Published online 24 July 2008

Nucleic Acids Research, 2008, Vol. 36, No. 14 ***4797–4807***
doi:10.1093/nar/gkn452

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 TGTGAN$_6$TCACA (i.e. TITIDN$_6$TCDCD). 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′-AAATGTGAN$_6$TCACATTT-3′ (16–22). Among the binding sites compared, the N_6 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 G_5, G_7 and A_8 within the symmetrical half-site: 5′-$A_1A_2A_3T_4G_5T_6G_7A_8N_9N_{10}N_{11}$-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, T_6 in the T_6G_7/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

*To whom correspondence should be addressed. Tel: +45 35327778; Fax: +45 35326042; Email: nem@imbg.ku.dk

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 $T_4G_5T_6G_7A_8$ 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 floor of the minor groove in B-DNA, but it also disrupts the pattern of hydration, alters the electrostatic potential in the minor groove and it is the only hydrogen bonding donor available in the minor groove.

Minor groove width is an important parameter for ligand–DNA recognition, and is to a first approximation correlated with the contents of A/T and G/C base pairs (absence or presence of the 2-amino group, respectively).

The two widely used nucleobase analogues, inosine (I) and 2,6-diaminopurine (D), offers the possibility to study the effect of the 2-amino group in DNA (39). In essence, the nucleobase analogues I and D keep the major groove information intact, whereas the minor groove properties including width are changed. With respect to minor groove width, we have previously shown that the minor groove width of I (i.e. guanine without the 2-amino group) rich sequences resemble that of A/T rich sequences and that D (i.e. adenine with a 2-amino group) rich sequences resemble that of G/C rich sequences (33,34).

In order to define regions of structural importance (as opposed to major groove recognition information) in the binding region of the CRP protein, we have studied the interaction of CRP with I and D substituted DNA in solution by employment of an *in vitro* selection system. Accordingly, substitutions in the minor groove may be a unique way to unravel the importance of the exocyclic 2-amino group in recognition of sequence-specific major groove binding proteins whose binding mechanism includes both direct and indirect recognition mechanisms.

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.

DNA oligos and plasmids

The primers used were 345: 5′-AGTGAATTCGAGCTC GGT-3′, 346: 5′-ATGACCATGATTACGCC-3′, M13 forward: 5′-TGTAAAACGACGGCCAGT-3′, M13 reverse: 5′-CAGGAAACAGCTATGAC-3′, Pre-bending primer 1: 5′-AGCTTGGTACCGAGCT-3′, Pre-bending primer 2: 5′-CGGCCGCCAGTGTGAT-3, *Lac* promoter 1: 5′-CATAAAGTGTAAAGCCT-3′ and *lac* promoter 2: 5′-GAAAGCGGGCAGTGAGC-3.

The sequence of the randomized *in vitro* selection template was: 5′-AGTGAATTCGAGCTCGGTATAT(N_{32}) ATATGGCGTAATCATGGTCAT-3′ where underlined sequence show the location of primer 345 and 346. N denotes any base. Oligos G8.05*[1-3] was derived from clone G8.05. In G8.05*[1], the right I-tract (5′-GGGG GG-3′) was changed to 5′-AGACAA-3′. In G8.05*[2], the left I-tract (5′-GGGG-3′) was changed to 5′-AGAC-3′. In G8.05*[3], both I-tracts 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′-GATCGCGAAAAGTGTGA CATATGTCACACTTTTCGC-3′ into the BamHI site of pUC 19 and Plasmid pICAP was constructed by cloning of a 75 bp PCR product containing the Berg–von Hippel CRP consensus sequence (17,18) 5′-AGTGAATTCGA GCTCGGTGCAACGCAATAAATGTGATCTAGATC ACATTTTAGGCACCGGCGTAATCATGGTCAT-3′ into pCR®2.1-TOPO vector (Invitrogen, Carlsbad, California, US). The two half-sites of the CRP-binding site are underlined in both plasmids.

^{32}P- labelled DNA fragments

All ^{32}P-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 20 ng ($\sim 5 \times 10^{11}$ 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 ^{32}P-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

pool was subjected to eight identical rounds of selection before the DNA was cloned and sequenced.

