 5 的 repeat 序列是一致的。

表 1 7 株极端嗜盐古菌中 CRISPR 结构概况

Table 1 The CRISPR loci in 7 halophilic archaea genomes

Name	Number of CRISPR loci	CRISPR locus name	CRISPR Loci	Number of repeats	Repeat size	Spacer size (bp) (min-max)	*cas* genes	Representative repeat sequence (5′→3′)
Haloarcula marismortui ATCC 43049	6	Hmari 1[b]	plasmid pNG300	3	38	43	No	TGTTGTGATTGCTTCATAAAGTTGGTATTCGGATGATT
		Hmari 2[c]	plasmid pNG300	48	30	35 (34 – 38)	No	GCTTCAACCCCACAAGGGTCCGTCTGAAAC
		Hmari 3	plasmid pNG400	5	30	36 (35 – 36)	No	GTTACAGACGGACCCTCGTGGGGTTGAAGC
		Hmari 4[c]	plasmid pNG400	26	30	36 (33 – 39)	8	GTTTCAGACGGACCCTTGGGCGGTTGAAGT
		Hmari 5[c]	plasmid pNG400	52	30	36 (34 – 38)	8	GTTACAGACGGACCCTCGTGGGGTTGAAGC
		Hmari 6[a]	chr I	4	23	25 (16 – 25)	No	GGCGGTCCCTGTTCGCTCTGGTT
Halobacterium salinarum R1	1	Hsal 1[a]	chr	3	32	40，22	No	TCGTCCATCTCCTCGTCATCGTCCTCGAAGTC
NRC-1	0	—	—	—	—	—	—	—
Haloquadratum walsbyi DSM 16790	2	Hwal 1[c]	chr	3	25	41，39	No	GTTTCAGATGAACCCTTGATGGGTT
		Hwal 2	chr	4	25	42	No	GTTTCAGATGAACCCTTGATGGGTT
Natronomonas pharaonis DSM 2160	5	Npha 1[c]	chr	5	30	36 (35 – 37)	No	GTTTCAGACGAACCCTTGTGGGGTTGAAGC
		Npha 2	chr	9	37	35 (34 – 38)	3	GTCGAGACGGACTGAAAACCCAGAACGGGATTGAAAC
		Npha 3[c]	plasmid PL131	3	30	36	No	GTTTCAGACGAACCCTTGTGGGGTTGAAGC
		Npha 4[b]	plasmid PL131	4	23	55 (48 – 56)	No	GCACCCCTCTATCGATGTGTACT
		Npha 5	plasmid PL23	8	37	35 (34 – 38)	No	GTCGAGACGGACTGAAAACCCAGAACGGGATTGAAAC
Haloferax mediterranii AS2087	10	Hmed 1[c]	chr	18	30	35 (33 – 38)	No	GTTACAGACGAACCCTAGTTGGGTTGAAGC
		Hmed 2[c]	chr	21	30	35 (34 – 38)	No	GTTACAGACGAACCCTAGTTGGGTTGAAGC
		Hmed 3[c]	chr	25	30	37 (34 – 54)	No	GCTTCAACCCAACTAGGGTTCGTCTGTAAC
		Hmed 4[b]	chr	2	25	39	No	GTTCGAATCCTGGTGGAACTATACT
		Hmed 5[c]	chr	2	30	37	No	GGTTACAGACGGACCCTCGTTGGGTTGAAG
		Hmed 6[a]	chr	2	39	51	No	GCCGCCGCCCATGCCGCCGGGGCCGCCACCGCCCATCAT
		Hmed 7[b]	chr	4	24	16	No	ACCTTACCTCACCTTACCTTACCT
		Hmed 8[a]	plasmid 100	2	38	33	No	TGTTTCAATCCCGTGTTGGGTTTCTACCGCACTGCGAC
Haloferax mediterranei AS2087	10	Hmed 9[c]	plasmid 500	11	30	36 (35 – 38)	8	GCTTCAACCCAATTAGGGTTCGTCTGTAAC
		Hmed 10[c]	plasmid 500	23	30	36 (34 – 39)	8	GTTACAGACGAACCCTAGTTGGGTTGAAGC
Haloferax volcanii DS2	9	Hvol 1[c]	chr	25	30	36 (34 – 39)	No	GCTTCAACCCCACAAGGGTTCGTCTGAAAC
		Hvol 2[a]	chr	2	42	27	No	GCCGCCACCGCCCATGCCGCCGGGGC CGCCGCCGCCCATCAT
		Hvol 3[a]	chr	2	24	42	No	CCGCTCGGCTCACGACGTTCGCCT
		Hvol 4[a]	chr	2	48	47	No	GGACCAAAAATAATCGCCAAGCGACTA CGCCAGCGCCGACAGGTCGAA
		Hvol 5[b]	plasmid pHV1	2	25	41	No	CACCCACCCCTCGTTTCCGACGTTC
		Hvol 6[c]	plasmid pHV4	40	30	36 (34 – 39)	8	GTTTCAGACGAACCCTTGTGGGATTGAAGC
		Hvol 7[c]	plasmid pHV4	12	31	35 (33 – 37)	8	GGTTTCAGACGAACCCTTGTGGGTTTGAAGC
		Hvol 8[a]	plasmid pHV4	2	30	33	No	GCGCGAAGCGCATCGTAGTACGTCCGTAAG
		Hvol 9[a]	plasmid pHV4	2	24	48	No	ACACCACGTTGTCCGCTGTTCTCG

