Dissecting direct and indirect readout of cAMP receptor protein DNA binding using an inosine and 2,6-diaminopurine in vitro selection system

Søren Lindemose, Peter Eigil Nielsen and Niels Erik Møllegaard*

Department of Cellular and Molecular Medicine, Panum Institute, University of Copenhagen, Blegdamsvej 3, DK-2200 Copenhagen N, Denmark

Received March 31, 2008; Revised June 28, 2008; Accepted June 30, 2008

ABSTRACT

The DNA interaction of the Escherichia coli cyclic AMP receptor protein (CRP) represents a typical example of a dual recognition mechanism exhibiting both direct and indirect readout. We have dissected the direct and indirect components of DNA recognition by CRP employing in vitro selection of a random library of DNA-binding sites containing inosine (I) and 2,6-diaminopurine (D) instead of guanine and adenine, respectively. Accordingly, the DNA helix minor groove is structurally altered due to the ‘transfer’ of the 2-amino group of guanine (now I) to adenine (now D), whereas the major groove is functionally intact. The majority of the selected sites contain the natural consensus sequence TGTGAN₆TCACA (i.e. TITIDN₆TCDCD). Thus, direct readout of the consensus sequence is independent of minor groove conformation. Consequently, the indirect readout known to occur in the TG/CA base pair step (primary kink site) in the consensus sequence is not affected by I–D substitutions. In contrast, the flanking regions are selected as I/C rich sequences (mostly I-tracts) instead of A/T rich sequences which are known to strongly increase CRP binding, thereby demonstrating almost exclusive indirect readout of helix structure/flexibility in this region through (anisotropic) flexibility of I-tracts.

INTRODUCTION

DNA-binding proteins achieve a large part of their specificity through direct hydrogen bonding and hydrophobic interactions between specific amino acid side chains and functional groups on the bases in the major and minor groove (1–5). However, these direct amino acid–base contacts (direct readout), is insufficient to fully explain the specificity of numerous DNA-binding proteins (6–12). In the indirect readout mechanism a local sequence-dependent DNA structure is recognized through protein contacts with the sugar–phosphate backbone and/or non-specific parts of the DNA bases. In this way, DNA features such as minor groove width, bending and flexibility/deformability of the helix adds another dimension to the recognition event. However, the contribution of the structural adaptations to binding affinity and thermodynamics is at present not fully understood.

One of the most extensively studied prokaryotic DNA-binding proteins is the cyclic AMP receptor protein (CRP) from Escherichia coli. The protein binds to DNA as a homodimer and regulates transcription initiation from more than 100 promoters (13,14 and references therein) by binding to DNA sequences located upstream from the RNA polymerase binding site (9,15). Comparative analysis of the CRP-binding sites in the E. coli genome has established a 22 bp 2-fold symmetrical consensus sequence 5'-AAATGTGAN₆TCACATTT-3' (16–22). Among the binding sites compared, the N₆ spacer sequence between the two half-sites seems to be only very weakly, if at all conserved (17,23).

Analysis of crystal structures of CRP–DNA complexes revealed features such as CRP-induced bending of the DNA helix and suggested a recognition mechanism including a combination of direct and indirect readout (19,24–26). Upon binding, each CRP monomer interacts directly with the DNA bases G₅, G₇ and A₈ within the symmetrical half-site: 5'-A₁A₂A₃T₄G₅T₆G₇A₈N₉N₁₀N₁₁-3' by means of a helix–turn–helix motif. The overall ~90° bending of the DNA in the CRP–DNA complex is a consequence of a primary and a secondary kink in each half-site. The preference for the remaining bases in the consensus is a consequence of an indirect readout mechanism in the sense that no direct protein–DNA nucleobase contacts have been identified. Especially, T₆ in the T₆G₇/CA base-pair step in the half-site is not in direct contact with the helix–turn–helix motif of the CRP monomer, but is nevertheless highly conserved and known to be
involved in the ~40° primary kink in the CRP–DNA complex observed in the crystals (25–27). In addition, two smaller secondary kinks are located in the flanking A/T rich sequences outside the T₄G₅T₆G₇A₈ sequence (9,19,24–27) and these sequences also appear to be important for DNA bending accommodated through electrostatic interactions between amino acids and phosphates in the DNA backbone (28–31).

Despite the fact that a consensus sequence has been deduced, most CRP-binding sites in the E. coli genome (13,32) deviate significantly from this, suggesting that the interaction with the protein cannot be determined by the specific base–amino acid contacts alone. In general, besides recognizing the bases within the half-sites, a critical factor for high affinity CRP binding relies on deformability of the DNA to accommodate an induced fit between protein and DNA (9,25–30).

