Chapter 17

Large-Scale Identification and Analysis of C-Proteins

Valery Sorokin, Konstantin Severinov, and Mikhail S. Gelfand

Abstract

The restriction-modification system is a toxin–antitoxin mechanism of bacterial cells to resist phage attacks. High efficiency comes at a price of high maintenance costs: (1) a host cell dies whenever it loses restriction-modification genes and (2) whenever a plasmid with restriction-modification genes enters a naïve cell, modification enzyme (methylase) has to be expressed prior to the synthesis of the restriction enzyme (restrictase) or the cell dies. These phenomena imply a sophisticated regulatory mechanism. During the evolution several such mechanisms were developed, of which one relies on a special C(control)-protein, a short autoregulatory protein containing an HTH-domain. Given the extreme diversity among restriction-modification systems, one could expect that C-proteins had evolved into several groups that might differ in autoregulatory binding sites architecture. However, only a few C-proteins (and the corresponding binding sites) were known before this study. Bioinformatics studies applied to C-proteins and their binding sites were limited to groups of well-known C-proteins and lacked systematic analysis. In this work, the authors use bioinformatics techniques to discover 201 C-protein genes with predicted autoregulatory binding sites. The systematic analysis of the predicted sites allowed for the discovery of 10 structural classes of binding sites.

Key words: Restriction-modification systems, C-proteins, DNA-binding proteins, bioinformatics, transcription regulation.

1. Introduction

1.1. Restriction-Modification Systems

The restriction-modification (RM) phenomenon was first discovered more than 50 years ago during the studies of bacterial antiphage defense (1, 2). Of the three types of RM systems, type II is the simplest and most prevalent. Restriction endonucleases encoded by type II systems are widely used in molecular cloning. A typical type II RM system is essentially a two-component

I. Ladunga (ed.), *Computational Biology of Transcription Factor Binding*, Methods in Molecular Biology 674,
DOI 10.1007/978-1-60761-854-6_17, © Springer Science+Business Media, LLC 2010

toxin–antitoxin system. A type II restriction endonuclease (toxin) cleaves unprotected or unmodified DNA at specific sites triggering DNA degradation and cell death. A methyltransferase (antitoxin) protects DNA from cleavage by methylating the same DNA sites (3, 4). Most RM systems' genes are plasmid borne and capable of horizontal spread through bacterial populations. A bacterial cell that acquires an RM system becomes resistant to infection by phages whose genomes contain unmethylated sites in their DNA. On the other hand, plasmids harboring RM systems' genes behave like selfish genetic elements, because a loss of an RM plasmid leads to cell death since methyltransferase (antitoxin) is shorter living than the corresponding restriction endonuclease (toxin).

The horizontal transfer of an RM system among naïve (i.e., lacking such genes) bacteria imposes constrains on RM genes expression. Since the corresponding genomic DNA sites in a naïve cell are unmethylated and therefore vulnerable to cleavage by the restriction endonuclease, the methylase must be synthesized first, while the appearance of restriction endonuclease activity must be delayed until all sites are methylated. On the other hand, once an RM system is established in the host cell, a steady-state ratio of restriction endonuclease and methyltransferase activity needs to be maintained. Too low endonuclease activity (or excessive methyltransferase activity) could lead to a loss of protection against bacteriophage infection and cell (and RM plasmid) death.

One of the most prevalent regulation strategies involves a dedicated transcription regulator encoded by a separate gene. These regulators are called control(C)-proteins and they influence the level of the endonuclease gene (and sometimes methylase gene) transcription in many RM systems. Since computational analysis of this type of regulation relies on the standard structure of C-containing RM-loci, we shall describe them in more detail.

1.2. Regulation of Transcription by C-Proteins

C-protein-dependent regulation was first described for the *Pvu*II system (5). The X-ray analysis of the C.*Ahd*I and C.*Bcl*I C-proteins revealed a 5-alpha-helical protein, which could be assigned to the Xre family of transcriptional regulators (6, 7). Of the five alpha helices, two represent a typical helix-turn-helix (HTH) domain. The remaining three alpha helices allow for effective dimerization of the protein. The similarity between C-proteins and the Xre family regulators (**Fig.** 17.1) suggests that like the latter, C-proteins activate transcription by directly interacting with the σ^{70} subunit of RNA polymerase.