[a]Not real CRISPR locus；[b]Questionable CRISPR locus；[c]has leader conservative motif around

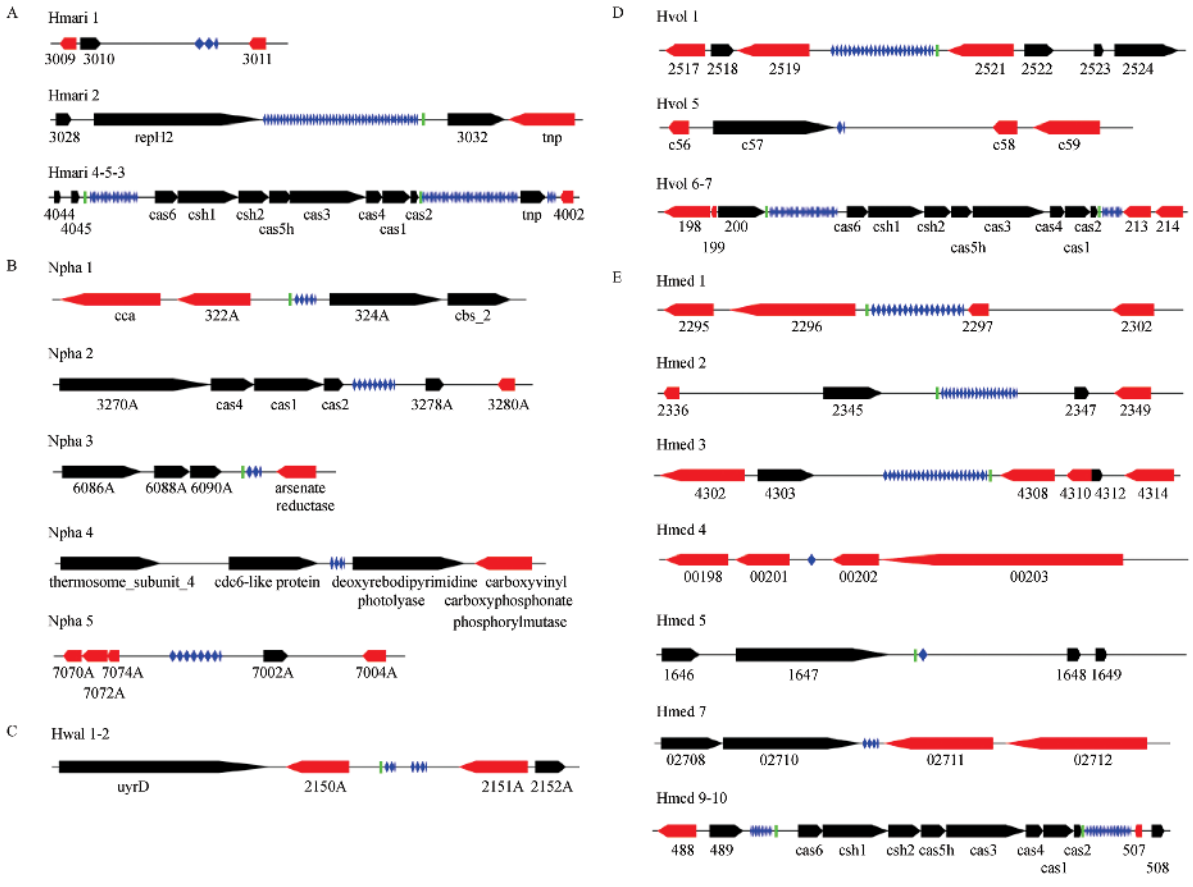

图 1　5 株极端嗜盐古菌中 24 个 CRISPR 位点结构示意图

Fig.1　Graphic representation of CRISPR loci in 5 halophilic archaea. The black and red box arrows represent the genes in different strand; the blue diamonds represent the CRISPR loci in repeat – spacer units. The green box represents the leader motif. The words below each gene indicate the name of that one. A. CRISPR loci in *Haloarcula marismortui* ATCC 43049; B. CRISPR loci in *Natronomonas pharaonis* DSM 2160; C. CRISPR loci in *Haloquadratum walsbyi* DSM 16790; D. CRISPR loci in *Haloferax volcanii* DS2; E. CRISPR loci in *Haloferax mediterranii* AS2087.