The exocyclic 2-amino group of guanine is an element of prime importance in DNA structure and recognition and has been shown to exert a significant influence on DNA bending, flexibility and intrinsic curvature (33–44). Not only does the 2-amino group obstruct access to the donor available in the minor groove and it is the only hydrogen bonding inter- action by employment of an

**DNA oligos and plasmids**

The primers used were 345: 5'-AGTGAATTCCAGCTCGTGGT-3', 346: 5'-ATGACCATGATTACGCC-3', M13 forward: 5'-GTGAAACCGACGCGCCGATG-3', M13 reverse: 5'-CAGGAAACAGCTATGAC-3'. Pre-bending primer 1: 5'-AGTTTGTACCCAGGCT-3', Pre-bending primer 2: 5'-CGGCGCCGATGTTATG-3', Lac promoter 1: 5'-CATAAAGTGTAAAGCCT-3' and lac promoter 2: 5'-GAAAGCGGGCAGTGGAC-3'.

The sequence of the randomized in vitro selection template was: 5'-AGTGAATTCCAGCTCGTGGTATAT(N₁₂)ATATGGCCGTAATCATGTCAT-3' where underlined sequence show the location of primer 345 and 346. N denotes any base. Oligos G8.05₅⁻¹₃ was derived from clone G8.05. In G8.05⁺¹, the right I-tract (5'-GGGG GG-3') was changed to 5'-AGCAAA-3'. In G8.05⁺², the left I-tract (5'-GGGG-3') was changed to 5'-AGAC-3'. In G8.05⁻², both I-tracks were changed with the same sequences. The plasmids used in the study were pUC19, plasmid p309 and plasmid pICAP. Plasmid 309 was constructed by cloning of a 36 bp oligo containing the CRP consensus sequence (24) 5'-GATCCGCAAAGTGTGACATATGCACACTTCTTCG-3' into the BamHI site of pUC19 and Plasmid pICAP was constructed by cloning of a 75 bp PCR product containing the Berg–von Hippel CRP consensus sequence (17,18) 5'-GTCGGCGGCAAGCTATGACATATTCGATACACATGTCAT-3' into pCR².1-TOPO vector (Invitrogen, Carlsbad, California, US). The two half-sites of the CRP-binding site are underlined in both plasmids.

**3²P-labelled DNA fragments**

All 3²P-labelled DNA fragments were produced by standard techniques (46) using either T4 polynucleotide kinase or Large Fragment of DNA Polymerase I (Klenow).

**In vitro binding site selection**

The in vitro selection assay was modelled after previous in vitro selections studies for protein-binding sites on DNA (47–49). The binding site selection experiments were initiated by use of 20ng (~5 x 10¹ⁱ molecules) of single-stranded in vitro selection template oligo. A double-stranded randomized DNA oligo pool was generated by PCR as described below except that 10 pmol ³²P-labelled primer 345, 10 pmol primer 346 and 100 µM of each nucleotide I, D, dCTP and dTTP (I–D mix) was used and only four PCR cycles were run. To enrich the randomized oligo pool for CRP-binding sites, the oligo pool was incubated with 50 nM CRP and subjected to EMSA. Following electrophoresis, the band shifts corresponding to CRP–DNA complexes were cut out and the DNA was purified. Before starting the next round of selection, the obtained DNA fragments were PCR amplified with natural dNTPs in a volume of 50 µl. After the PCR amplification 20 µl of the reaction was stored at ~20°C as a ‘CRP-binding site DNA library’. The remaining 30 µl was gel purified before a new round of selection with I–D was initiated. In total, the double-stranded randomized oligo

**MATERIALS AND METHODS**

**Protein purification**

The wild-type CRP protein was purified as previously described (45) using an overproducing E. coli strain and cAMP affinity columns.
All PCR reactions in the study used a similar protocol. In each case, the template under study was PCR amplified in a total volume of 50 µl containing 10 mM Tris–HCl, pH 8.3, 50 mM KCl and 1.5 mM MgCl₂ and 2.5 U of Taq DNA polymerase (Fermentas, St. Leon-Rot, Germany) using either 200 µM dNTPs or I–D mix (from Roche and TriLink Biotechnologies, respectively). After an initial denaturing step of 2 min at 94 ºC, amplification cycles were performed with each cycle consisting of the following segments: 94 ºC for 30 s, 48 ºC for 30 s and 72 ºC for 30 s. After the last PCR cycle, the extension segment was continued for 7 min at 72 ºC before cooling down to room temperature. The PCR products were gel purified and resuspended in either 10 µl H₂O or CRP binding buffer depending on future use of the DNA (PCR or EMSA).

