Evolutionary conservation of sequence and secondary structures in CRISPR repeats

Victor Kunin, Rotem Sorek and Philip Hugenholtz[¶]

DOE Joint Genome Institute

2800 Mitchell Drive, Walnut Creek,

94598, CA

[¶]Corresponding author

emails:

Victor Kunin: vkunin@lbl.gov

Rotem Sorek: rsorek@lbl.gov

Philip Hugenholtz: phugenhotlz@lbl.gov

DRAFT MANUSCRIPT

SEPTEMBER 2006

Abstract

Clustered Regularly Interspaced Palindromic Repeats (CRISPRs) are a novel class of direct repeats, separated by unique spacer sequences of similar length, that are present in ~40% of bacterial and all archaeal genomes analyzed to date. More than 40 gene families, called CRISPR-associated sequences (CAS), appear in conjunction with these repeats and are thought to be involved in the propagation and functioning of CRISPRs. It has been proposed that the CRISPR/CAS system samples, maintains a record of, and inactivates invasive DNA that the cell has encountered, and therefore constitutes a prokaryotic analog of an immune system. Here we analyze CRISPR repeats identified in 195 microbial genomes and show that they can be organized into multiple clusters based on sequence similarity. All individual repeats in any given cluster were inferred to form characteristic RNA secondary structure, ranging from non-existent to pronounced. Stable secondary structures included G:U base pairs and exhibited multiple compensatory base changes in the stem region, indicating evolutionary conservation and functional importance. We also show that the repeat-based classification corresponds to, and expands upon, a previously reported CAS gene-based classification including specific relationships between CRISPR and CAS subtypes.

Introduction

Clustered Regularly Interspaced Palindromic Repeats (CRISPRs) are repetitive structures in Bacteria and Archaea comprised of exact repeat sequences of typically 21 to 48 bases long (herein repeats) separated by unique spacers of similar length (herein spacers) (Jansen et al. 2002; Mojica et al. 1995). The CRISPR sequences appear to be among the most rapidly evolving parts of the genome, to the point that closely related species and strains, sometimes more than 99% identical on the DNA level, differ in their CRISPR composition (Bolotin et al. 2004; Pourcel et al. 2005).

Up to 45 gene families, called CRISPR-associated sequences (CAS), appear in conjunction with these repeats and are hypothesized to be responsible for CRISPR propagation and functioning (Haft et al. 2005; Jansen et al. 2002; Makarova et al. 2006). It has been proposed that CAS can be divided into seven or eight subtypes, according to their operon organization and gene phylogeny (Haft et al. 2005; Makarova et al. 2002). Phylogenetic analysis additionally indicates that CAS have undergone extensive horizontal gene transfer, as very similar CAS genes are found in distantly related organisms (Godde and Bickerton 2006; Makarova et al. 2006). CRISPRs and CAS have been found on mobile genetic elements, such as megaplasmids, transposons, and even prophages, suggesting a possible distribution mechanism for the system (Godde and Bickerton 2006; Sebaihia et al. 2006).

Initially, CRISPRs were suggested to play a role in replicon partitioning (Mojica et al. 1995) and later speculated to be a part of a DNA repair system specific for thermophilic Archaea and Bacteria (Makarova et al. 2002). However, it was recently reported that the spacers are often highly similar to fragments of extrachromosomal DNA, such as phage or plasmid DNA (Mojica et al. 2005; Pourcel et al. 2005). It was therefore suggested that the CRISPR/CAS system participates in an antiviral response probably by an RNAi-like mechanism. As the proposed mechanism for CRISPR function involves sampling, maintaining a record of, and destroying invasive DNA elements, it is speculated that the CRISPR/CAS system is a prokaryotic analog of an immune system (Mojica et al. 2005).

Despite in-depth analyses of CAS, the nature of the repeat sequences has not been examined closely. This is presumably because repeats, as short DNA sequences, have less comparative potential than protein-coding genes. Previous studies have only noted that repeats are highly variable, and do not appear to be similar between organisms (Godde and Bickerton 2006; Jansen et al. 2002). However, we show that repeats from diverse organisms can be grouped into clusters based on sequence similarity, and that some clusters have pronounced secondary structures with compensatory base changes. We further show that there is a clear correspondence between CAS subtypes and repeat clusters. Our findings have important implications for CRISPR function and diversity.

Results

To obtain a set of CRISPR arrays we employed the PILERCR program [http://www.drive5.com/pilercr/] on 439 currently available bacterial and archaeal genomes in IMG version 1.50 (Markowitz et al. 2006). We found 561 arrays, ranging in size from 3 to 220 repeats, in 195 genomes (44% of the genomes tested). These results

are in agreement with the results of Godde et al (Godde and Bickerton 2006), who found CRISPR arrays in 40% of the genomes they tested. Overall, our set of CRISPR contained 561 repeat sequences (as repeats are generally identical within an array) and 13,372 spacers.