在 *N. pharaonis* 和 *H. walsbyi* 中各自发现了 5 个和 2 个 CRISPR 位点(表 1 ;图 1-B,图 1-C)。在 *N. pharaonis* 中 5 个位点分别位于染色体和质粒,只发现有 3 个 *cas* 基因在位点 Npha 2 附近。其中位点 Npha 1 和 Npha 3,位点 Npha 2 和 Npha 5 的 repeat 序列是一致的。而在 *H. walsbyi* 中,2 个位点均位于染色体但没有发现 *cas* 基因的存在,而这两个位点的 repeat 序列也完全一致。

在 *H. volcanii* 和 *H. mediterranii* 中,CRISPR 结构彼此类似。在染色体和质粒上均发现有 CRISPR 结构的存在,同时两株菌各自含有 8 个 *cas* 基因在最大的质粒上(表 1 ;图 1-D,图 1-E)。与另 3 株菌不同的是在这两株菌的染色体上也发现了大的 CRISPR 位点,在 *H. volcanii* 中是 Hvol 1,其 repeat 序列数为 25,而在 *H. mediterranii* 中是 Hmed 1-3,repeat 序列数为 18,21 和 25。而文献报道称 CRISPR 结构通常位于进化速率较快的结构中,如质粒[6]。在 *H. mediterranii* 中有 3 个位点的 repeat 序列是一致的(Hmed 1,Hmed 2 和 Hmed 10),而在 *H. volcanii* 中则没有发现 repeat 序列彼此一致的位点。

2.2　Leader 序列

在进行 leader 序列分析时,选取了 24 个位点上游 650 bp 范围的序列进行比对分析。经过 Clustal W2 [http://www.ebi.ac.uk/Tools/clustalw2/]的多轮比较后,在 5 个物种的 15 个 CRISPR 位点(Hmari 2, Hmari 4, Hmari 5, Hwal 1, Npha 1, Npha 3, Hmed 1, Hmed 2, Hmed 3, Hmed 5, Hmed 9, Hmed 10, Hvol 1, Hvol 6 和 Hvol 7)的侧翼序列中发现了一个较为保守且具有部分回文结构的 motif。该保守结构起始自第一个 repeat 序列大约 95 个碱基,以“TCGACC”的 6 碱基排列为特征。在距离大约 25 个碱基后,出现与之互补的另一 6 碱基排列“GGTCGA”,形成回文结构。而第一个 repeat 序列距离该回文结构大约 58 个碱基。

但该 motif 并不完全保守，比如，在 Npha 1 位点中，该 motif 以碱基“A”代替“TCGACC”中第三位碱基“G”；而在 Npha 3 中以碱基“T”代替“GGTCGA”倒数第三位碱基“C”。通过分析发现，这些非保守位点基本上都来源于拥有较少 repeat 序列数的 CRISPR 结构，提示当 CRISPR 结构的 repeat 序列数较少即一些小 CRISPR 结构不如拥有较多 repeat 序列数的大 CRISPR 结构保守。

因此将 15 个位点中的 4 个小位点(Npha 1，Npha 3，Hwal 1 和 Hmed 5)排除。将剩余的 11 个大 CRISPR 位点进行多序列比对。比对结果如图 2 所示。