The endonuclease gene is usually localized immediately downstream of the C-protein gene, forming a single CR (C-protein-restriction endonuclease) transcription unit (8). The methylase gene constitutes a separate transcription unit, either convergent or divergent with respect to the CR unit. In all

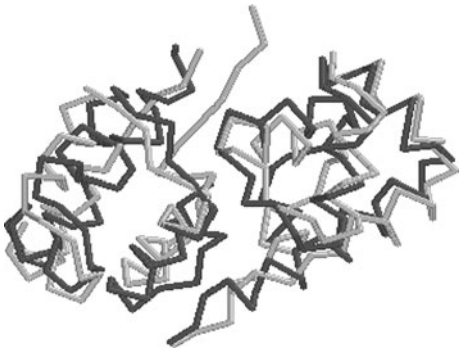

Fig. 17.1. Overlaid structures and protein alignment of C-protein C.*Ahd*I (*black*) and cytosine regulator CylR2 (*gray*), an Xre-subfamily protein.

experimentally studied systems, the region upstream of the C-protein gene contains two C-protein binding sites. The distal site (located further from the C-protein translation start site) has a higher affinity. A weak promoter is usually positioned downstream of the two binding sites. Thus, the basal level of the CR unit transcription is low. After a sufficient amount of C-protein has been produced, the C-protein dimerizes and binds to the distal site activating the CR promoter. Upon further C-protein accumulation, a second dimer binds to the weaker proximal site and represses the further transcription of the CR unit. These positive and negative feedback loops allow for maintaining a steady level of the endonuclease transcript (9, 10).

In some RM systems, C-proteins also affect the transcription of the methylase gene. One (indirect) way of doing this operates when the *M* and *CR* transcription units are divergent. The distance between the transcription units is so small that the methylase gene promoter overlaps with the C-protein binding sites. Upon binding to the distal site, the C-protein dimer prevents the RNA polymerase from binding to the methylase promoter. This regulatory mechanism is operational in the *Eco*RV RM system (11).

Another direct way of affecting the methylase gene transcription has been recently described for the *Esp*1396I RM system (12), where the *M* and *CR* transcription units are convergent. A single, high-affinity C-protein binding site was found upstream of the methylase gene. Because of the high affinity, the C-protein dimer binds to this site early on, repressing the methylase gene transcription. As larger quantities of C-protein are produced, the C-protein binds first to the distal and then to the proximal sites located upstream of the C-protein gene, causing activation then repression of endonuclease gene transcription as described above. The schema of genetic organization of several C-protein-dependent RM system loci is shown in **Fig.** 17.2.

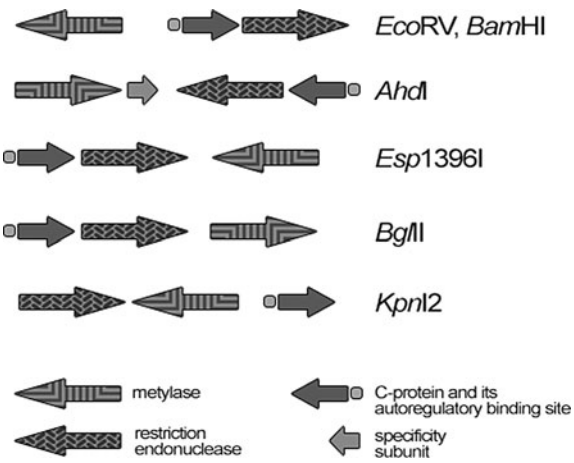

Fig. 17.2. The genetic organization of several known C-protein-dependent RM systems.

1.3. Computational Analysis of Transcription Regulation by C-Proteins

C-protein binding sites were first reported in 1995 (13). The analysis of upstream sequences of six then known C-protein genes revealed a conserved 12-bp region with the consensus sequence of ACTTATAGTCTG, later extended to 18 bp (aCTYATaGTCYGTNGNYt) (14). Newly discovered C-proteins changed the binding site motif (15). The authors argued that the 18-bp sites were unnecessarily long for monomer binding while lacking the dyad symmetry expected for dimer-binding sites. Based on these considerations, short binding sites (C.*Sma*I: AATGCTACT; C.*Nme*SI: TGCTACTTATAG; C.*Bgl*II: GATACTTATAGTC) were proposed. These sites were considered to interact with C-protein monomers.

As of July 2007, the main repository of RM systems, Rebase, contained as few as 48 C-proteins and 8 confirmed C-protein binding sites. The binding sites formed three structural groups, each named after its archetypical representative, with two groups containing only one member. The largest group with six binding sites was named after C.*Pvu*II. It contained binding sites that lacked a pronounced palindromic structure expected for the dimeric form of C-proteins. The remaining binding sites of the C.*Eco*RV and C.*Eco*O109I group were palindromic.