PCR

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 $MgCl_2$ 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°C, amplification cycles were performed with each cycle consisting of the following segments: 94°C for 30 s, 48°C for 30 s and 72°C for 30 s. After the last PCR cycle, the extension segment was continued for 7 min at 72°C before cooling down to room temperature. The PCR products were gel purified and resuspended in either 10 µl H_2O or CRP binding buffer depending on future use of the DNA (PCR or EMSA).

PCR for $K_{relative}$ 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 the same plasmid where one PCR fragment (75 bp) contained I–D mix and the other PCR fragment (180 bp) contained dNTPs. Clone DNA fragments containing dNTPs were generated using Pre-bending primer 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 ^{32}P-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 $MgCl_2$, 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°C. 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°C 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, $K_{relative}$, 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: $K_{relative} = (K_{mutant})/(K_{wild\text{-}type}) = (K_{clone})/(K_{ICAP})$, where K_{clone} is the ratio of protein-bound clone DNA divided by free clone DNA, and K_{ICAP} is the same ratio for the ICAP–DNA. The binding free energy change, $\Delta\Delta G$, which is the difference between the binding free energy for CRP–DNA_{clone} complex formation versus the binding free energy for CRP–DNA_{ICAP} complex formation, was calculated from the general assumption: $\Delta\Delta G = RT\ln(K_{d;\,clone}) - RT\ln(K_{d;\,ICAP}) = -RT\ln[(K_{d;\,clone})/(K_{d;\,ICAP})]$. This is in our system equivalent to: $\Delta\Delta G = -RT\ln(K_{relative})$ where $K_{relative}$ is the relative equilibrium binding constant described above, R is the gas constant [8.3145 joule/(mol × K)] and T is the temperature in Kelvin. The average $K_{relative}$ obtained from at least triplicate experiments was used in the expression. Note that positive $\Delta\Delta G$ 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 $\ln(f_a) - \ln(f_c)$ representing the differential cleavage at each bond relative to the control (where f_a is the fractional cleavage at any bond in the presence of the protein, and f_c 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 TCACATTT-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 ($K_{relative} = 0.014 \pm 0{,}002$; $\Delta\Delta G = 2.50$ 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).

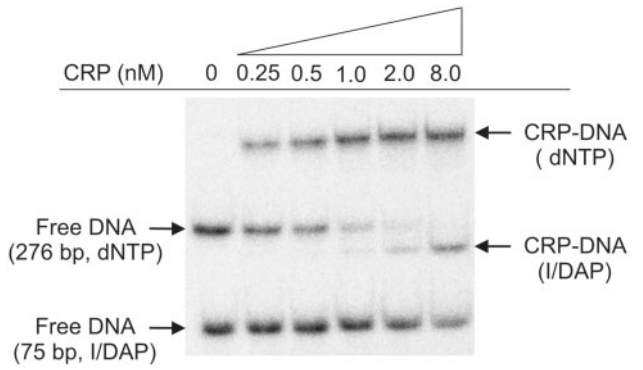

Figure 1. Gel retardation assay for the determination of the relative binding constant of ICAP DNA fragments that contain dNTPs (276 bp) or I and D (75 bp). Both fragments were mixed and titrated with increasing concentration of CRP protein as indicated above the figure. The CRP–DNA complexes and the free DNAs were resolved by gel electrophoresis and the relative binding constants, $K_{relative}$, was calculated as described in Materials and methods section. Arrows indicate the location and identity of the DNA and CRP–DNA complexes. By use of these substitutions we measure that the affinity of the CRP–DNA complex containing I and D is 70-fold lower compared to the complex containing dNTP ($K_{relative} = 0.014 \pm 0.002$; $\Delta\Delta G = 2.50$ kcal/mol).