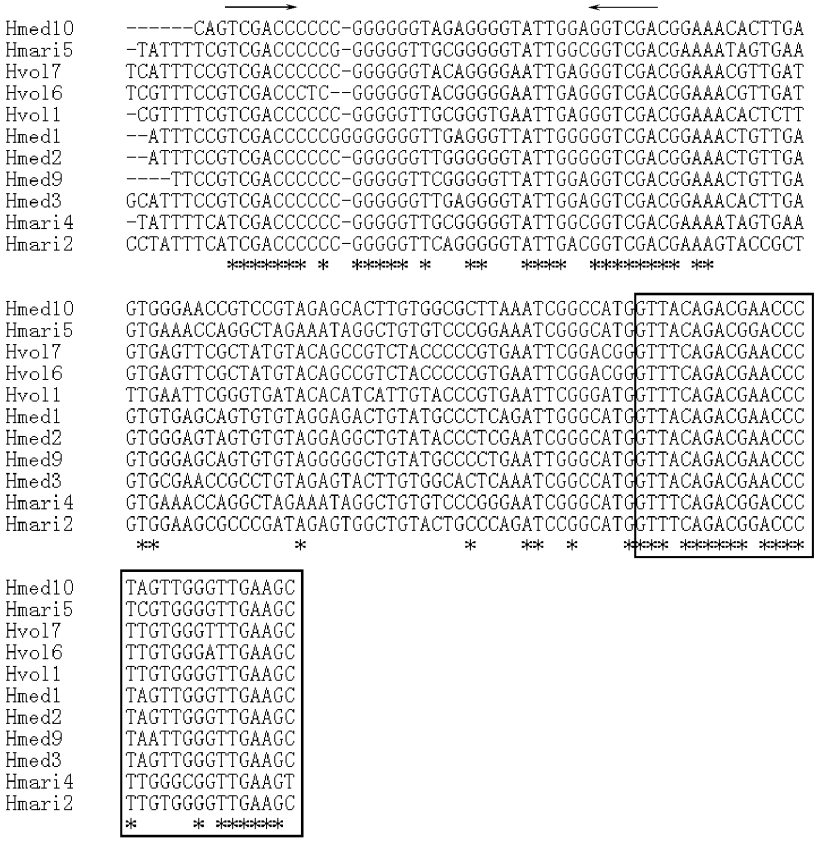

图 2 3 株嗜盐古菌中 11 个 CRISPR 位点的 Leader 序列保守性比对分析

Fig.2 Alignment of the 11 conserved flanking sequences and the first repeats (in black box) of 3 halophilic archaea species. The arrow indicates the palindromic structure of the alignment. The asterisk represents the conserved base in all the 11 loci.

据图 2 所示，在去除由小 CRISPR 位点产生的非保守碱基后，在 3 个物种的 11 个 CRISPR 位点中，存在一较为保守的 motif。该 motif 以连续出现的“C”或“G”碱基为特征。回文结构的前后两字符串“TCGACC”和“GGTCGA”均表现出完全的保守性。同样，该 motif 距离第一个 repeat 序列大约 95 个碱基，并且在该区域范围外未发现其他保守结构存在。回文结构前后两字符串相隔大约 25 个碱基，第一个 repeat 序列距离该 motif 大约 58 个碱基。

leader 序列的 GC 含量分析显示，拥有该 motif 的 leader 序列倾向于高 GC 含量(表 2)，并且在 CRISPR 位点上下游序列中经常呈现出较明显的 GC 含量差异。含有 motif 的一侧 GC 含量要高于另一侧。虽有例外存在，如 Npha 1 中含 leader motif 的侧翼序列的 GC 含量比另一侧低接近 20%(表 2)，但总

表 2 15 个 leader-positive CRISPR 位点的两端侧翼序列 GC 含量分析

Table 2 The GC content of two flanking regions of each leader-positive CRISPR locus.

Locus	GC% 100 bp upstream	GC% 100 bp downstream	Discrepancy/%
Hmari 2^{L}	60.6	**61.6**	1.6
Hmari 4^{L}	**57.6**	51.5	11.7
Hmari 5^{L}	**57.6**	56.6	1.7
Hmed 1^{L}	**59.6**	50.5	18
Hmed 2^{L}	**61.6**	48.5	27
Hmed 3^{L}	47.5	**59.6**	25.4
Hmed 5	**57.6**	46.5	23.8
Hmed 9^{L}	53.5	**61.6**	15
Hmed 10^{L}	**60.6**	40.4	50
Hvol 1^{L}	57.6	**53.5**	−7
Hvol 6^{L}	**59.6**	52.5	13.4
Hvol 7^{L}	**59.6**	62.6	−4.8
Npha 1	**53.5**	65.7	−18.5
Npha 3	**59.6**	63.6	−6.2
Hwal 1	**48.5**	43.4	11.6

[L]indicates large CRISPR array.

The bold and italic figure represents that the conserved leader motif is in that region.

体上当一个 CRISPR 位点侧翼序列中含有 leader motif 时,该侧翼区域的 GC 含量会比另一侧高大约 10%或以上。

2.3 repeat 序列

根据对 leader 序列的分析,将所有 24 个 CRISPR 位点分为两组,一组中的 CRISPR 位点均在其侧翼序列内含有保守 motif,将该组命名为 leader-positive 组,另一组中的 CRISPR 位点则不含有该 motif,命名为 leader-negative 组。通过分类发现,在 leader-positive 组中几乎所有的 CRISPR 结构中的 repeat 序列长度均为 30 bp,并且所有的 11 个大 CRISPR 位点均归为该组。而在 leader-negative 组中,repeat 序列的长度和 repeat 序列的数目均很波动,序列长度从 23 bp 到 38 bp 不等,数量从 2 到 9 不等且数目均较小(表 3)。

表 3 24 个嗜盐古菌 CRISPR 位点 leader 序列划分表

Table 3 Classification of 24 CRISPR arrays in 5 halophilic archaea based on leader sequences