The C.*Pvu*II-like group of C-protein binding sites was subsequently extended to include 24 binding sites for known C-proteins from Rebase (16). A typical C.*Pvu*II-like site was found to have a complex structure with a highly conserved tetranucleotide between two copies of the palindrome. Each palindrome arm is called a C-box, and the entire site thus contains four C-boxes. Interestingly, the proximal (3′) palindrome was less similar to the consensus, i.e., was "weaker"

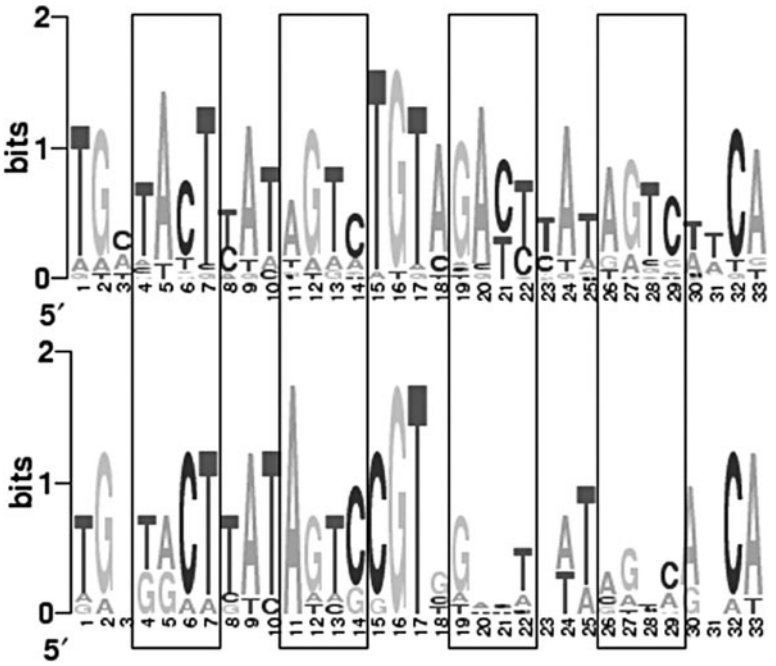

Fig. 17.3. Sequence logo showing the structure of C.*Pvu*II-like binding sites (16). Framed tetranucleotides represent C-boxes. A pair of C-boxes constitutes a palindrome. Additionally, the motifs contain highly conserved GT-core nucleotides and self-complementary dinucleotides at motif termini.

than the distal (5′) one, consistent with the above transcription regulatory mechanism. The observed palindromic structure (**Fig.** 17.3) likely reflects the dimeric form of C-proteins binding to the sites.

2. Methods

2.1. Large-Scale Identification and Analysis of C-Proteins

A systematic study of candidate C-proteins and their binding sites was reported in (17). All data related to the study can be found at http://iitp.bioinf.fbb.msu.ru/vsorokin.

Forty-six C-proteins from Rebase were used as queries in a BLAST (18) search against the non-redundant GenBank (19) nucleotide collection (tblastn, threshold: 1e-05, *see* **Note 1**). After manual curation, 245 unique hits were retained for further analysis.

The upstream sequences of candidate C-protein genes were analyzed in order to identify conserved regions that could correspond to the binding sites. Since there is no reason to expect a single conserved motif for all C-proteins, this was done separately in groups of closely related C-proteins. Specifically, a multiple

alignment of all 291 proteins (46 known Rebase C-proteins and 245 putative C-proteins) was built using MUSCLE (20) (default parameters) and the maximum likelihood tree was constructed using the PROML procedure from the PHYLIP (21) package using the default parameter settings (*see* **Note 2**). Examination of the resulting phylogenetic tree revealed several large, separate branches. A slightly reduced variant of the original tree, containing only those proteins for which the putative binding motifs could be identified (see below), is shown in **Fig.** 17.4.

2.2. Identification of Candidate C-Protein Binding Motifs

Each major branch of the C-protein phylogenetic tree was analyzed independently. Hundred-base pairs long sequences upstream of most closely related genes were aligned using