The original report that coined the term CRISPRs described a 'weak palindromic signal' associated with the repeat (Jansen et al. 2002). We hypothesized that the observed palindromic signature might be indicative of a functional RNA secondary structure within the repeat. This hypothesis is supported by the experimental demonstration that CRISPR repeat-spacer pairs are transcribed and processed into non-messenger RNAs in several Archaea (Tang et al. 2002), indicating that they are active through an RNA intermediate.

To assess the possibility that CRISPR repeats form stable RNA secondary structures, we used the RNAfold software (Hofacker et al. 1994) (see Methods) to predict the intramolecular RNA structure for each of the repeats in our set. This software provides a bit-score that reflects the stability of each secondary structure. We compared the stability of the predicted secondary structure of repeats and spacers to that of similarly sized sequences selected randomly from bacterial genomes [Fig 1A]. We found that the folding-score distribution deviates from the scores for random sequences, indicating on a tendency of repeats to form stable secondary structure.

The trimodal pattern of the RNA folding distribution for CRISPR repeats [Fig 1A] suggests that they are not homogeneous, and that a large subset form stable secondary structures, in contrast to spacers and random sequences. To identify repeat subtypes we first attempted to align each of the 561 repeats in our set to all other repeats using the Smith-Waterman algorithm (Smith and Waterman 1981). The sequence similarity results were then clustered using the MCL algorithm (Van Dongen 2000) (see Methods). This procedure generated 33 clusters, 12 of which contained 10 or more members, with the largest cluster (#1) containing 94 repeat sequences. Some clusters contained repeats from organisms as distantly related as archaea and bacteria, supporting the inference that CRISPR/CAS systems can be horizontally transferred between microorganisms (Godde and Bickerton 2006; Haft et al. 2005; Makarova et al. 2006).

As an independent measure for the validity of the clustering, we examined the RNA stability scores in each of the MCL-defined clusters (note that RNA stability was not taken into account in the clustering procedure). As seen in Figure 1B, clusters #2 and #3 comprise repeats with consistently high folding scores, indicating pronounced secondary structure. By contrast, clusters #1, #6, #7, #9, #10 and #11 contain repeats with consistently poor folding scores. Clusters #4, #5, #8 and #12 show intermediate folding scores, suggesting weak secondary structures. Together, these groups explain the trimodal distribution observed in Figure 1A. The homogeneity of RNA structure stability scores within each cluster, along with the dramatic difference in scores between clusters, suggests that our clustering method is valid.

To further explore the observation that repeats form stable RNA secondary structures, we examined sequence alignments of the repeat clusters. CRISPR repeats are generally considered to be highly dissimilar to each other (Godde and Bickerton 2006), except for similar repeats in strains of the same species or in closely related species (Mojica et al. 1995). However, repeats within the clusters we generated, although often containing sequences from vastly different phylogenetic groups, were generally more similar to each other and hence alignable. Figure 2A presents a multiple alignment of a subset of the repeats in cluster #3. A highly stable stem-loop structure was consistently predicted for repeats in this cluster by RNAfold (Hofacker et al. 1994), [Fig 1B]. Notably, substitutions in the predicted stem structure are consistently accompanied by compensatory changes that preserve the base pairing [Fig 2A]. This mutational pattern, together with the presence of G:U base pairs [Fig. 2A], is typical of conserved RNA secondary structures and highlights the importance of the stem-loop in the repeats for the functionality of CRISPRs.

A summary of the repeat similarity space is presented in Figure 3. As with cluster #3 [Fig 2], repeats in other clusters with high and intermediate folding scores also form stem-loop structures [Fig 3], and display compensatory mutations. While the stem-loop motif is seen in all of these clusters, the actual sequence, as well as the length of the stem, its position relative to the unstructured region, and the size of the unstructured sequence varies between clusters. For example, while the stem in cluster #4 is typically 5 bp long and is found in the middle of the repeat, the stem in cluster #3 is typically 7 bp long, and is found towards the 5' end of the repeat [Figs 2&3]. Notably, most repeat clusters have a conserved 3' terminus of GAAA(C/G) possibly acting as a binding site for one of the conserved CAS proteins.

Some clusters, such as #2, #3 and #4, were discrete in the sequence similarity space, whereas the boundaries of other clusters such as #1, #6 and #7 were not clearly defined. The discrete clusters were generally comprised of structure-forming repeats, and the less well-defined clusters were comprised of unstructured repeats. This may be a reflection of the evolutionary constraints on the stem structure.