Leader motif positive				Leader motif negative			
No.	Locus	Repeat size	Repeat amount	No.	Locus	Repeat size	Repeat amount
1	Hmari 2	30	48	16	Hmari 1	38	3
2	Hmari 4	30	26	17	Hmari 3	30	5
3	Hmari 5	30	52	18	Hmed 4	25	2
4	Hmed 1	30	18	19	Hmed 7	24	4
5	Hmed 2	30	21	20	Hvol 5	25	2
6	Hmed 3	30	25	21	Npha 2	37	9
7	Hmed 5	30	2	22	Npha 4	23	4
8	Hmed 9	30	11	23	Npha 5	37	8
9	Hmed 10	30	23	24	Hwal 2	25	4
10	Hovl 1	30	25				
11	Hovl 6	30	40				
12	Hovl 7	31	12				
13	Hpha 1	30	5				
14	Hpha 3	30	3				
15	Hwal 1	25	3				

从表 3 中可以看出 30 bp 这一 repeat 序列的长度在嗜盐古菌的 CRISPR 结构中可能具有重要意义。

有文献报道 repeat 序列与 RNA 二级结构形成之间具有关联性[6],故选取了 leader-positive 组中 15 个 CRISPR 位点进行 RNA 结构分析。因在同一 CRISPR 位点中,repeat 序列无论是序列组成还是长度均一致[4],因此每一位点只选取一条代表性的 repeat 序列予以分析,结果如图 3 所示。从图中可看出 15 个位点的 RNA 二级结构可分为两类,一类由 Hmed 1,Hmed 2,Hmed 3,Hmed 5,Hmed 9,Hmed 10,Hvol 7,Hmari 5 和 Hwal 1 等 9 个 CRISPR 位点组成。该类 CRISPR 位点形成的 RNA 二级结构的特征是,在整个结构的中部含有唯一的茎部分,一大一小两个环分布在两端,该结构以“环”为主。而另一类结构由剩余的 6 个 CRISPR 位点组成,该类位点形成的 RNA 二级结构的特征是茎与环依次形成,两个环大小类似,以“茎”结构为主(图 3-B,图 3-C)。对 leader-positive 组中 311 条长度为 30 或 31 bp 的 repeat 序列所进行的一致性分析显示,在 30 个碱基中有超过 25 个以上的碱基非常保守(图 3-A)。而对于不保守的第 4,第 11,第 17 和第 20 四个位点,在形成以“环”为主的 RNA 二级结构的 9 个 CRISPR 位点中 7 个位点的 repeat 序列在第 4 位上的碱基是“A”,占 77.7%;而形成以“茎”为主的 RNA 二级结构的 6 个 CRISPR 位点中,repeat 序列在第 4 位上均为碱基“T”,为 100%,提示第 4 位上的碱基对整个 repeat 序列所形成的 RNA 二级结构影响最大。分析发现当该位点上的碱基为“A”时,repeat 序列形成的 RNA 二级结构倾向于以“环”为主的茎环结构(图 3-B),而当该位点上的碱基为“T”时,则更趋于形成以“茎”为主的茎环结构(图 3-C)。虽然这两类结构在第 17 和第 20 位上也有较大不同但并未发现对二级结构的形成有直接的影响作用。

根据 Victor Kunin 等的研究[6],当 repeat 序列能形成以“环”为主的二级结构时,在转录为 RNA 的过程中 spacer 序列与外源 DNA 或 RNA 发生作用不是通过相邻的两个 repeat 序列互补完成而是通过单个的“repeat-spacer”单元完成,在该过程中推测 repeat 序列可能介导外源遗传组分(DNA 或 RNA)与 CAS 编码蛋白的相互作用。本研究得到的 RNA 二级结构结果与文献报道一致故作相同推测。

