doi:10.1016/j.jmb.2011.01.056 J. Mol. Biol. (2011) 409, 47–53

Contents lists available at www.sciencedirect.com

Journal of Molecular Biology

journal homepage: http://ees.elsevier.com.jmb

REVIEW

The Dinucleotide CG as a Genomic Signalling Module

Adrian Bird

The Wellcome Trust Centre for Cell Biology, University of Edinburgh, Michael Swann Building, The King's Buildings, Edinburgh EH9 3JR, UK

Received 13 December 2010; received in revised form 27 January 2011; accepted 28 January 2011
Available online 3 February 2011

Edited by M. Yaniv

Keywords:
CpG islands;
DNA methylation;
MeCP2;
Cfp1

The operon model proposed the existence of a category of proteins that control gene expression by interacting with specific DNA sequences. Since then, a large number of transcription factors recognizing a diversity of sequence motifs have been discovered. This article discusses an unusually short protein recognition sequence, 5′CG, which is read by multiple DNA binding proteins. CG exists in three distinct chemical states, two of which bind mutually exclusively to proteins that modulate chromatin structure. Non-methylated CG, which is highly concentrated at CpG island promoters, recruits enzymes that create the mark of promoter activity, trimethyl-lysine 4 of histone H3. Methylated CG, on the other hand, is a gene silencing mark and accordingly recruits enzymes that deacetylate histones. Thus, CG, despite its simplicity, has the properties of a genome-wide signalling module that adds a layer of positive or negative control over gene expression.

Introduction

Cells contain a complete set of genetic information encoded in their DNA, but not all genes are expressed at any one time. In bacteria, for example, certain genes become active only in response to an environmental (e.g., nutritional) signal. Mammalian cells have large numbers of silent genes, as many proteins are specific for a particular cell type and must not be expressed in other kinds of cells. β-Globin, growth hormone, and opsin, to name only a few, are each repressed in the vast majority of mammalian cell types. This gene expression programme is a key end point of the process of development, ensuring that irrelevant genes are shut down, whereas required genes are either active or poised for expression when called upon. Jacob and Monod were the first to make headway in understanding how differential gene expression is achieved, and the conclusions of their seminal paper in 1961[1] continue to reverberate through contemporary molecular biology. A major revelation was that there are proteins whose role is not to facilitate metabolic processes directly, but which recognise and interact in *trans* with specific DNA sequences to control gene expression. Even today, when chromatin structure and DNA/histone modifications are high on the research agenda, proteins that recognise specific DNA sequences are still considered the primary instigators of differential activity across the genome. To direct any process to a specific region of the genome, it is necessary to be able to distinguish the target DNA from a vast excess of non-target DNA. Sequence-specific DNA binding proteins and nucleic acid polymers can do this, but it is not clear that anything else can. Therefore, changing patterns of transcription, replication, and recombination with concomitant alterations in histone/DNA modification and chromatin conformation are almost certainly secondary to a targeted triggering event based on the recognition of DNA sequence.[2,3]

The genome of the bacterium *Escherichia coli* is estimated to encode ~250 DNA binding proteins. Not all the DNA sequences that they recognise are yet known, but the length of DNA that is read is often in the range of 20–30 bp, albeit with considerable

E-mail address: a.bird@ed.ac.uk.

Abbreviations used: CGIs, CpG islands; H3K4me3, trimethylation of lysine 4 of histone H3; MBD, methyl-CpG binding domain.

variability.[4] The lac repressor discovered by Jacob and Monod, for example, binds a 21-bp site.[5] As genomes get bigger across the phylogenetic spectrum, the discriminatory power needed to target specific genomic locations becomes more daunting. To cope with this, one might expect that DNA binding motifs would become longer. In fact, the opposite is the case: the average length of DNA recognised by mammalian transcription factors is usually 6–8 bp.[6] To circumvent this logistical problem, multiple factors bind to clustered sites and collaborate to regulate gene expression. The β-globin locus control region, for example, binds the erythroid DNA sequence-specific transcription factors GATA1, EKLF/KLF1, and NF-E2, among others.[7] The probability that an appropriate cluster of binding sites will occur by chance is low. Combinatorial binding of transcription factors thereby restores the missing specificity.