PCR for Krelative experiments: ICAP DNA fragments (276 bp) were obtained by PCR using primer M13 R, primer M13 F, dNTPs and plasmid pICAP as template. The different in vitro selection clone and mutant DNA fragments (75 bp) were similarly obtained by PCR using primer 345, primer 346, I–D mix and individual plasmids containing the cloned DNA sequences as template. PCR conditions for natural dNTP versus I–D experiments: two DNA fragments of different size were generated by PCR from plasmid pUC19 using the primers lac promoter 1 and 2. Finally, Lac P1 DNA fragment was generated by PCR from plasmid pUC19 using the primers lac promoter 1 and lac promoter 2.

Electrophoretic mobility shift assay

Double-stranded 32P-labelled DNA fragments and CRP protein were incubated in 10 or 30 µl CRP binding buffer (10 mM Tris–HCl, pH 8.0, 50 mM KCl, 2.5 mM MgCl₂, 1 mM EDTA, 55 µg/ml bovine serum albumin, 1 mM dithiothreitol, 0.05% NP-40, 2 µg/ml calf thymus DNA and 50 µM cAMP) containing 100 µM freshly made cAMP for 30 min at 23 ºC. After incubation, 3 or 9 µl loading buffer (CRP binding buffer containing 50% glycerol and 0.1 mg/ml bromophenol blue) was added and samples were immediately loaded on 5% (55:1) polyacrylamide gels and run at 6-8 V/cm at 23 ºC for 90–120 min. Following electrophoresis, the CRP–DNA complexes were detected by autoradiography or exposure to phosphor imager storage screens.

Relative binding constants and binding free energy change

The relative equilibrium binding constants, Krelative, of 26 individual clones, lac P1, ICAP with I–D and three mutants of G8.05 were measured by EMSA in a competition assay as previously described (30). All experiments were performed at least in triplicates. In this assay, a mixture of two different sized DNAs (5–20 pM), both containing a binding site for CRP, competes for a limited amount of CRP protein simultaneously. After incubation, the CRP–DNA complexes are resolved from each other and free DNAs by electrophoresis. Following exposure to phosphor imager storage screens, four different bands were clearly visible and the amount of radioactivity in each band was quantified using STORM Phosphor Imager scanner and Image Quant 5.2 software from Molecular Dynamics, Sunnyvale, California, US. The relative equilibrium binding constants were calculated by the formula: Krelative = (Kmutant)/(Kwild-type) = (Kclone)/(KICAP), where Kclone is the ratio of protein-bound clone DNA divided by free clone DNA, and KICAP is the same ratio for the ICAP–DNA. The binding free energy change, ΔΔG, which is the difference between the binding free energy for CRP–DNAclone complex formation versus the binding free energy for CRP–DNAICAP complex formation, was calculated from the general assumption: ΔΔG = RTln[(KICAP)/Krelative] – RTln(KICAP). This is in our system equivalent to: ΔΔG = –RTln(Krelative) where Krelative is the relative equilibrium binding constant described above, R is the gas constant [8.3145 joule/(mol x K)] and T is the temperature in Kelvin. The average Krelative obtained from at least triplicate experiments was used in the expression. Note that positive ΔΔG values indicate a reduction of binding affinity.

Cloning

PCR products were cloned directly into the pCR® 2.1-TOPO vector and transformed into the E. coli TOP10 strain using the TOPO TA Cloning kit (Invitrogen) according to the manufacturer’s recommendations.

Sequencing

Inserts from 89 individual white colonies were sequenced with ABI PRISM BigDye Terminator v1.1 Cycle Sequencing Kit (Applied Biosystems, Foster city, California, US) using a 3100 Genetic Analyser (Applied Biosystems).

Uranyl photo-cleavage and DNase I footprinting

The uranyl photo-cleavage and DNase I digestion was performed as previously described (50,51). A Molecular Dynamics STORM PhosphorImager was used to collect data from the phosphor storage screens and base-line corrected scans were obtained by using Image Quant version 5.2 software. Differential cleavage plots were calculated from the expression ln(fa) – ln(fc) representing the differential cleavage at each bond relative to the control (where fa is the fractional cleavage at any bond in the presence of the protein, and fc is the fractional cleavage of the same bond in the control). Using this expression, positive values indicate enhanced cleavage, whereas negative values indicate cleavage inhibition (footprints).

RESULTS

CRP binding to I and D substituted ICAP consensus sequence

Initially, by use of a gel based competition assay, we tested the effect of I and D substitutions on CRP binding to the...
strongest known CRP-binding site, the symmetric ICAP consensus DNA sequence 5‘-AAATGTGATCTAGA TCACTATT-3‘, which binds CRP much stronger than the naturally occurring binding sites in E. coli (15,18,31).