Two recent studies identified between 20 and 45 gene families of CRISPR-associated sequences (CAS) (Haft et al. 2005; Makarova et al. 2006). Based on the tendency of CAS genes to appear together, Haft et al (Haft et al. 2005) defined eight CAS subtypes (named Ecoli, Ypest, Nmeni, Dvulg, Tneap, Hmari, Apern and Mtube). We sought to determine whether our CRISPR repeat clusters corresponded to particular CAS subtypes. For this, we searched 20kb of sequence flanking each side of the repeat array for CAS genes using the 45 CAS families TIGRFAM HMMs defined by (Haft et al. 2005).

We found that the Ecoli CAS subtype genes appear exclusively in the proximity of structured repeat cluster #2, and, similarly, the Dvulg and Ypest CAS subtypes strictly correspond to our structured clusters #3 and #4, respectively [Tables 1 & S1]. Presumably, specific and different sets of genes are needed in order to recognize, bind and process the different repeat types. Despite the overall pronounced correspondence between the CAS subtypes and repeat clusters, particularly for structured clusters, there

are notable exceptions. For example, the reported frequent cooccurence of the Mtube subtype with other CAS subtypes (Haft et al. 2005) is consistent with its promiscuous association with numerous repeat clusters (Table 1). Another interesting exception is the cooccurence of the Tneap and Apern subtypes in the *Thermococcus kodakaraensis* genome with cluster #6, which is apparently due to a fusion of the Tneap and Apern subtypes [Fig. S1, Table S1]. This genome contains 3 CRISPR arrays, all with identical repeat sequences classified as cluster #6 [Table S1]. In some cases the CAS subtype for one or more repeat cluster members differs from the consensus for that cluster [Table S1]. This suggests that the association between CRISPR repeat subtypes and CAS subtypes is somewhat flexible.

We also identified a repeat cluster (#5) that is not associated with any of the recognized CAS subtypes. We found that it is associated with most of the core CAS (cas1-4 and cas6), but lacks any of the additional type-defining genes. Cluster #5 occurs exclusively in genomes that contain other CRISPR repeat subtypes and it is possible that it employs at least part of their CAS machinery.

Discussion

This study shows that CRISPR repeats are not structurally homogeneous and can be divided into distinct types based on sequence similarity and ability to form stable secondary structures. This explains why previous attempts to align all repeats resulted in a poorly defined consensus sequence (Godde and Bickerton 2006). We observed compensatory base changes in the stems of the structured repeat clusters including G:U base pairs, indicating that the CRISPR system likely functions through an RNA intermediate.

The inference of stem-loop formation within individual CRISPR repeats is in contrast to the speculation that pairs of repeats form duplexes, and are subsequently cleaved to release spacers (Makarova et al. 2006). Such hypothesized duplexing would unlikely require the ubiquitous presence of the less conserved interior nucleotides, which would form a loop in the single repeat folding model [Fig 2] and an unpaired bulge in the duplex repeat folding model. The folding of individual repeats is also supported by the observation that CRISPR arrays in Archaea are transcribed and processed into non-messenger RNAs that contain a single repeat and spacer (Tang et al. 2002; Tang et al. 2005). The repeat stem-loop may be specifically recognized (Cusack 1999) by RNA-binding CAS-encoded proteins.

A previous report suggested that spacer regions contribute to the formation of secondary structures in CRISPR arrays (Makarova et al. 2006). However, we could not detect a significant deviation of spacer secondary structures from random sequences [Fig 1], indicating that spacers are unlikely to be selected based on their secondary structure. In fact, the spacers appear to have slightly weaker structures than random sequences. This is probably due to the AT richness of spacers (46% GC) relative to average bacterial genomic sequences (53% GC), as AT base pairs form less stable structures than GC pairs. The lower spacer GC content is consistent with a proposed viral origin of spacer

sequences (Pourcel et al. 2005), as viruses are on average 7% lower in GC content than bacteria.

Previous attempts to classify CRISPR/CAS systems were based on CAS gene content and phylogeny (mostly of cas1) (Haft et al. 2005; Makarova et al. 2006). We add a further dimension to this classification by showing that the repeat sequence itself is also a classifying feature. This can be advantageous in instances where CRISPR arrays occur in the absence of CAS genes. For example, *Thermoplasma acidophilum* contains a CRISPR array but lacks CAS genes (Haft et al. 2005), so it cannot be classified based on CAS. Our clustering indicates that the *T. acidophilum* repeat belongs to (Euryarchaeal） cluster #6 (Fig. 3, Table S1). In some instances, the repeat classification was able to provide higher resolution to the existing CAS classification. For example, the Nmeni subtype was reported to have an optional gene csn2 (Haft et al. 2005). Our clustering divides this subtype to 3 clusters (#10, #16 and #22). The csn2 gene is invariably present in one cluster (#10) and absent in the other two. The finding of a repeat cluster (#5) that cannot be readily resolved by associated CAS genes (see Results) further demonstrates the power of CRISPR-based classification.