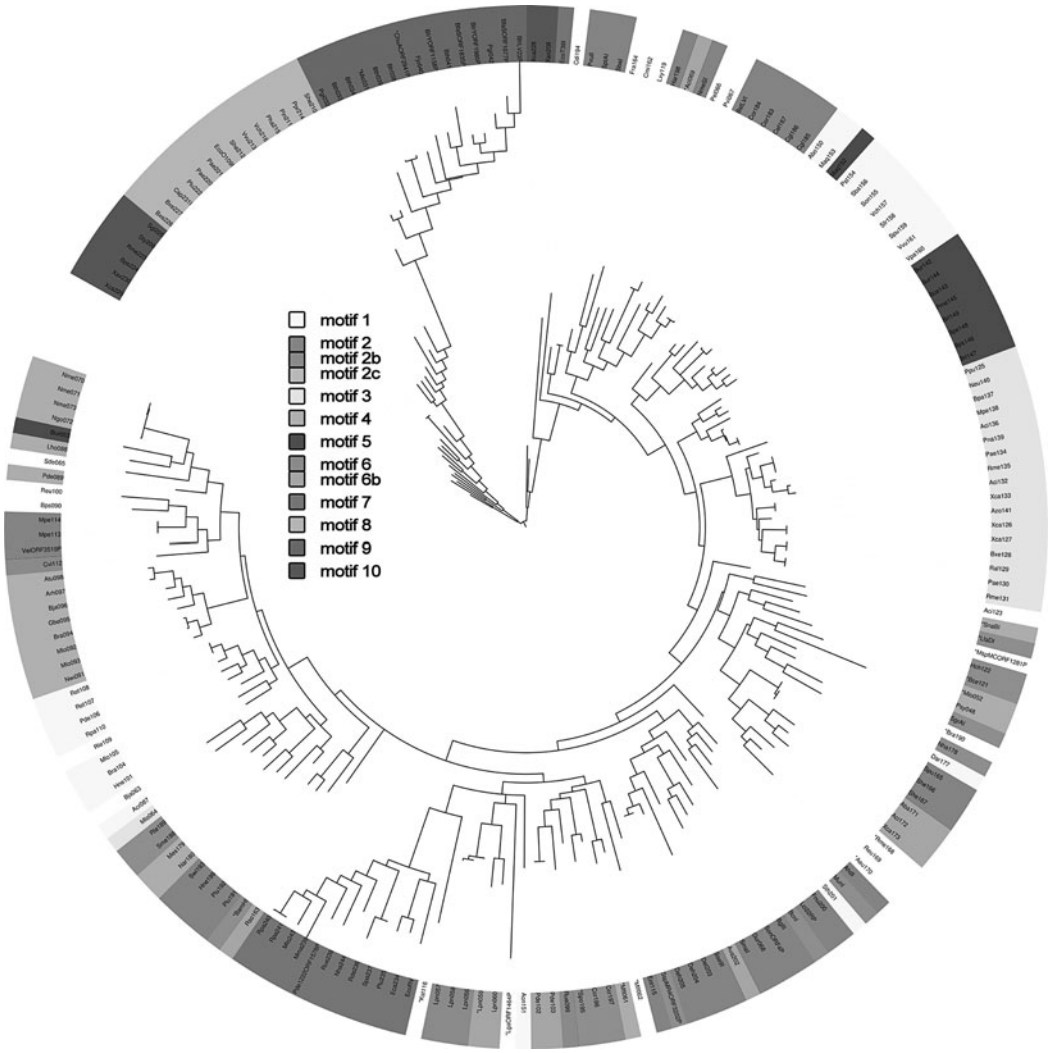

Fig. 17.4. The structure of motifs of predicted C-protein binding sites (17). Clustering was performed using ClusterTree-RS (22). C-boxes are *highlighted*, *arrows* represent palindromes formed by pairs of C-boxes.

MUSCLE (20) with the default parameter settings. If the alignment contained many highly conserved regions, the alignment was further extended by adding more distant members of the branch (*see* **Note 3**). The extension process terminated when alignments deteriorated completely. The resulting alignments were analyzed manually and the conserved islands that remained were considered to correspond to C-protein binding sites.

To identify additional binding sites, and, in particular, to account for a possibility of incorrectly annotated start codons of some predicted C-proteins, conserved sites obtained as described above were used to build HMM profiles (http://hmmer.wustl.edu, *see* **Note 4**), and the latter were used to scan by hmmer using default parameters for matches in regions from −100 to +50 relative to annotated start codons of all predicted C-protein genes. As expected, the majority of matches coincided with already predicted binding sites. However, several new matches were found and some matches were at a different location than the initially identified sites. A small overall number of corrections indicated the overall consistency of the prediction procedure. In total, 201 binding sites were predicted.

2.3. Validation of Predicted C-Protein Binding Sites

Candidate binding sites were predicted for 201 of the total of 291 C-proteins. As mentioned above, C-protein binding sites for eight Rebase RM systems have been identified experimentally and 24 more sites computationally (16). All these sites were present among the sites identified by our procedure.

2.4. The Structure of Binding Motifs

Previously known binding sites were assigned to one structural group named after its archetype C.*Pvu*II and two single-member groups: C.*Eco*RV and C.*Eco*O109I. The C.*Pvu*II-group sites had a structure of two short palindromes separated by highly conserved tetranucleotides (16). Two palindrome arms represent the so-called C-boxes, which are likely binding sites of individual C-protein monomers when they form dimers. Pairs of conserved, complementary positions outside the palindromes were disregarded in (16). Unlike the C.*Pvu*II-like sites, the binding sites of C.*Eco*RV and C.*Eco*O109I are single palindromes.