Two DNA fragments of different sizes, one containing normal nucleobases and the other I and D instead of guanine and adenine, respectively, were constructed by PCR and incubated with CRP. After incubation, where the two types of DNA competed for a limited amount of CRP, protein-bound DNA was separated from non-bound DNA by gel-electrophoresis (EMSA). The competition experiment demonstrates that the relative affinity of CRP for normal DNA (dNTP) is approximately 70 times higher ($K_{\text{relative}} = 0.014 \pm 0.002; \Delta G = 2.50 \text{kcal/mol}$) than for I and D containing DNA (Figures 1 and 5).

In addition to the primary change of width and chemical properties of the minor groove, and to a much lesser extent the effect on the accessibility and structure of the major groove (34–36), the I and D substitutions may also change structural parameters such as the bending and flexibility of the DNA helix (35,36,43,44). The result, therefore, strongly suggests that inherent structural parameters such as minor groove width and bending and flexibility, which either may be anisotropic or isotropic in some region of the binding site, are of significant importance for binding. Alternatively, protein-induced deformability may in some regions be diminished upon I and D substitutions. However, the substitutions still allow strong CRP binding due to maintenance of direct amino acid–base pair contacts in the major groove of the I and D substituted DNA fragment. Consequently, the experiment demonstrates that both direct and indirect readout are involved in CRP–DNA interactions as also previously suggested (9, 25–27).

**In vitro selection of CRP-binding sites containing nucleobase analogues**

In order to gain more detailed information on the observed consequence of changing the amino substituents in the minor groove and on the recognition mechanism of CRP, a modified PCR-based *in vitro* selection method was developed, in which I and D triphosphates were incorporated instead of dGTP and dATP, respectively, into the PCR products.

The starting material for CRP selection was a population of $\sim 5 \times 10^{11}$ different DNA fragments of 75 bp in which the central 32 bp had been randomized. After incubation with the CRP protein, EMSA was employed to separate CRP–DNA complexes from free DNA. The DNA from the CRP–DNA complex was purified and PCR amplified before the next round of selection. After eight rounds of selection, the obtained DNA fragments were used as a template for a final round of PCR, employing normal nucleotides, in order to clone and sequence the selected CRP-binding sites.

A total of 89 individual clones were sequenced and 49 different DNA sequences were obtained (Figure 2). Thus, several of the sequences were found more than once and a single sequence was present 12 times indicating a relatively stringent selection. To simplify the comparison with natural CRP-binding sites, the selected sequences in Figure 2 are presented with normal nucleobases instead of I and D. Interestingly, several of the clones harbour a TGT insertion (i.e. TITIDN 6TCDCD) sequence, which contains exactly the two 5-bp half-sites spaced by 6 nt as found in the ICAP consensus sequence obtained when naturally occurring CRP-binding sites are compared (16,17,23). In fact all the selected sequences contain a perfect consensus sequence or minor variations thereof. Therefore, it appeared natural to align all the sequences with respect to this consensus sequence.

**Footprint analysis of in vitro selected binding sites**

Although the occurrence of a consensus sequence indicates CRP binding to this particular region, DNase I and uranyl photo-footprinting analysis were performed to verify and map the binding. In addition to the ICAP consensus sequence three clones (G8.03, G8.05, and G8.29) were analysed and a typical autoradiograph is shown in Figure 3A. As expected the DNase I results (exemplified by clone G8.05 in Figure 3A) clearly demonstrate binding of CRP to the half-sites, and specific CRP phosphate backbone interactions were confirmed by uranyl photofootprinting. The unique protein–phosphate contacts probed by uranyl in each of the analysed binding sites were determined and are presented as differential cleavage plots in Figure 3B. From these plots it is evident that 4–6 phosphates flanking each half-site is protected (black bars), whereas, most interestingly, phosphates in between the two half-sites show hypersensitivity towards uranyl cleavage (arrows). Even though this hypersensitivity towards uranyl is most noticeable for the ICAP sequence, it is evident that the uranyl cleavage pattern of the binding sites analysed is rather similar.

The fact that uranyl footprinting analysis indicates strong protein–phosphate interactions in the flanking A/T rich regions in the symmetrical ICAP binding site...
as well as in the I/C-rich regions in the selected binding sites (black bars in Figure 3B) is in full agreement with X-ray crystallography structures, which have demonstrated that CRP contacts the phosphates in the two protected regions outside the half-sites (9). In contrast, the strong hypersensitivity towards uranyl cleavage observed in the N6 spacer region between the two half-sites (arrows in Figure 3B) is not readily explained by the X-ray crystallography data, which indicate that the same phosphates interact with CRP in the crystals (19, 24).

However, overall the DNase I and uranyl footprinting experiments demonstrate that CRP binds in a nearly identical fashion to the symmetrical ICAP binding site and to the selected I and D containing sites.