This article concerns a eukaryotic DNA binding motif that is atypically short by any standards—just 2 bp long. At first sight, this seems much too simple to be biologically interesting because a 2-bp sequence will occur very frequently by chance. For example, the sequence AG would be expected once every 17 bp, on average, in a genome with a base composition of 40% G+C. Here, I will discuss evidence that the dinucleotide sequence CG, despite its simplicity, has the properties of a genomic signalling module that participates in local and global regulation of gene expression through interaction with DNA binding proteins. The focus will be on animal genomes, but many of the same arguments apply to plants, where CG is also a genomic signal.[8]

CG exists in chemically distinct forms

CG is a self-complementary DNA sequence, but not uniquely so, as it shares this property with three other dinucleotides: AT, TA, and GC. CG differs from the others in that it exists in three chemically distinct forms: unmethylated, methylated, and hydroxymethylated.[9,10] CG methylation involves modification of the cytosine base at the 5 position of the pyrimidine ring. Viewed along the axis of the DNA double helix, it is evident that the methyl group lies in the major groove of B-form DNA and does not sterically interfere with the exquisite base pairing between G and C, which would compromise coding specificity. An important advantage of self-complementarity is that modifications at CG can be copied at DNA replication.[11,12] In the case of methylated CG, the maintenance DNA methyltransferase Dnmt1 only methylates the newly synthesised progeny strand if the parental strand bears a methyl group. Heritability means that DNA methylation patterns are stable through cell division.

The most important biological consequence of DNA methylation that has been observed in many systems is long-term silencing of transcription.[13] This can be accomplished in two ways: (1) repulsion of transcription factors that require non-methylated CG in their binding site[14]; (2) attraction of proteins that specifically bind m5CG (see below). Set against this useful property, there is a downside to the presence of 5-methylcytosine—its mutability. Both 5-methylcytosine and cytosine are prone to hydrolytic deamination, giving rise to thymine and uracil, respectively. Uracil, being an alien DNA base, is recognised by a dedicated DNA repair system and restored to cytosine.[15] Thymine, on the other hand, is an authentic DNA component, albeit incorrectly base-paired after m5C deamination, and this appears to interfere with its efficient repair. Glycosylases that can remove T from a T:G mismatch have been identified and, in the case of MBD4, shown to contribute to repair, but their efficiency appears to be inadequate to eliminate the problem altogether.[16] As a result, many C-to-T transitions arise at sites of CG methylation and cause mutations.

Given the existence of glycosylases such as MBD4[17] and TDG,[18,19] it is interesting to speculate about why they fail to rectify all 5mC deaminations. A possible reason is that enzymes removing T from DNA must not be overzealous or they will cause more mutations than they repair. After all, there are nearly 10^9 bona fide T residues in a haploid genome, only one of which is likely to be due to m5C deamination at any one time. The scope for potentially mutagenic error is therefore large. This may be why uracil glycosylases, which remove U from a U:G mismatch, are inert when faced with a T:G mismatch. Given that the double helix breathes continually, the presence of a mismatch may not be enough to distinguish which T residues should be excised. Interestingly, it has been suggested that replacement of U by T in the presumed ancestral RNA genome during the early evolution of life was driven by the need to distinguish the deamination product of C from a normal DNA base so that it could be recognised for repair.[20] By methylating C, genomes recreate the dilemma that was so effectively circumvented by the invention of T. Mutability is the legacy of this evolutionary step. Medically, for example, it is evident that about 30% of all point mutations causing human disease arise at CG sites, almost certainly as a result of DNA methylation.[21]

Hydroxymethylation also occurs at the 5 position of the cytosine ring.[9,10] This is not coincidental, as it is created by enzymes that specifically hydroxylate the methyl group of m5C to create this altered form.[10] Little is yet known about the functional significance of this modification, but two alternatives will be important to distinguish: (1) hmC is a chromatin signal in its own right with distinct biological consequences; (2) it is an intermediate in the demethylation of DNA. With respect to the latter possibility, it is clear that demethylation of m5C