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 GAN_6TCACA (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

Clone number	Sequence	*f*
G8.01	TGGCACGGGG**CGTGA**GACAGG**TCACA**GGGGAGGATA	1
G8.02	ATATAAAGGG**TGAGA**GGTCTG**TCACG**GGGCATGTGG	1
G8.03	TTACAGAGGG**CGTGA**CAAGGA**TCACA**GCTGGCGATA	12
G8.04	ACGAGGAGGG**TGTGA**CACGGA**TCACG**GATATGGCGT	4
G8.05	TATGACGGGG**CGTGA**CAAGGA**TCACA**GGGGGGAGCA	1
G8.06	GCAGAGCAGA**TGTGA**GAGGAG**TCACA**ATATGGCGTA	1
G8.07	TATATAGGGG**CGTGA**CAGGAG**TCACG**GGGAGACGCG	1
G8.08	ATCACGAGGG**AGTGA**CAGGGA**TCACA**GGGGTGCAAT	1
G8.09	ATCACAGGGG**CGGGA**CAACAA**TCACA**CGGGGACCAT	1
G8.11	TACCGGGGGG**CGTGA**GAAAGG**TCACG**GGCGTCAATA	1
G8.16	ATATAGGGGA**CGGGA**TACGAG**TCACA**GGGCAAACGG	1
G8.17	GACAAAAACA**CGTGA**CACACA**TCACA**TATGGCGTAA	2
G8.19	TTAAGAGGGG**CGTGA**TCGGAT**TCACG**GGGGAGCATA	1
G8.20	GAGAATGGAA**TGTGA**TCGGAG**TCACA**GCCTGAATAT	2
G8.21	TGGGAGGGCG**CGTGA**CGAATA**TCACA**AGACGGCATA	1
G8.23	CGGACAGGGG**TGTGA**GAGCAG**TCACA**TATGGCGTAA	1
G8.26	TAGCGGGAGG**TGTGA**GACGGA**TCACG**GGGGTGTATA	4
G8.27	TGCGTCAAGG**CGTGA**GAAGAA**TCACG**GGGGCAGATA	4
G8.28	AATAAGAGGG**TGTGA**TGAGAG**TCACG**CGATATGGCG	2
G8.29	ATATGAGGGG**TGGGA**CACAAC**GCACA**GGGCTAAGAC	1
G8.30	TGAGCGAGGG**TGGGA**GACGCA**TCACA**GGGCCGAATA	5
G8.33	AACAGGAGGA**GGTGA**CAGGAA**TCACA**GGGGATATGG	1
G8.38	TATCAAAAGA**TGTGA**GCGTAT**TCACA**GGGGGAGTAT	1
G8.41	ATATGCGGGG**AGTGA**CAGGGA**TCACG**GGCTAAGCAC	1
G8.43	TCAGCAGAGC**TGTGA**CAGGGG**TCACG**CCCGATCATA	1
G8.46	TGGCGGCAAT**TGTGA**CAGAGC**GCACA**GCCCCAGATA	1
G8.48	GCTCGGTATA**TGTGA**GAACAG**TCACA**GGGCCGCCGC	1
G8.50	GGCGGCGAGG**TGTGA**CACAGG**TCACA**GGGCATATGG	2
G8.52	ACAGAGAGGA**CGTGA**CACGGG**TCACA**TATGGCGTAA	1
G8.54	ATGATGAGGG**TGTGA**CGAGTA**TCGCG**AGGGGTGCAT	1
G8.56	TCAAAAGAAG**CGTGA**CACCGG**TCACA**GGGCTGGATA	1
G8.59	GTATATAGGG**TGTGA**CGTAGG**TCACA**AGCAAGCGCG	1
G8.60	ATGATGGGGG**TGTGA**CGAGTA**TCGCG**AGGGGTGCAT	1
G8.61	GGCATAGGAA**TGTGA**CAAGAA**TCACA**CGATATGGCG	1
G8.64	ATATGGAGGA**CGTGA**TGTGGA**TCACA**GGCAACGGGA	1
G8.65	GGCTGGAACA**TGTGA**CAGAGA**TCACG**GTGCAATATG	1
G8.68	ATATACAGGG**CGTGA**GCAGAG**TCACG**GGCTCAATCG	1
G8.70	AGCGGCGGCC**TGTGA**GGCGGA**TCACA**TATGGCGTAA	1
G8.72	TGAAAAGAAG**CGTGA**CACCGG**TCACA**GGGCTGGATA	2
G8.74	CAAGGGGAAA**TGTGA**GACGGA**TCACA**TATGGCGTAA	1
G8.76	GGGGTGGGGG**CGTGA**TGCGTA**TCACA**GTGGATATGG	1
G8.78	ATGCAACAGA**TGTGA**GCTGTG**TCACA**AATATGGCGT	1
G8.80	ATGGCGGGGG**TGAGA**CACGAG**TCACA**GGGTACCCAT	1
G8.82	TATATGGGGG**CGGGA**GACGAG**TCACA**GGGGCTGAGC	1
G8.84	CAGGGAAAGG**TGAGA**CTCGTG**TCACA**TATGGCGTAA	1
G8.85	GGGCTAGGGG**CGTGA**CACGGG**TCACG**GGGGTATATG	1
G8.86	GGTATATGGG**TGTGA**CACGGA**TCACA**AGGAGGGCCG	1
G8.88	AAAAGCGAGG**CGTGA**CGAGGG**TCACA**TATGGCGTAA	1
G8.89	ATATGCAAAA**TGTGA**CAGGGG**TCACG**GGCCAACACA	1
Lac P1	TAA**TGTGA**GTTAGC**TCACT**CAT	-
ICAP	AAA**TGTGA**TCTAGA**TCACA**TTT	-