The significant differences between CRISPR/CAS subtypes, both in CRISPR repeat sequence and structure, and in CAS gene content and phylogeny, raises the possibility that these systems also differ functionally - e.g., in their specificity for different types of invading extrachromosomal DNA. Support for this hypothesis could be the fact that frequently several CRISPR/CAS subtypes are found in the same genome and at least two functions have been hypothesized for these elements (chromosome segregation (Mojica et al. 1995) and destruction of foreign DNAs (Mojica et al. 2005)). The study of CRISPRs is in its infancy, and their mode and function is still highly speculative. Our results provide another step toward a comprehensive understanding of these intriguing elements.

Acknowledgments

This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48, Lawrence Berkeley National Laboratory under contract No. DE-AC02-05CH11231 and Los Alamos National Laboratory under contract No. DE-AC02-06NA25396.

Methods

Identification of CRISPR arrays. All genome sequences available through the IMG database version 1.50 (Markowitz et al. 2006) were analyzed for CRISPR arrays using the PILERCR program [http://www.drive5.com/pilercr/].

Delineation of repeat clusters. Pairwise similarities between repeats was calculated using an in-house implementation of the Smith-Waterman algorithm (Smith and Waterman 1981). The best scoring similarity from the two possible repeat pair orientations, and only scores >7, were used for further analysis. Clustering of pairwise similarities was performed using the MCL program with default parameters (Van Dongen

2000). Multiple alignments were performed using muscle (Edgar 2004). Sequence logos for each cluster were generated using WebLogo (Crooks et al. 2004). The similarity space of repeats was visualized using BioLayout Java (Goldovsky et al. 2005).

Determination of repeat secondary structures. Structural predictions were performed using the RNA Vienna Package (Mathews et al. 1999) downloaded from [http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi] (Hofacker 2003). Folding scores for all repeats or individual repeat clusters were divided into bins of 2 score units and plotted as percentages. Random sequence strings with the same length distribution as repeats were generated from the analyzed genomes. The average GC contents were calculated for archaeal, bacterial and viral genomes in the IMG database ver. 1.50, and the average GC content was calculated for all spacers in all genomes.

CAS gene identification. The HMMs for CAS genes described in (Haft et al. 2005) were obtained from the TIGRFAM database version 6.0 [http://www.tigr.org/TIGRFAMs/]. To identify CAS genes, all coding sequences within 20 Kb of the identified CRISPR arrays were searched with the CAS HMMs using hmmpfam [http://hmmer.janelia.org/] with the thresholds of an e-value <0.001 and a positive score.

References:

Bolotin, A., B. Quinquis, P. Renault, A. Sorokin, S.D. Ehrlich, S. Kulakauskas, A. Lapidus, E. Goltsman, M. Mazur, G.D. Pusch, M. Fonstein, R. Overbeek, N. Kyprides, B. Purnelle, D. Prozzi, K. Ngui, D. Masuy, F. Hancy, S. Burteau, M. Boutry, J. Delcour, A. Goffeau, and P. Hols. 2004. Complete sequence and comparative genome analysis of the dairy bacterium Streptococcus thermophilus. *Nat Biotechnol* **22:** 1554-1558.

Crooks, G.E., G. Hon, J.M. Chandonia, and S.E. Brenner. 2004. WebLogo: a sequence logo generator. *Genome Res* **14:** 1188-1190.

Cusack, S. 1999. RNA-protein complexes. *Curr Opin Struct Biol* **9:** 66-73.

Edgar, R.C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32:** 1792-1797.

Godde, J.S. and A. Bickerton. 2006. The repetitive DNA elements called CRISPRs and their associated genes: evidence of horizontal transfer among prokaryotes. *J Mol Evol* **62:** 718-729.

Goldovsky, L., I. Cases, A.J. Enright, and C.A. Ouzounis. 2005. BioLayout(Java): versatile network visualisation of structural and functional relationships. *Appl Bioinformatics* **4:** 71-74.

Haft, D.H., J. Selengut, E.F. Mongodin, and K.E. Nelson. 2005. A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes. *PLoS Comput Biol* **1:** e60.

Hofacker, I.L. 2003. Vienna RNA secondary structure server. *Nucleic Acids Res* **31:** 3429-3431.

Hofacker, I.L., W. Fontana, P.F. Stadler, S. Bonhoeffer, M. Tacker, and S. P. 1994. Fast Folding and Comparison of RNA Secondary Structures. *Monatshefte f. Chemie* **125:** 167-188.

Jansen, R., J.D. Embden, W. Gaastra, and L.M. Schouls. 2002. Identification of genes that are associated with DNA repeats in prokaryotes. *Mol Microbiol* **43:** 1565-1575.

Makarova, K.S., L. Aravind, N.V. Grishin, I.B. Rogozin, and E.V. Koonin. 2002. A DNA repair system specific for thermophilic Archaea and bacteria predicted by genomic context analysis. *Nucleic Acids Res* **30:** 482-496.