The set of newly predicted sites from (17) was split into clusters [ClusterTree-RS procedure (22)], which we will refer to as motifs. This procedure yielded 10 stable motifs (**Fig.** 17.5), which comprised 181 (90%) of the 201 predicted binding sites. While the remaining sites resembled motifs 1–6 (see below) the procedure failed to cluster them, probably due to the search stringency. The motifs logos (**Fig.** 17.5) are described below, while their main features are listed in **Table** 17.1.

Motifs 7 and 8

Motifs 7 and 8 correspond to the previously recognized C.*Eco*RV and C.*Eco*O109I groups, respectively. However, the

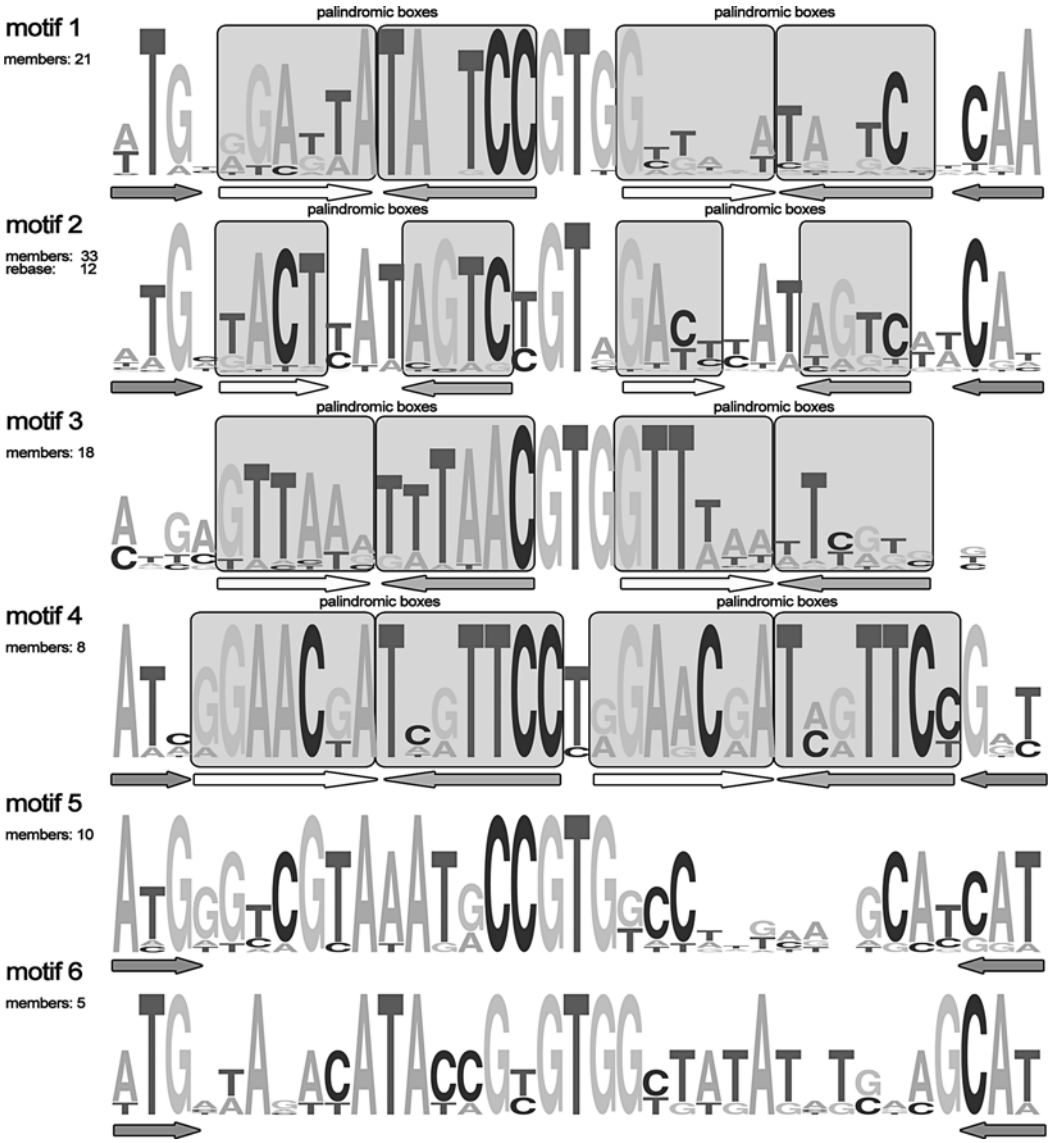

Fig. 17.5. The structure of motifs of predicted C-protein binding sites (17). Clustering was performed using ClusterTree-RS (22). C-boxes are highlighted. Arrows represent palindromes formed by pairs of C-boxes.

constructed motifs are longer due to additional conserved positions.