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-N_6-TCACA (bold letters).Underlined sequences indicate the location of fixed sequence in the *in vitro* selection template. For comparison, the symmetric ICAP consensus sequence and the wild-type *lac* P1 site are shown.

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 N_6 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.

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 nucleobase 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

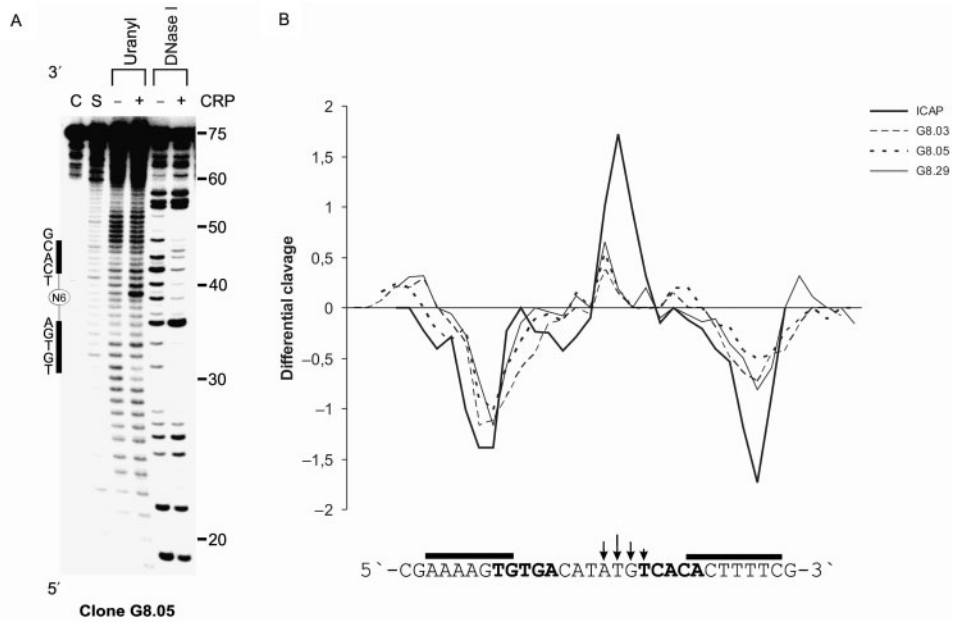

Figure 3. DNase I and uranyl footprints of CRP–DNA complexes. (**A**) Autoradiograph showing the DNase I digestion and uranyl photo-cleavage pattern of the I and D containing clone G8.05. The sequence of this clone is 5′-CCCCC-TGTGA-TCCTTG-TCACG-CCCCG-3′ where the half sites are underlined. Note that 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 $\ln(f_a) - \ln(f_c)$.