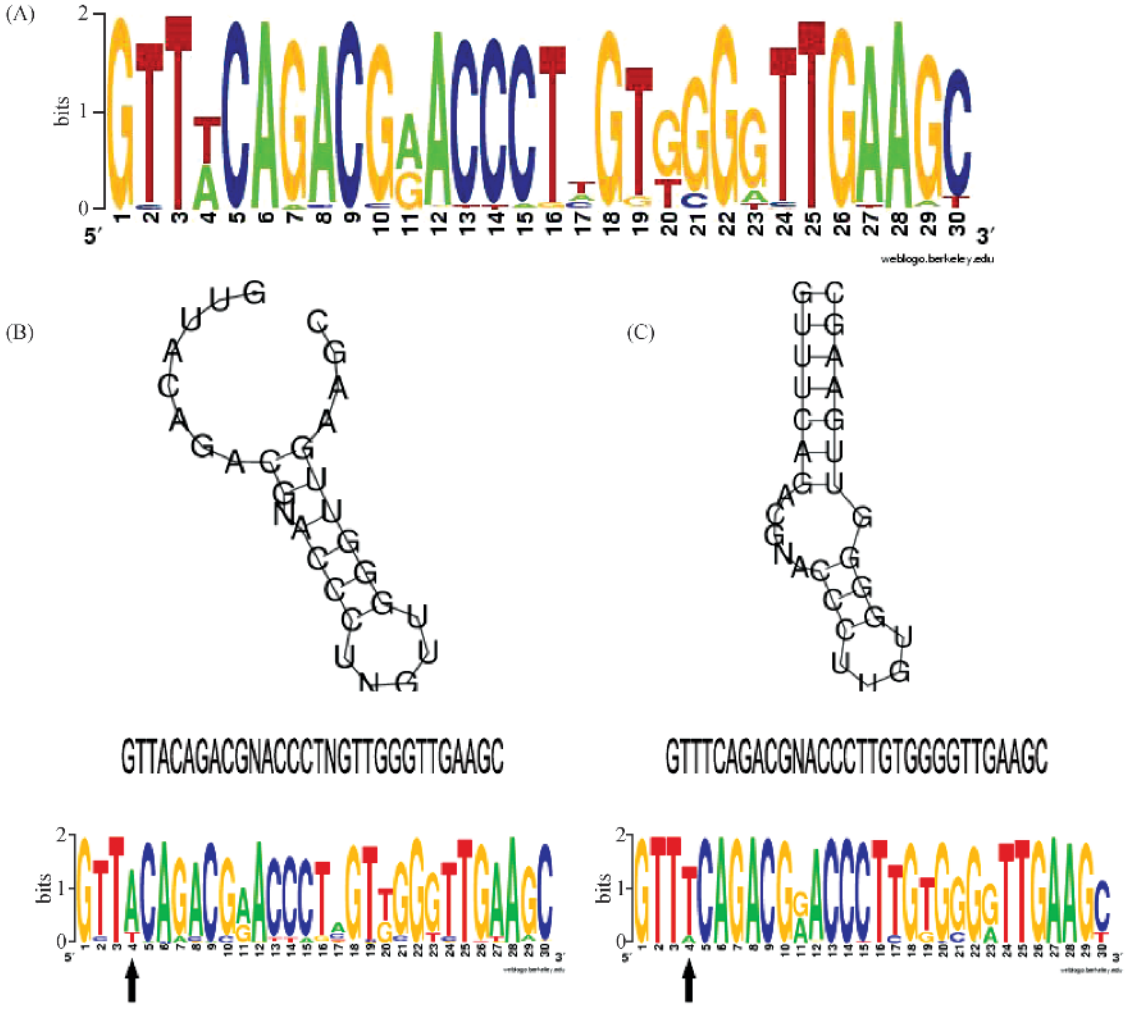

图 3　5 株嗜盐古菌中 311 条 repeat 序列保守性分析及 RNA 二级结构预测

Fig.3　The alignment of 311 repeats in 5 halophilic archaea and the predicted secondary structure of the consensus sequences. The two consensuses and conservation analyses are shown at the bottom of the figure and the non-conserved base is replaced by“N”in the sequences. The black arrow represents the 4th locus of the conservation. A. Sequence logo for all the 311 repeats in 5 halophilic archaea. B. Predicted secondary structure of repeat consensus with“A”at the 4th locus. C. Predicted secondary structure of repeat consensus with“T”at the 4th locus.

2.4　插入片段序列(spacer)及 *cas* 基因

在 24 个 CRISPR 位点中总共有 331 条 spacer 序列，长度从 16 bp 到 56 bp 不等。在这 331 条序列中有少数序列之间是彼此相同的，如 Hmari 3 位点的第 1 条 spacer 序列与 Hmari 5 位点的第 40 条 spacer 序列。而值得注意的是 Npha 2 位点除第一条 spacer 序列外其余 7 条 spacer 序列和 Npha 5 的全部 7 条 spacer 序列完全相同，并且顺序一致。同时 Npha 2 与 Npha 5 的 repeat 序列也均一致，强烈提示 Npha 5 位点来源于 Npha 2 位点的复制。此外，在对 331 条 spacer 序列进行的同源性搜索中没有发现具有显著相似性的外源片段。

在编码 *cas* 基因的 4 株古菌中，*cas* 基因的排列顺序是一致的为 *cas*6-*csh*2-*csh*2-*cas*5*h*-*cas*3-*cas*4-*cas*1-*cas*2(图 1)，这与文献报道一致[2]。在这 8 个基因中，只有 *cas*1 在 4 株菌中均保守。在这 4 株菌中，*H. marismortui* 和 *H. mediterranii* 之间的 *cas* 基因保守性较其他 2 株菌强，平均达到 61%，最高为 *cas*1 基因，达到 76%，最低为 *cas*6，约为 47%。

3　讨论

目前 CRISPR 结构作为一种抵御噬菌体侵染的细胞防御机制，对它的各种作用机理并不完全清楚。但有关 CRISPR 的生物利用前景却十分广泛并且有的已经得到实际应用。