![Figure 2](image-url)

**Figure 2.** Isolation of CRP-binding sites. The sequences of cloned CRP-binding sites obtained after eight rounds of selection are shown. Eighty-nine individual clones were analysed and contained 49 different sequences. For simplicity, the I and D content has been replaced by guanine (G) and adenine (A), respectively. The frequencies of the cloned sequences are indicated to the right of the table. The sequences are aligned about the core consensus of the two half-sites: TGTGA-N6-TCACA (bold letters). Underlined sequences indicate the location of fixed sequence in the N6 spacer region between the two half-sites (arrows). The frequencies of the cloned sequences are indicated to the right of the table. The sequences are aligned about the core consensus of the two half-sites: TGTGA-N6-TCACA (bold letters). Underlined sequences indicate the location of fixed sequence in the N6 spacer region between the two half-sites (arrows).
Comparison of the selected sequences

When the selected sequences are aligned, the similarity to the CRP consensus sequence is striking. Thus, although not all selected binding sites have a perfect consensus sequence, it is anticipated that the two half-sites in all the selected sequences direct binding of CRP. This assumption is supported by the footprint analysis in Figure 3. The nucleotide frequencies at each position of the selected sequences are compiled in Figure 4, which also for comparison includes the symmetrical ICAP consensus sequence with depiction of amino acid and phosphate contacts (dots and ovals, respectively) deduced from X-ray crystallography. It is noticed that the variations in the 5-bp half-sites, where specific nucleo-base amino acid contacts take place, only occur at specific positions. In the left half-site TGTGA variations exclusively occur at positions −4 and −6 (underlined) and in the right half-site TCACA variation predominantly occurs in positions +6 and +4 (underlined). The other positions (−5, −7, −8, +8, +7 and +5) in the half-sites are extremely well conserved. This observation is fully in agreement with the X-ray crystallographic data obtained from several CRP–DNA complexes where the bases at position −4, −6, +6 and +4 are not engaged in direct amino acid

Figure 3. DNase I and uranyl footprints of CRP–DNA complexes. (A) Autoradiograph showing the DNase I digestion and uranyl photo-cleavage pattern of the I and D containing clone G8.05. The sequence of this clone is 5′-CCCCC-TGTGA-TCCTTG-TCACG-CCCCG-3′ where the half sites are underlined. Note that the probing result shown is obtained from the complementary and reversed strand of the sequence in Figure 2. On top of the figure, C indicates the untreated DNA (75 bp), S is a Maxam–Gilbert DMS G reaction and +/- denotes presence and absence of 100 nM CRP protein. Black bars to the left of the figure show the position of the two half-sites, whereas numbering from the labelling is shown to the right. (B) Differential cleavage plots comparing the susceptibility of G8.03, G8.05, G8.29 and the ICAP consensus sequence to uranyl photo-cleavage in the absence and presence of CRP protein. As a reference, only the ICAP consensus sequence is shown below the plots where the two half-sites are denoted with bold letters. Black bars denote phosphate protection (footprints) and arrows indicate hypersensitive phosphates in the CRP–ICAP complex. Note that the vertical axis is in units of ln(fa) − ln(fb).

Figure 4. Nucleotide frequencies of the 49 selected sequences from Figure 2. The central 22 bases from each clone have been aligned. Nucleotides from the 5′ end are numbered −1 to −11 and the 3′ end is numbered +11 to +1. At every nucleotide position, a frequency was calculated and a threshold of 0.5 was used to deduce the consensus sequence. N means any base. Note that I and D have been replaced by guanine (G) and adenine (A), respectively. For comparison, the ICAP consensus sequence is shown below. CRP-ICAP data from X-ray crystallography (9) has been added with depiction of amino acid (ovals) and phosphate contacts (black dots).
contacts leaving them less important for CRP binding (9,24–27).

The finding that all of the selected I and D sequences contain two half-sites with highly conserved bases, as also seen for unmodified DNA, strongly support the consensus that the direct readout, i.e. nucleobase–amino acid contacts, in the two 5-bp half-sites is indispensable for high-affinity binding. In contrast, the position of the 2-amino group on guanine and consequently, the width of the minor groove in the 5-bp half-sites seem to be of minor importance in this region. Furthermore, the selected sites have the TI/CD step at the primary kink site, which indicates that CRP is able to create the primary kink deformation involving compression of the major groove (positive roll angle) independent on the position of the 2-amino group in the TG/CA step.

Finally, sequences flanking the two 5-bp half-sites have been shown to contribute significantly to CRP binding (29,30). Interestingly, several of the binding sites are I-rich in the flanking sequences on both sides of the two half-sites as opposed to A/T-rich sequences normally found in naturally occurring CRP-binding sites (Figures 2 and 4).