Makarova, K.S., N.V. Grishin, S.A. Shabalina, Y.I. Wolf, and E.V. Koonin. 2006. 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. *Biol Direct* **1:** 7.

Markowitz, V.M., N. Ivanova, K. Palaniappan, E. Szeto, F. Korzeniewski, A. Lykidis, I. Anderson, K. Mavrommatis, V. Kunin, H. Garcia Martin, I. Dubchak, P. Hugenholtz, and N.C. Kyrpides. 2006. An experimental metagenome data management and analysis system. *Bioinformatics* **22:** e359-367.

Mathews, D.H., J. Sabina, M. Zuker, and D.H. Turner. 1999. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J Mol Biol* **288:** 911-940.

Mojica, F.J., C. Diez-Villasenor, J. Garcia-Martinez, and E. Soria. 2005. Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *J Mol Evol* **60:** 174-182.

Mojica, F.J., C. Ferrer, G. Juez, and F. Rodriguez-Valera. 1995. Long stretches of short tandem repeats are present in the largest replicons of the Archaea Haloferax mediterranei and Haloferax volcanii and could be involved in replicon partitioning. *Mol Microbiol* **17:** 85-93.

Pourcel, C., G. Salvignol, and G. Vergnaud. 2005. CRISPR elements in Yersinia pestis acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. *Microbiology* **151:** 653-663.

Sebaihia, M., B.W. Wren, P. Mullany, N.F. Fairweather, N. Minton, R. Stabler, N.R. Thomson, A.P. Roberts, A.M. Cerdeno-Tarraga, H. Wang, M.T. Holden, A. Wright, C. Churcher, M.A. Quail, S. Baker, N. Bason, K. Brooks, T. Chillingworth, A. Cronin, P. Davis, L. Dowd, A. Fraser, T. Feltwell, Z. Hance, S. Holroyd, K. Jagels, S. Moule, K. Mungall, C. Price, E. Rabbinowitsch, S. Sharp, M. Simmonds, K. Stevens, L. Unwin, S. Whithead, B. Dupuy, G. Dougan, B. Barrell, and J. Parkhill. 2006. The multidrug-resistant human pathogen Clostridium difficile has a highly mobile, mosaic genome. *Nat Genet* **38:** 779-786.

Smith, T.F. and M.S. Waterman. 1981. Identification of common molecular subsequences. *J Mol Biol* **147:** 195-197.

Tang, T.H., J.P. Bachellerie, T. Rozhdestvensky, M.L. Bortolin, H. Huber, M. Drungowski, T. Elge, J. Brosius, and A. Huttenhofer. 2002. Identification of 86 candidates for small non-messenger RNAs from the archaeon Archaeoglobus fulgidus. *Proc Natl Acad Sci U S A* **99:** 7536-7541.

Tang, T.H., N. Polacek, M. Zywicki, H. Huber, K. Brugger, R. Garrett, J.P. Bachellerie, and A. Huttenhofer. 2005. Identification of novel non-coding RNAs as potential antisense regulators in the archaeon Sulfolobus solfataricus. *Mol Microbiol* **55:** 469-481.

Van Dongen, S. 2000. Graph Clustering by Flow Simulation. University of Utrecht.

Figure legends

Figure 1. Distributions of folding scores of (A) all CRISPR repeats and all spacers, as compared to random sequences and (B) individual repeat clusters. X-axis, negative folding scores; Y-axis, fraction (percent) of total.

Figure 2. Evidence for secondary structure in cluster #3. (A) Multiple alignment of a subset (for clarity) of repeats in cluster #3. Numbers 1 to 7 and 7 to 1 indicate the residues involved in stem base-pairing, some compensatory mutations in the stem are highlighted with circles. Note G:U base pairing at position 5 in *Xanthomonas oryzae* and relaxed conservation of loop residues typical of RNA secondary structure in which the structure is functional rather than the sequence. (B) Sequence logo for all repeats in cluster #3. (C) Predicted secondary structure of *Syntrophus acidotrophicus* repeat using RNAfold. Stem positions are numbered in accordance with the alignment.

Figure 3. The sequence similarity space of CRISPR repeats. Dots denote individual repeat sequences; distances between dots represent Smith-Waterman similarities, such that closer dots represent more similar sequences. Dot colors denote cluster association as derived from MCL clustering. The 12 largest clusters are indicated by circles together with their sequence logos, coarse phylogenetic composition, and sample secondary structures where applicable.

Figure S1. Arrangement of the CAS cassette in the *Thermococcus kodakaraensis* genome. Chromosomal coordinates are given at the top of the figure. A CRISPR array is shown to the left of the figure as red vertical lines (1 line = 5 repeats). Core cas genes are shown in black, Apern subtype genes are shown in blue and Tneap subtype genes in red as predicted by TIGRFAM analysis (see methods).