Motifs 1–6

Motifs 1–6 have the same length and share common structural features, described by the scheme

$$\text{Z-X-N-X}^{*}\text{-[GT-rich spacer]-x-n-x}^{*}\text{-Z}^{*},$$

where Z denotes conserved complementary trinucleotides, X represents palindrome-forming C-boxes, and asterisks denote

Table 17.1
The classification of C-proteins and their binding motifs

Group 1 C.PvuII-like sites	Group 2 Palindromic sites	Group 3 Possible false positives
Motifs 1–4 • Length: 35 bp • GT-rich central area • Double palindromes (four C-boxes) • Conserved terminal complementary trinucleotides (except motif 3)	Motifs 7, 8,10 • Single palindromes (two C-boxes) • Additional downstream palindromes • Form three separate, individual branches on the C-protein tree	Motif 9 • Unusually short • No pronounced palindromic structure • A separate branch on the C-protein tree
Motifs 5,6 • Length: 35 bp • GT-rich central area • No pronounced C-boxes • Conserved terminal complementary trinucleotides		

complementary elements. The uppercase X-N-X* indicates that the 5′-copy is much closer to the overall palindromic consensus than the 3′-copy. The conserved spacer between the copies is unique for each motif.

Motifs 1, 2, and 4 fit the above scheme exactly. Motif 3 lacks external trinucleotides. Motifs 5 and 6 lack the palindromic structure but retain the highly conserved complementary external trinucleotides. All previously identified C.*Pvu*II-like binding sites conform to the motif 2 structure.

Motif 9

Motif 9 is rather short (10 bp) but well conserved. It lacks any palindromic symmetry and reveals no other structural features. While these could be false positive, predicted binding sites for six C-proteins from Rebase belong to this motif. Hence, this prediction, while tentative, warrants experimental verification.

Motif 10

This single palindromic motif is not related to other motifs. Rebase contains no C-proteins predicted to bind to this motif. Again, this prediction requires experimental verification.

2.5. Downstream Sites

X-ray analysis revealed the dimeric nature of C-proteins C.*Ahd*I and C.*Bcl*I (7, 8). Indeed, all experimentally studied C-protein binding events involved paired binding sites. Activation required their interaction with the high-affinity promoter-distal site. This is followed by repression through interaction with the low-affinity promoter-proximal site of the CR transcription unit.

Thus, if the predicted sites are functional, additional weaker sites could be expected in close vicinity and downstream of predicted "single" sites.

Motifs 1–4 already consisted of two palindromes representing promoter-distal and promoter-proximal binding sites. Consistent with the existing model, the consensus of the proximal palindrome was weaker than the consensus of the distal palindrome.

Motifs 7, 8, and 10 include single palindromes. A special procedure was applied to search for additional sites downstream of these motifs. First, HMM profiles were constructed for each motif (*see* **Note 4**). Next, upstream sequences of C-protein genes containing sites forming a motif were searched for profile matches. As expected, a strong match coinciding with already predicted binding site was observed in all cases. Importantly, all "second best" matches were downstream of primary matches and were predicted to be the downstream binding sites. The consensus of proximal (downstream) sites was weaker than that of distal sites. This result agrees with the established model of C-protein transcription regulation involving activation upon C-protein binding to the upstream site and subsequent repression following binding to the downstream site.

Unlike the case for motifs 1–4, the distance varies between the distal and proximal copies for motifs 7, 8, and 10. This suggests

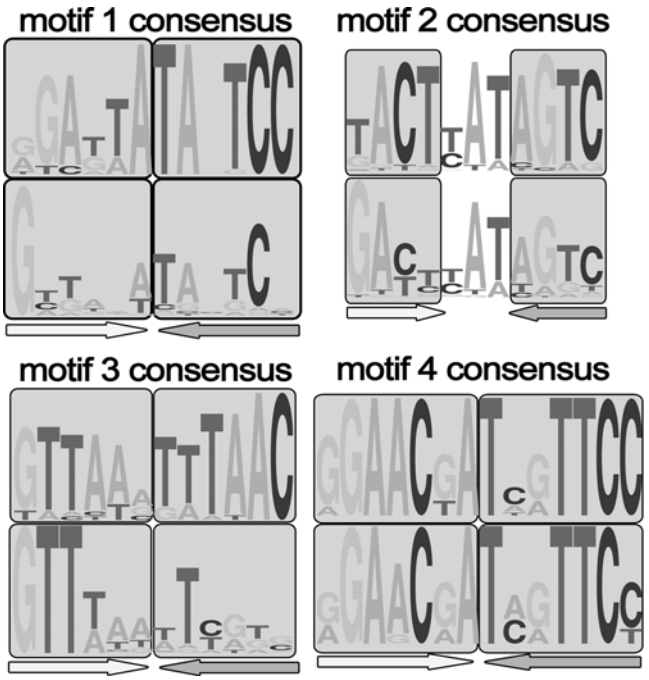

Fig. 17.6. Sequence logo of distal (*top*) and proximal (*bottom*) pairs of C-boxes from identified binding sites (17). Notation is as in **Fig**. 17.5.