**Relative binding constants**

To decipher the importance of the different DNA segments of the binding sites, such as half-sites and flanking sequences, we measured the relative binding constants ($K_{\text{relative}}$) for a subset of selected clones using ICAP, which is the strongest CRP-binding site known, as internal standard (Figure 5). As a reference point, we found that CRP binds to ICAP with a $K_d = 1 \times 10^{-10}$ M (data not shown). Twenty-six of the clones shown in Figure 2 were chosen for further analysis on the basis of variations in the two half-sites and in the number of I/C base pairs in sequences flanking the 5-bp half-sites. From these experiments it is revealed that there is a 30- to 40-fold difference in the measured $K_{\text{relative}}$ values between the best and the weakest sites (Figure 5), and the strongest binding site isolated, clone G8.85, binds CRP only 13 times weaker than ICAP ($\Delta \Delta G_{G8.85} = 1.48$ kcal/mol).

In comparison to this, the wild-type CRP-binding site lac P1 from the *E. coli lac* promoter, which is one of the stronger CRP-binding sites in the natural genome (15,18,31,52) was estimated to bind CRP approximately 80 times weaker than ICAP ($\Delta \Delta G_{\text{lac P1}} = 2.60$ kcal/mol). Thus, binding of CRP to several of the strongest I and D containing binding sites is markedly stronger than all known naturally occurring binding sites.

**Effect of variations in the 5-bp half-sites**

Nine of the 26 clones analysed in Figure 5 have the perfect 5-bp consensus sequence TGTGAn/C1/C1, whereas the other clones contain either one or two variations in the half-sites. Interestingly, it is noted that none of the 9 clones with perfect half-sites are among the strongest...
binding sites. In fact five of the six weakest clones analysed have a perfect TGTTGAN<sub>n</sub>TCACA sequence. Thus, the two half-sites cannot exclusively govern the strength of the CRP–DNA interactions.

The effect of a C instead of a T in position −4 and a G instead of an A in position +4 seems not to be affecting the affinity negatively as they occur frequently. Indeed the strongest site isolated (G8.85) has both a C and a G in these two positions (−4 and +4, Figure 5) and a positive effect of a G at position +4 is clearly demonstrated when we compare clone G8.85 with G8.01. These two clones are nearly identical in sequence, but clone G8.01 has an A in position +4 instead of a G. This may explain the ∼3-fold reduction (ΔΔG<sub>G8.01</sub> − ΔΔG<sub>G8.85</sub> = ∼0.62 kcal/mol) in relative affinity of clone G8.01 compared to clone G8.85.

In contrast an A in position −4 seems to reduce binding since clone G8.08 represents a relatively weak binding site compared to stronger binding sites with a T or C in position −4 (compare G8.08 with e.g. G8.03 and G8.05, and also with G8.50).

Even though the T in the TG step at positions −6 and −7 in the left half-site is not in direct contact with any residues of the CRP protein, it is highly conserved and known to be involved in a ∼40° kink in the CRP–DNA complexes observed in the crystals (25–27). Despite the use of I and D we observe that the high preference for a T in this position is maintained. In other words, the base step TI/CD seems to be able to undergo the same major groove compression as the TG/CA step in normal DNA.

However, in a few cases other base pair steps than the TI/CD at the primary kink site were found among the selected sequences. Specifically, some of the binding sites (clones G8.09, G8.29, G8.30 and G8.82) have a left half-site with the sequence TGGGA or CGGGA, i.e. II/CC at the primary kink site. Selection of these sites may well be explained by the high deformability of the region, since I, due to the absence of the 2-amino group in the minor groove, allows local deformability of the DNA (34–40,43,44). In fact, our data led us to conclude that a TI/CD base pair step is as deformable as the II/CC base pair step. This is evident when comparing clone G8.05 (TI/CD) with G8.82 (II/CC) and clone G8.01 (TI/CD) with G8.82 (II/CC). These clones are pairwise nearly identical in sequence but contain a different base at position −6 (a T versus an I). Nevertheless, there is no significant difference in relative affinity (Figure 5).

Finally, a DI/CT base pair step is found at the primary kink site in clones G8.80 and G8.84, further emphasizing that base pair steps other than TG/CA as in normal DNA can undergo the deformation needed for high-affinity CRP binding.