首先，CRISPR 是一种进化速率非常快的结构，在同种内的不同菌株间即使是基因组其余部分均一致的情况下在 CRISPR 结构中也能发现差异，这就为菌

株分型(strain typing)提供了依据[4]。目前依据CRISPR结构进行的菌株分型方法已经得到了具体的应用,其中基于spacer序列的Spoligotyping方法已经成为结核分枝杆菌(*M. tuberculosis*)基因分型的标准方法[4]。其次,作为天然的抗噬菌体机制,CRISPR可以为生物发酵工业如乳制品业和酿酒业中噬菌体污染带来一种全新的解决方法,即通过将已知的侵染发酵菌的噬菌体中保守的序列片段人工插入这些发酵菌的CRISPR结构中,使之产生相应的抗性以抵御这些噬菌体的侵染。同时CRISPR作为一种可遗传的结构,为这一应用更增添了实际意义。最后,如前所述CRISPR结构是一种原核生物中类似真核生物RNA干扰的机制,spacer片段能与外源片段互补配对继而导致其降解失效[3]。假如这一机制得到确证,无疑将具有重大意义。而基因敲除这一费时费力的实验将可以通过另一种快捷的方式来完成,即通过将带有CRISPR结构的质粒转入菌株中,而这一CRISPR结构携带有与目标基因配对的短片段,这样即可达到基因失效的目的。同时更重要的是,根据CRISPR结构的天然特性,一个CRISPR结构中可以带有很多spacer片段,而这些片段又可以来自多个基因的各自不同片段,这就能够实现多个基因的同时敲除,这是传统基因敲除方法所不能实现的。

随着测序技术的发展有越来越多的古菌完成了全基因组测序,在这些已测序的古菌中,CRISPR结构的分布非常广泛,除了嗜盐菌外在耐热(嗜热)菌(thermophiles),产甲烷古菌(methanoarchaea)中也均发现有该结构的存在[8]。一些位于染色体上的CRISPR结构经常呈现出多而大的状态,有的甚至超过基因组的1%,如*Sulfolobus*[8]。

在CRISPR结构的4个组成部分leader序列,repeat序列,spacer片段,*cas*基因中,对于leader序列的了解相对较少。文献报道新插入的spacer片段总是加入在leader序列与紧邻的末端repeat序列之间并认为其可能为蛋白结合位点[9]。本研究在5个物种15个CRISPR位点的leader序列中发现了较为保守的回文结构,在结构数据上为这一假设提供了进一步的理论依据,推测该回文结构很可能为蛋白的结合位点。而在这15个位点中,除了拥有多repeat数的大CRISPR结构以外,也存在只拥有2个repeat序列的小CRISPR结构。这一现象更进一步提示leader序列在CRISPR结构延伸扩展中发挥的重要作用。另外,在发掘leader序列的过程中GC含量在一个CRISPR结构两侧的含量差可起到一定的参考作用(表2)。

本研究未发现与spacer序列具有显著相似性的外源片段。其原因推测为:首先,spacer序列长度短且异常多样化,能够搜索到具有显著相似性外源片段的概率比较小,例如,Mojica等对4500条spacer序列的搜索中只有2%(88条)的序列找到了已知的显著相似性序列[22]。而本研究的331条spacer序列基数相对较少,在很低的概率下难以发现显著相似性的外源片段。其次,根据文献报道[23],目前已知的噬菌体仍然只占非常小的一部分,尚有数量巨大的未知噬菌体存在,而本研究所涉及的极端嗜盐古菌的已知病毒和噬菌体的序列更少,因此未能从现有数据库中发现spacer序列对应的外源片段是可以理解的。

本工作是对嗜盐古菌CRISPR结构首次进行系统的生物学信息分析,为古菌CRISPR的演化及功能机制研究提供了基础数据。

参考文献

[1] Jansen R, van Embden JDA, Gaastra W, et al. Identification of genes that are associated with DNA repeats in prokaryotes. *Molecular Microbiology*, 2002, 43(6): 1565 – 1575.

[2] Haft DH, Selengut J, Mongodin EF, et al. A Guild of 45 CRISPR-Associated (Cas) Protein Families and Multiple CRISPR/Cas Subtypes Exist in Prokaryotic Genomes. *PLoS Computational Biology*, 2005, 1(6): e60.

[3] Makarova KS, Grishin NV, Shabalina SA, et al. A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. *Biology Direct*, 2006, 1: 7.

[4] Sorek R, Kunin V, Hugenholtz P. CRISPR—a widespread system that provides acquired resistance against phages in bacteria and archaea. *Nature Reviews Microbiology*, 2008, 6: 181 – 186.

[5] Grissa I, Vergnaud G, Pourcel C. The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. *BMC Bioinformatics*, 2007, 8: 172.

[6] Kunin V, Sorek R, Hugenholtz P. Evolutionary conservation of sequence and secondary structures in CRISPR repeats. *Genome Biology*, 2007, 8: R61.

[7] Tang TH, Bachellerie JP, Rozhdestvensky T, et al. Identification of 86 candidates for small non-messenger RNAs from the archaeon *Archaeoglobus fulgidus*. *PNAS*, 2002, 99(11): 7536 – 7541.