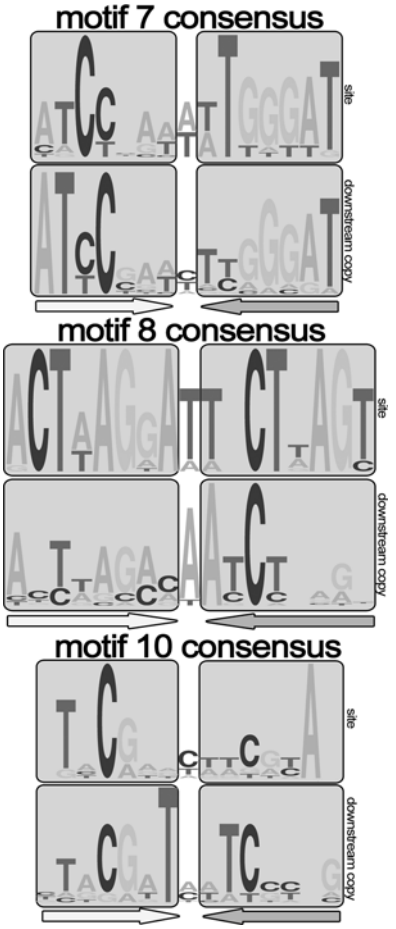

Fig. 17.6. (continued)

that C-protein dimers bound to these sites do not interact with each other, unlike C-proteins with constant distances.

No downstream copies were observed for C.*Pvu*II-like motifs 5 and 6 (lacking palindromic structure) as well as for non-palindromic motif 9. It is therefore unclear whether these binding sites, if real, operate using the same activation-repression mechanism as other C-proteins. The distal and proximal pairs of C-boxes are shown in **Fig.** 17.6.

3. Conclusions

Candidate C-proteins and predicted binding sites were consistently identified by a set of computational methods. Having started with known C-proteins from Rebase, we retrieved from

GenBank nucleotide sequences which could represent C-protein genes. A multi-step manual analysis was applied to remove possible false positives followed by the prediction of autoregulatory binding sites in upstream regions of C-protein genes. Manual analysis of the multiple alignments of closely related upstream sequences allowed for the identification of highly conserved islands, which were predicted to be the binding sites. Indeed, the predicted binding sites completely matched all known sites. The predicted binding sites were clustered into 10 motifs, of which only three had been known previously.

The work resulted in 201 predicted C-protein genes with predicted autoregulatory binding sites.

4. Notes

1. The use of such a liberal BLAST *e*-value threshold requires considerable care. On the other hand, since the typical C-protein length is only about 70 amino acids, stricter thresholds could yield loss of relevant hits. In (17), the authors controlled for several factors to avoid false positives. First, multiple alignment of all candidates was constructed and the presence of a pronounced HTH domain was verified. Second, manual analysis of hit annotations demonstrated that the set of candidates did not contain transcription factors with known, unrelated function. The third filter was the requirement of upstream regulatory sites, as described in the text.
2. The construction of phylogenetic trees requires high-quality multiple alignments. In case of distantly related proteins, computer-generated alignments need to be inspected manually. Alignments lacking a relatively conserved region (i.e., regions with a small number of mismatches and gaps) were refined as follows. The most distant sequences were removed until the conserved region appeared. Also, constructing a maximum likelihood tree of 300 proteins is a highly CPU-intensive task. If the purpose is to obtain a guide for iterative alignment of gene upstream regions, as here, the time may be decreased by temporarily removing very similar proteins. Subtrees for individual groups may then be constructed with complete data.
3. The "phylogenetic footprinting" approach (23) is used when candidate sites are expected to occur upstream of orthologous genes. This procedure is based on the assumption that the binding sites are more conserved than the

surrounding upstream regions. To achieve the sharpest contrast between the putative sites and the rest of the region, one needs to find the best set of sequences to be aligned. This is done heuristically by gradually adding more and more distant sequences until the alignment disintegrates. The alignment constructed at the last step before that is the one, where the conserved islands likely correspond to the binding sites. It is useful to constrain the alignment by retaining in the sequences to be aligned some part of the protein-coding region, which normally is sufficiently strongly conserved to be uniformly alignable at the nucleotide level.