**Effect of I-tracts outside the 5-bp half-sites**

By inspection of Figures 2 and 4 it is evident that the flanking regions proximal to the consensus half-sites are very inosine rich and most contain pure I-tracts without IC or CI steps. It is also noted that the clone with the highest affinity, G8.85, has I-tracts on both sides of the consensus site. This is also true for clones G8.07, G8.19, G8.05, G8.82 and G8.01. Other clones like G8.80, G8.29, G8.09, G8.23, G8.26, G8.11, G8.76 and G8.08 have only a single I-tract flanking either the left or the right side of the consensus site. This is in contrast to several of the weakest binding sites, e.g. clones G8.06 and G8.74, which have a perfect consensus sequence but no flanking I-tract. Thus, to a first approximation it appears that I-tracts outside the consensus sequence increase the overall affinity. Based on the measured K<sub>relative</sub> values in Figure 5, the significance of the I-tracts may be estimated. Firstly, in order to identify the contribution of a single I-tract clones G8.06 and G8.23 are compared. These two clones have identical half-sites and a nearly identical intervening N<sub>6</sub> spacer region and right-flanking sequence. From this comparison it seems plausible that addition of an I-tract to clone G8.06 on the left side of the first half-site could be responsible for the 4-fold higher relative affinity of clone G8.23 (ΔΔG<sub>G8.06</sub> − ΔΔG<sub>G8.23</sub> = 0.85 kcal/mol).

Likewise clones G8.76 and G8.01 are very similar, but a T (at position +2) in clone G8.76 interrupts the right I-tract, which may be responsible for the ∼3-fold weaker relative affinity (ΔΔG<sub>G8.76</sub> − ΔΔG<sub>G8.01</sub> = 0.55 kcal/mol).

The importance of the I-tracts is obvious if clone G8.23 and G8.74 are compared. Despite minor differences in the N<sub>6</sub> spacer region between the two half-sites, it seems reasonable to assume that the presence of the I-tract in clone G8.74 is responsible for the 8-fold higher relative affinity of clone G8.23 (ΔΔG<sub>G8.74</sub> − ΔΔG<sub>G8.23</sub> = 1.26 kcal/mol).

In general, sequences with two I-tracts flanking the two 5-bp half-sites constitute significantly better binding sites than sequences without flanking I-tracts. Thus, the presence of the I-tracts may account for the observed difference between strong and weak binding sites.

These examples clearly show that I-rich flanking sequences increase binding, and to further support this conclusion clone G8.05 was used as a scaffold for synthesizing three mutants in which the right, left or both I-tracts were substituted with a random sequence (Figure 6A). The results from these K<sub>relative</sub> measurements were a 2-fold reduction in relative affinity when the right I-tract was replaced (ΔΔG<sub>G8.05</sub> = 0.47 kcal/mol) and a 3-fold reduction in relative affinity when the left I-tract was replaced (ΔΔG<sub>G8.05</sub> = 0.67 kcal/mol). When both I-tracts were replaced, a 4-fold decrease (ΔΔG<sub>G8.05</sub> = 0.88 kcal/mol) was observed (Figure 6A).

By removal of both I-tracts from G8.05 we end up with a DNA sequence (i.e. G8.05<sup>−2</sup>) that looks very much like clone G8.17 where both half-sites are identical. In fact the K<sub>relative</sub> values of these two clones are very close to each other.

Thus I-tracts in the flanking regions apparently have structural characteristics that facilitate CRP binding. Upon binding, CRP wraps the DNA-binding site around the protein surface in order to optimize the fit between the partners (9,24–30). DNA wrapping around CRP is accompanied by a compression of the minor groove in the flanking regions 10–11 bases away from the dyad axis just outside the 5-bp half-sites (9,24–30). Therefore, for a high affinity CRP-binding site this DNA region must be either statically bent towards the
minor groove or be (anisotropically) flexible. We note a pronounced strand asymmetry of the I-tracts as for virtually all clones the I-tracts are on the same strand (and thus the C-tracts on the other strand). Since the tracts are 20 bp apart corresponding to two helical turns, this arrangement would indicate that the effect is indeed due to directional bending and/or anisotropic flexibility of these I-tracts.

It is noteworthy that I-tracts, A-tracts and in general A/T-rich sequences, which have all been shown to increase CRP binding if present proximal to the consensus site, are characterized by a narrow minor groove in solution (29,30,33,34,53). However, in contrast to pure A-tracts (without TA step), I-tracts (and general A/T-rich sequences) in phase with the helical pitch do not give pronounced strand asymmetry of the I-tracts. Since the tracts are 20 bp apart corresponding to two helical turns, this arrangement would indicate that the effect is indeed due to directional bending and/or anisotropic flexibility of these I-tracts.

CRP binding to in vitro selected sequences containing natural dNTP

Despite the use of I and D in the selection we find strong binding sites with a normal consensus sequence, or minor variations of that, very strongly indicating that the direct readout of CRP in the half-sites is not significantly affected by I and D. However, substituting I and D for the natural G and A bases in the ICAP sequence resulted in a ~70-fold reduction (\(\Delta \Delta G = 2.50 \text{kcal/mol} \)) in relative affinity of CRP (Figures 1 and 5). As discussed above, flanking I-tracts may make significant contributions to the CRP–DNA binding energy similarly to what was previously found for flanking A/T-rich sequences (29,30).