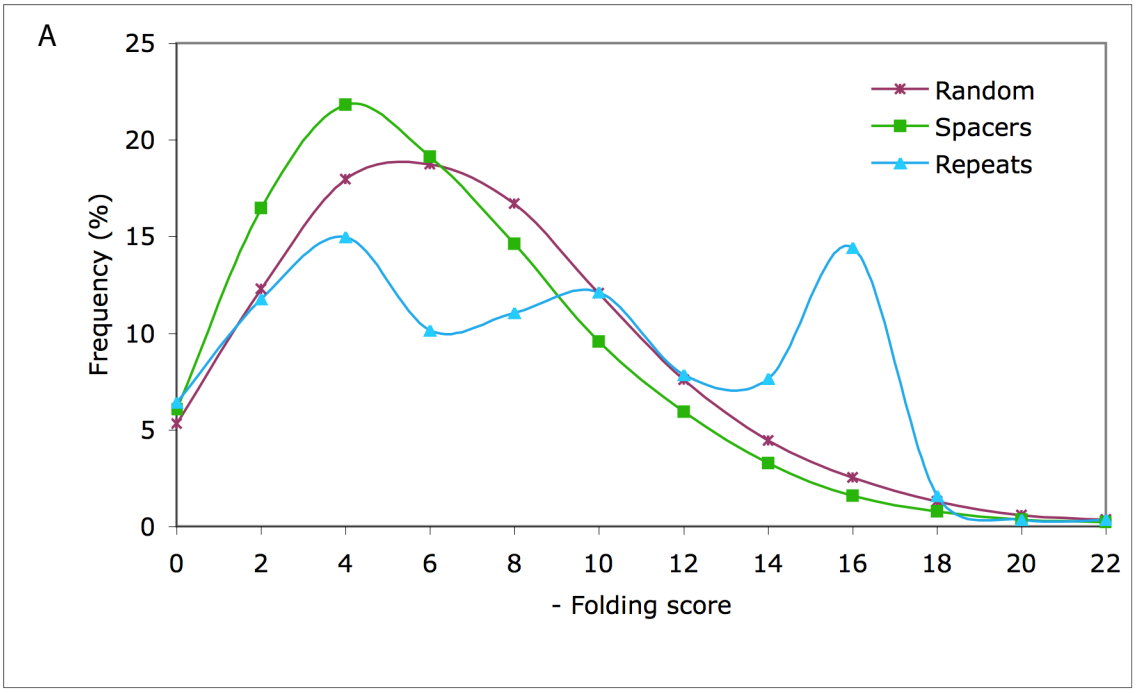
Tables

Table 1. Occurrence of CAS subtypes (Haft et al. 2005) in the proximity (+/-20kb) of the twelve largest repeat clusters. Associations are indicated by an X. An instance of a putative fusion between two CAS subtypes is indicated by an F.

CAS	***Repeat cluster***											
subtype	***#1***	***#2***	***#3***	***#4***	***#5***	***#6***	***#7***	***#8***	***#9***	***#10***	***#11***	***#12***
Ecoli		X										
Ypest				X								
Nmeni										X		
Dvulg			X									
Tneap	X					X						
Hmari	X								X			
Apern						F	X				X	
Mtube	X					X		X				X

Table S1. CAS genes in the neighborhood of CRISPR arrays, as predicted by TIGRFAM (see methods). Core and type-specific genes are indicated, each genome is given both with its full name and an IMG accession. IMG gene OIDs are given for each protein.