	−1	−2	−3	−4	−5	−6	−7	−8	−9	−10	−11										
A	17	7	14	2	0	3	0	**49**	0	**33**	13	9	**17**	**21**	0	0	**47**	0	**33**	7	9	4
G	**32**	**37**	**32**	1	**49**	5	**49**	0	15	10	13	**33**	**23**	**24**	2	0	2	0	16	**31**	**33**	**26**
C	0	4	2	**20**	0	0	0	0	**28**	5	19	6	2	2	0	**49**	0	**49**	0	4	4	9
T	0	1	1	**26**	0	**41**	0	0	6	1	4	1	7	2	**47**	0	0	0	0	7	3	10
												+11	+10	+9	+8	+7	+6	+5	+4	+3	+2	+1
Frequency	0.65	0.76	0.65	0.94	1.0	0.84	1.0	1.0	0.57	0.67	-	0.67	0.82	0.92	0.96	1.0	0.96	1.0	0.67	0.63	0.67	0.53
Consensus	G	G	G	**C/T**	**G**	**T**	**G**	**A**	C	A	N	G	G/A	G/A	**T**	**C**	**A**	**C**	**A**	G	G	G

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 ($K_{relative}$) 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 $K_{relative}$ 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 ($\Delta\Delta G_{lac\ P1} = 2.60$ 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$_6$TCACA, 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

Clone number	Sequence	$K_{relative}$	$1/K_{relative}$	$\Delta\Delta G$ (kcal/mol)
ICAP (dNTP)	CGCAATAAA**TGTGA**TCTAGA**TCACA**TTTTAGGCA	–	–	[0]
ICAP (I/DAP)	CGCAATAAA**TGTGA**TCTAGA**TCACA**TTTTAGGCA	0,014 ± 0,002	71	2.50
Lac P1 (dNTP)	CGCAATTAA**TGTGA**GTTAGC**TCACT**CATTAGGCA	0,012 ± 0,002	83	2.60
G8.85	GGCTAGGGG**CGTGA**CACGGG**TCACG**GGGGTATAT	0,080 ± 0,008	13	1.48
G8.50	GCGGCGAGG**TGTGA**CACAGG**TCACA**GGGCATATG	0,068 ± 0,005	15	1.58
G8.07	ATATAGGGG**CGTGA**CAGGAG**TCACG**GGGAGACGC	0,047 ± 0,013	21	1.80
G8.80	TGGCGGGGG**TGAGA**CACGAG**TCACA**GGGTACCCA	0,041 ± 0,005	24	1.88
G8.19	TAAGAGGGG**CGTGA**TCGGAT**TCACG**GGGGAGCAT	0,041 ± 0,003	24	1.88
G8.05	ATGACGGGG**CGTGA**CAAGGA**TCACA**GGGGGGAGC	0,040 ± 0,008	25	1.89
G8.03	TACAGAGGG**CGTGA**CAAGGA**TCACA**GCTGGCGAT	0,036 ± 0,008	28	1.95
G8.82	ATATGGGGG**CGGGA**GACGAG**TCACA**GGGGCTGAG	0,035 ± 0,005	29	1.97
G8.20	AGAATGGAA**TGTGA**TCGGAG**TCACA**GCCTGAATA	0,034 ± 0,002	29	1.99
G8.29	TATGAGGGG**TGGGA**CACAAC**GCACA**GGGCTAAGA	0,034 ± 0,003	29	1.99
G8.01	GGCACGGGG**CGTGA**GACAGG**TCACA**GGGGAGGAT	0,028 ± 0,002	36	2.10
G8.84	AGGGAAAGG**TGAGA**CTCGTG**TCACA**TATGGCGTA	0,028 ± 0,005	36	2.10
G8.09	TCACAGGGG**CGGGA**CAACAA**TCACA**CGGGGACCA	0,020 ± 0,003	50	2.30
G8.61	GCATAGGAA**TGTGA**CAAGAA**TCACA**CGATATGGC	0,018 ± 0,002	56	2.36
G8.23	GGACAGGGG**TGTGA**GAGCAG**TCACA**TATGGCGTA	0,017 ± 0,001	59	2.39
G8.26	AGCGGGAGG**TGTGA**GACGGA**TCACG**GGGGTGTAT	0,017 ± 0,001	59	2.39
G8.30	GAGCGAGGG**TGGGA**GACGCA**TCACA**GGGCCGAAT	0,016 ± 0,005	63	2.43
G8.11	ACCGGGGGG**CGTGA**GAAAGG**TCACG**GGCGTCAAT	0,015 ± 0,001	67	2.47
G8.76	GGGTGGGGG**CGTGA**TGCGTA**TCACA**GTGGATATG	0,011 ± 0,002	91	2.65
G8.08	TCACGAGGG**AGTGA**CAGGGA**TCACA**GGGGTGCAA	0,010 ± 0,001	100	2.71
G8.78	TGCAACAGA**TGTGA**GCTGTG**TCACA**AATATGGCG	0,010 ± 0,001	100	2.71
G8.48	CTCGGTATA**TGTGA**GAACAG**TCACA**GGGCCGCCG	0,009 ± 0,001	111	2.77
G8.17	ACAAAAACA**CGTGA**CACACA**TCACA**TATGGCGTA	0,006 ± 0,001	167	3.01
G8.70	GCGGCGGCC**TGTGA**GGCGGA**TCACA**TATGGCGTA	0,005 ± 0,001	200	3.11
G8.06	CAGAGCAGA**TGTGA**GAGGAG**TCACA**ATATGGCGT	0,004 ± 0,001	250	3.24
G8.74	AAGGGGAAA**TGTGA**GACGGA**TCACA**TATGGCGTA	0,002 ± 0,001	500	3.65