Therefore, we ascribe the reduced binding of CRP to I–D containing ICAP predominantly (or exclusively) to the binding contribution of the A-tracts which is lost upon I–D substitution.

In order to further dissect the relative contribution of the flanking sequences, we decided to study the effect on CRP binding strength of reverting three of the I–D selected clones (G8.06, G8.29 and G8.50 having no, one or two flanking I-tracts) to normal nucleobases (Figure 6B). The relative affinity of CRP binding to the sequence in clone G8.06 is only slightly affected by the presence of normal nucleobases instead of I and D, whereas the relative affinity of CRP binding to clone G8.50 is decreased approximately 25 times (\(\Delta \Delta G = 1.86 \text{kcal/mol} \)) and that of clone G8.29 >2000-fold (\(\Delta \Delta G > 4.4 \text{kcal/mol} \)). These results confirm that the indirect readout of the I-tracts contributes very significantly to the binding energy. From the data of Figure 6A the contribution is estimated to ~1 kcal/mol, whereas the data of clone G8.29 in Figure 6B would indicate >4 kcal/mol. The latter is clearly an overestimate since in this case the G → I substitution may also affect the consensus recognition through the TG → GG change at the primary kink site, which is expected to reduce binding by ~1.4 kcal/mol per half-site (20).

CONCLUSION

The present results clearly demonstrate that in vitro selection employing modified nucleobases such as inosine and 2,6-diaminopurine is a powerful tool for analysing the contribution of direct and indirect interactions in protein-DNA recognition.

Specifically, we find that with I–D-substituted DNA, CRP selects binding sites that have a preferred TITIDN_{5–6}TCGCAC consensus sequence, corresponding to TGTGAN_{9–10}TCACA and thus identical to that previously found to be optimal for CRP binding to non-modified DNA. Therefore, the results emphasize that direct readout (occurring in the major groove) in the two half-sites is not significantly influenced by the position of the 2-amino group in the minor groove (and consequently by minor groove width). Furthermore, all the selected sequences...
contain an inosine in position 7 and the majority of the sequences have thymine in position 6. Thus the 2-amino group does not seem to affect the indirect readout interaction at the $T_{DG}$ step, which is responsible for primary kinking in the half sites.

In contrast, it is clearly demonstrated that the 2-amino group affects the indirect readout component of CRP–DNA interactions in the flanking regions one helical turn from the centre of the binding site, where A/T rich sequences have been shown to increase affinity in normal CRP-binding sites. Selection of inosine-rich sequences in these regions emphasizes the importance of flexibility/deformability, known to be present in sequences containing I/C base pairs, as opposed to direct sequence readout. Flexibility introduced by inosine substitutions has previously been shown to account for strong affinity of both the FIS and HMG proteins for their respective binding targets (35,36). Furthermore the results are in full accordance with previous conclusions on the structural similarities between A/T-rich sequences and I-tracts (33,34,36), and finally our results additionally suggest that pure I-tracts are anisotropically flexible.

In a simplified model one may consider the DNA recognition of CRP to be divided into four partially independent (half-site) components: the pentameric consensus element, the primary kink site, the proximal flanking region and the intervening ($N_6$) region. From the present results it seems quite clear that while the consensus element is recognized through a direct readout mechanism, the flanking regions clearly contribute almost exclusively via indirect readout, as most probably does the kink site, while the contribution of the $N_6$ region still remains to be established.

Very strong CRP-binding sites were selected in this study indicating that selection has been rigorous. In the E. coli genome most, if not all, CRP-binding sites appear to have half-sites strongly deviating from the consensus. In some cases, as in the gal operon half-site sequences may even exist without the two important guanines in the TGTGA (i.e. TITID) sequence found in all the selected sequences. Therefore, some weaker CRP-binding sites severely deviating from the consensus half-site TGTGA (i.e. TITID) may be obtained with a less rigorous selection. This could reveal new interesting DNA structural alternatives (e.g., increased indirect readout contacts that compensate for loss of direct readout contacts) that are capable of forming the CRP–DNA complex as exemplified by the sequence of clone G8.29. Such studies are in progress.

We also foresee that the in vitro selection system presented in this study could be very useful for categorizing different DNA-binding proteins with respect to the contribution and magnitude of indirect versus direct readout components to the recognition process.

ACKNOWLEDGEMENT

This work was supported by the Novo Nordisk Foundation, the Lundbeck Foundation and the Danish Medical Research Council. Funding to pay the Open Access publication charges for this article was provided by University of Copenhagen.

Conflict of interest statement. None declared.

REFERENCES