Since for the RM-systems, due to frequent horizontal transfer, orthology relationships are challenging to establish, the evolutionary distance between the C-proteins (the regulators encoded by the regulated genes at the same time) was used as a guide for the progressive alignment of the upstream regions. The phylogenetic tree of C-proteins sets the order in which the upstream regions are considered. However, when the alignment starts to deteriorate, one should try to add several different sequences, since the exact topology of the tree is not very reliable.

4. Given the short extent of the binding sites (here, at most 35 bp), HMM profiles needs to be calibrated by the *hmmcalibrate* procedure. This allows one to increase precision of the *e*-value estimation when the e-value falls within the range of [1e-05, 1].

References

1. Bertani, G., Weigle, J.J. (1953) Host controlled variation in bacterial viruses. *J Bacteriol* 65, 113–121.
2. Lurlia, S.E., and Human, M.L. (1952) A nonhereditary, host-induced variation of bacterial viruses. *J Bacteriol* 64, 557–569.
3. Bickle, T.A., and Krueger, D.H. (1993) Biology of DNA restriction. *Microbiol Rev* 57, 434–450.
4. King, G., and Murray, N.E. (1994) Restriction enzymes in cells, not eppendorfs. *Trends Microbiol* 2, 465–469.
5. Knowle, D., Lintner, R., Touma, Y.M., and Blumenthal, R.M. (2005) Nature of promoter activated by C. PvuII, an unusual regulatory protein conserved among restriction-modification systems. *J Bacteriol* 187, 488–497.
6. Sawaya, M.R., Zhu, Z., Mersha, F. et al. (2005) Crystal structure of the restriction modification system control element C.Bcll and mapping of its binding site. *Structure* 13, 1837–1847.
7. McGeehan, J.E., Streeter, S.D., Papapanagiotou, I. et al. (2005) High-resolution crystal structure of the restriction-modification controller protein C.AhdI from *Aeromonas hydrophila*. *J Mol Biol* 346, 689–701.
8. Bart, A., Dankert, J., and van der Ende, A. (1999) Operator sequences for the regulatory proteins of restriction modification systems. *Mol Microbiol* 31, 1277–1278.
9. Bogdanova, E., Djordjevic, M., Papapanagiotou, I. et al. (2008) Transcription regulation of type II restriction-modification system AhdI. *Nucleic Acids Res* 36, 1429–1442.
10. Semenova, E., Minakhin, L., Bogdanova, E. et al. (2005) Transcription regulation of the EcoRV restriction modification system. *Nucleic Acids Res* 33, 6942–6951.
11. Zheleznaya, L.A., Kainov, D.E., Yunusova, A.K. et al. (2003) Regulatory C protein of

the EcoRV modification–restriction system. *Biochemistry (Moscow)* 68, 125–132.
12. Cesnaviciene, E., Mitkaite, G., Stankevicius, K. et al. (2003) Esp1396I restriction-modification system: structural organization and mode of regulation. *Nucleic Acids Res* 31, 743–749.
13. Rimseliene, R, Vaisvila, R, and Janulaitis, A. (1995) The eco72IC gene specifies a trans-acting factor which influences expression of both DNA methyltransferase and endonuclease from the Eco72I restriction-modification system. *Gene* 157, 217–219.
14. Anton, B.P., Heiter, D.F., Benner, J.S. et al. (1997) Cloning and characterization of the BglII restriction-modification system reveals a possible evolutionary footprint. *Gene* 187, 19–27.
15. Bart, A., Dankert, J., and van der Ende, A. (1999) Operator sequences for the regulatory proteins of restriction-modification systems. *Mol Microbiol* 31, 1275–1281.
16. Mruk, I., Rajesh, P., and Blumenthal, R.M. (2007) Regulatory circuit based on autogenous activation-repression: roles of C-boxes and spacer sequences in control of the PvuII restriction-modification system. *Nucleic Acids Res* 35, 6935–6952.
17. Sorokin, V., Severinov, K., and Gelfand, M.S. (2009) Systematic prediction of control proteins and their DNA binding sites. *Nucleic Acid Res* 37, 441–451.
18. Altschul, S.F., Madden, T.L., Schaffer, A.A. et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25, 3389–3402.
19. Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J. et al. (2000) Genbank. *Nucleic Acids Res* 28, 15–18.
20. Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32, 1792–1797.
21. Felsenstein, J. (1989) PHYLIP – Phylogeny inference package (version 3.2). *Cladistics* 5, 164–166.
22. Stavrovskaia, E.D., Makeev, V.I., Mironov, A.A. (2006) ClusterTree-RS: the binary tree algorithm for identification of co-regulated genes by clustering regulatory signals. *Mol Biol (Moscow)* 40, 524–532.
23. Wasserman, W.W., and Fickett, J.W. (1998) Identification of regulatory regions which confer muscle-specific gene expression. *J Mol Biol* 278, 167–181.