Figure 1


A
Frequency (%)
- Folding score
Random
Spacers
Repeats

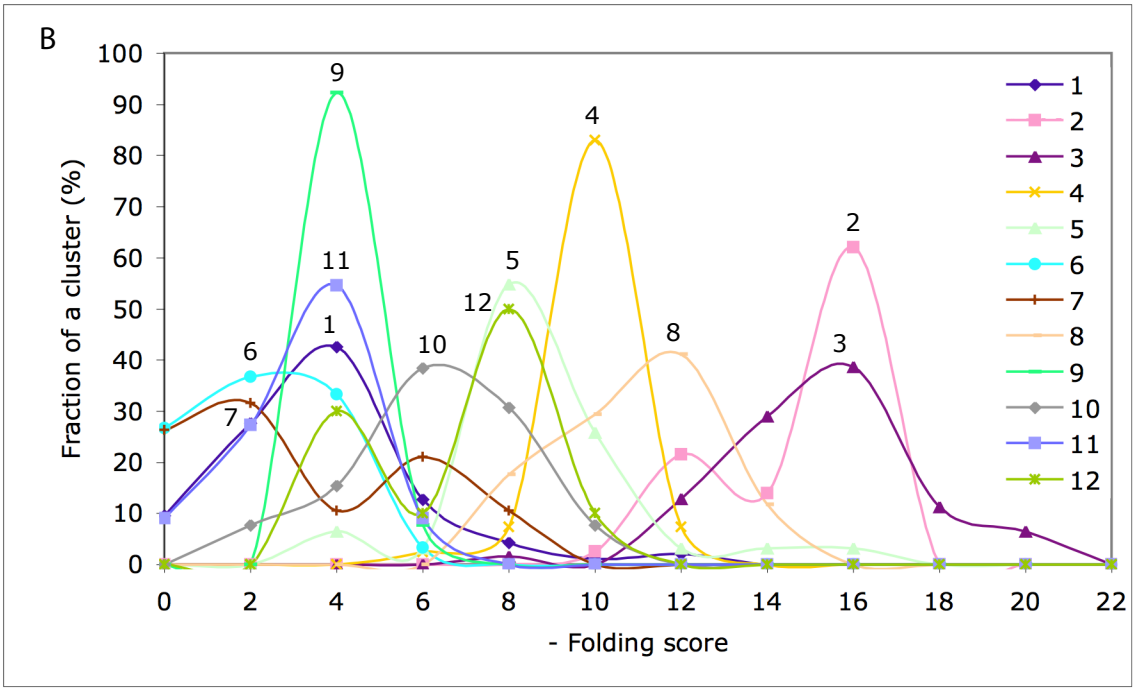

B
Fraction of a cluster (%)
- Folding score
1
2
3
4
5
6
7
8
9
10
11
12

Figure 2

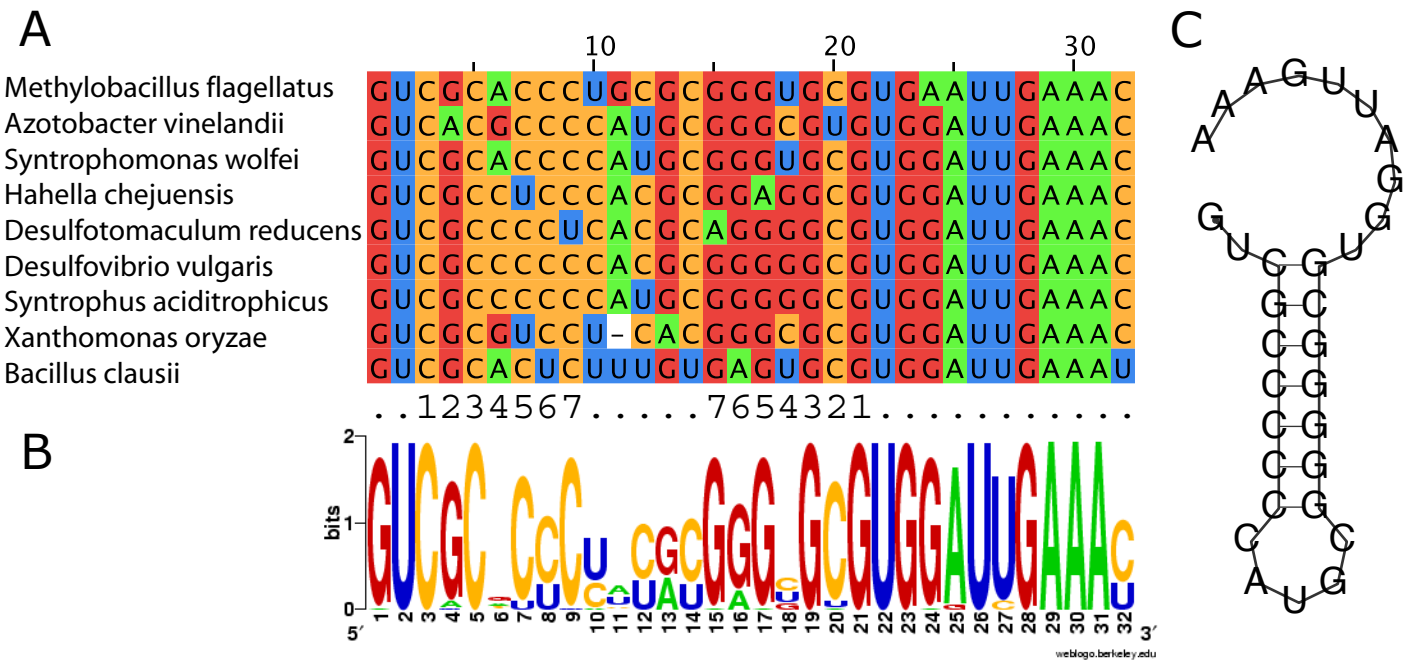
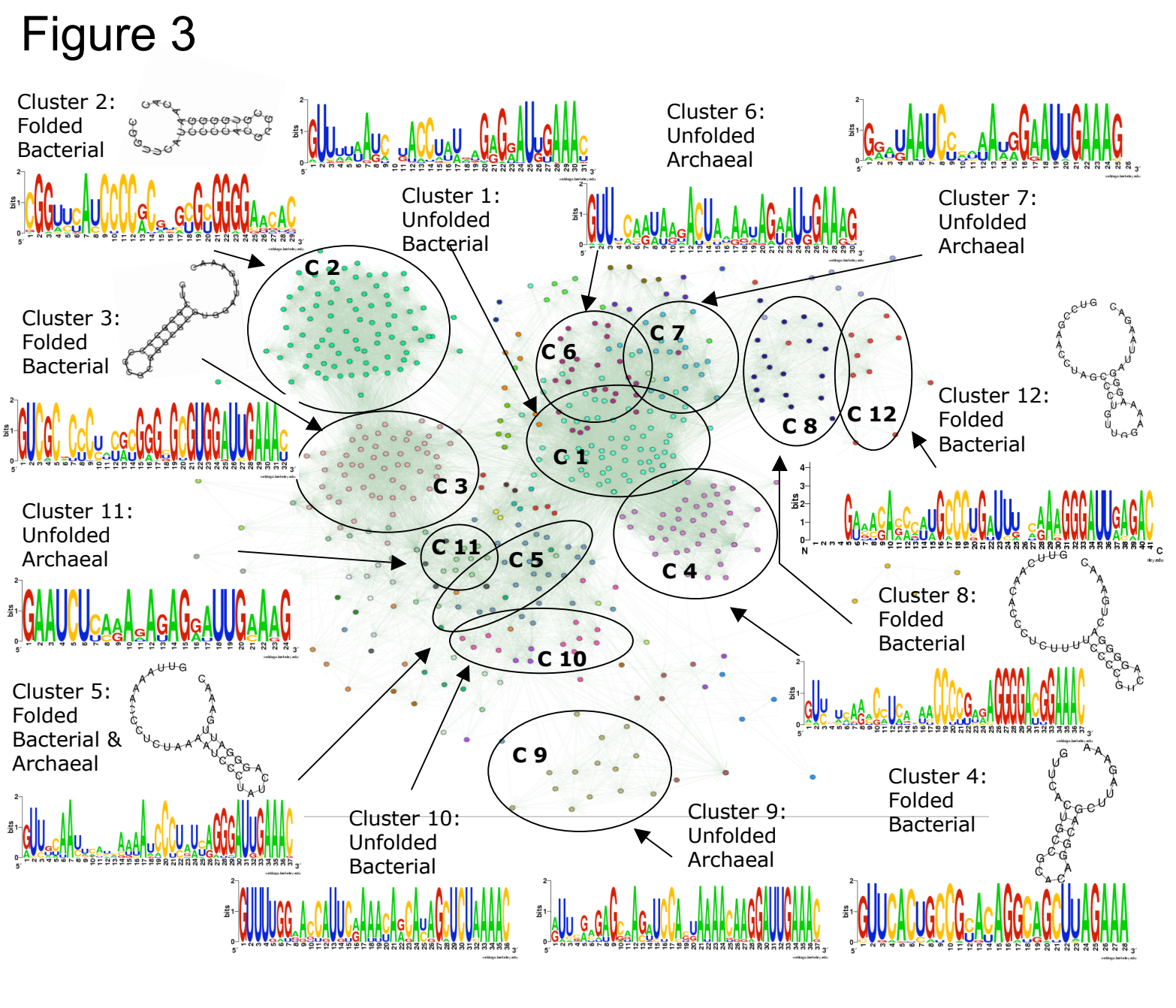

A
10
20
30
Methylobacillus flagellatus
Azotobacter vinelandii
Syntrophomonas wolfei
Hahella chejuensis
Desulfotomaculum reducens
Desulfovibrio vulgaris
Syntrophus aciditrophicus
Xanthomonas oryzae
Bacillus clausii
..1234567.....7654321..........
B
bits
5′
3′
weblogo.berkeley.edu
C

Figure 3


Cluster 2:
Folded
Bacterial
Cluster 1:
Unfolded
Bacterial
Cluster 6:
Unfolded
Archaeal
Cluster 7:
Unfolded
Archaeal
Cluster 3:
Folded
Bacterial
Cluster 12:
Folded
Bacterial
Cluster 11:
Unfolded
Archaeal
Cluster 8:
Folded
Bacterial
Cluster 5:
Folded
Bacterial &
Archaeal
Cluster 4:
Folded
Bacterial
Cluster 10:
Unfolded
Bacterial
Cluster 9:
Unfolded
Archaeal
C 1
C 2
C 3
C 4
C 5
C 6
C 7
C 8
C 9
C 10
C 11
C 12