[8] Lillestøl RK, Redder P, Garrett RA, et al. A Putative Viral Defence Mechanism in Archaeal Cells. *Archaea*, 2006, 2: 59 – 72.

[9] Barrangou R, Fremaux C, Deveau H, et al. CRISPR Provides Acquired Resistance against Viruses in Prokaryotes. *Science*, 2007, 315: 1709 – 1712.

[10] Deveau H, Barrangou R, Garneau JE, et al. Phage Response to CRISPR-Encoded Resistance in *Streptococcus thermophilus*. *Journal of Bacteriology*, 2008, 190(4): 1390 – 1400.

[11] Baliga NS, Bonneau R, Facciotti MT, et al. Genome sequence of *Haloarcula marismortui*: A halophilic archaeon from the Dead Sea. *Genome Research*, 2004, 14: 2221 – 2234.

[12] Pfeiffer F, Schuster SC, Broicher A, et al. Evolution in the laboratory: The genome of *Halobacterium salinarum* strain R1 compared to that of strain NRC-1. *Genomics*, 2008, 91: 335 – 346.

[13] Victor Ng W, Kennedy SP, Mahairas GG, et al. Genome sequence of *Halobacterium species* NRC-1. *PNAS*, 2000, 97(22): 12176 – 12181.

[14] Bolhuis H, Palm P, Wende A, et al. The genome of the square archaeon *Haloquadratum walsbyi*: life at the limits of water activity. *BMC Genomics*, 2006, 7: 169.

[15] Falb M, Pfeiffer F, Palm P, et al. Living with two extremes: Conclusions from the genome sequence of *Natronomonas pharaonis*. *Genome Research*, 2005, 15: 1336 – 1343.

[16] Mullakhanbhai MF, Larsen H. *Halobacterium volcanii* spec. nov., a Dead Sea halobacterium with a moderate salt requirement. *Archives of Microbiology*, 1975, 104(3): 207 – 214.

[17] Grissa I, Vergnaud G, Pourcel C. CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Research*, 2007, 35: w52 – 57.

[18] Delcher AL, Harmon D, Kasif S, et al. Improved microbial gene identification with GLIMMER. *Nucleic Acids Research*, 1999, 27(23): 4636 – 4641.

[19] Altschul SF, Gish W, Miller W, et al. Basic Local Alignment Search Tool. *Journal of Molecular Biology*, 1990, 215: 403 – 410.

[20] Larkin MA, Blackshields G, Brown NP, et al. ClustalW and ClustalX version 2. *Bioinformatics*, 2007, 23(21): 2947 – 2948.

[21] Mathews DH, Sabina J, Zuker M, et al. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *Journal of Molecular Evolution*, 1999, 288: 911 – 940.

[22] Mojica FJM, Diez-Villasenor C, Garcia-Martinez J, et al. Intervening Sequences of Regularly Spaced Prokaryotic Repeats Derive from Foreign Genetic Elements. *Journal of Molecular Evolution*, 2005, 60: 174 – 182.

[23] Edwards RA, Rohwer F. Viral metagenomics. *Nature Reviews Microbiology*, 2005, 3: 504 – 510.

Comparative analysis of Clustered Regularly Interspaced Short Palindromic Repeats (CRISPRs) loci in the genomes of halophilic archaea

Fan Zhang[1], Bing Zhang[1], Hua Xiang[2], Songnian Hu[1*]

([1] Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100029, China)

([2] State Key Laboratory of Microbial Resources, Institute of Microbiology, Chinese Academy of Sciences, Beijing 100101, China)

Abstract: [Objective] Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) is a widespread system that provides acquired resistance against phages in bacteria and archaea. Here we aim to genome-widely analyze the CRISPR in extreme halophilic archaea, of which the whole genome sequences are available at present time. **[Methods]** We used bioinformatics methods including alignment, conservation analysis, GC content and RNA structure prediction to analyze the CRISPR structures of 7 haloarchaeal genomes. **[Results]** We identified the CRISPR structures in 5 halophilic archaea and revealed a conserved palindromic motif in the flanking regions of these CRISPR structures. In addition, we found that the repeat sequences of large CRISPR structures in halophilic archaea were greatly conserved, and two types of predicted RNA secondary structures derived from the repeat sequences were likely determined by the fourth base of the repeat sequence. **[Conclusion]** Our results support the proposal that the leader sequence may function as recognition site by having palindromic structures in flanking regions, and the stem-loop secondary structure formed by repeat sequences may function in mediating the interaction between foreign genetic elements and CAS-encoded proteins.

Keywords: CRISPR; leader; repeat; halophilic archaea; motif; palindromic

(本文责编：王晋芳)

* Corresponding author. Tel/Fax: + 86-10-82995362; E-mail: husn@big.ac.cn

Received: 28 April 2009/Revised: 21 May 2009