Figure 5. Relative binding constants and binding free energy calculations for 26 selected clones. The sequences are aligned about the core consensus of the two half-sites: TGTGA-N$_6$-TCACA (bold letters). Underlined sequences indicate the location of fixed sequence in the *in vitro* selection template. The data were obtained in competition assays where ICAP DNA fragments (reference) and selected binding sites (clones) competed for a limiting amount of CRP protein. $K_{relative}$ and $\Delta\Delta G$ was calculated as described in the Materials and methods section. For comparison, the wild-type *lac* P1 binding site was also included in the analysis.

binding sites. In fact five of the six weakest clones analysed have a perfect $TGTGAN_6TCACA$ 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 ($\Delta\Delta G_{G8.01} - \Delta\Delta G_{G8.85} = \sim 0.62$ 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 $K_{relative}$ 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_6 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 ($\Delta\Delta G_{G8.06} - \Delta\Delta G_{G8.23} = 0.85$ 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 ($\Delta\Delta G_{G8.76} - \Delta\Delta G_{G8.01} = 0.55$ 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_6 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 ($\Delta\Delta G_{G8.74} - \Delta\Delta G_{G8.23} = 1.26$ 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 $K_{relative}$ measurements were a 2-fold reduction in relative affinity when the right I-tract was replaced ($\Delta\Delta G_{G8.05*1} - \Delta\Delta G_{G8.05} = 0.47$ kcal/mol) and a 3-fold reduction in relative affinity when the left I-tract was replaced ($\Delta\Delta G_{G8.05*2} - \Delta\Delta G_{G8.05} = 0.67$ kcal/mol). When both I-tracts were replaced, a 4-fold decrease ($\Delta\Delta G_{G8.05*3} - \Delta\Delta G_{G8.05} = 0.88$ kcal/mol) was observed (Figure 6A).

By removal of both I-tracts from G8.05 we end up with a DNA sequence (i.e. $G8.05^{*3}$) that looks very much like clone G8.17 where both half-sites are identical. In fact the $K_{relative}$ 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

Clone number	Sequence	$K_{relative}$	$1/K_{relative}$	G (kcal/mol)
A				
G8.05	ACGGGG**CGTGA**CAAGGA**TCACA**GGGGGG	0.040 ± 0.008	25	1.89
G8.05*[1]	ACGGGG**CGTGA**CAAGGA**TCACA**AGACAA	0.018 ± 0.004	56	2.36
G8.05*[2]	ACAGAC**CGTGA**CAAGGA**TCACA**GGGGGG	0.013 ± 0.001	77	2.56
G8.05*[3]	ACAGAC**CGTGA**CAAGGA**TCACA**AGACAA	0.009 ± 0.002	111	2.77
B				
G8.06 (dNTP)	AGCAGA**TGTGA**GAGGAG**TCACA**ATATGG	0.611 ± 0.039	1.64	0.29
G8.29 (dNTP)	GAGGGG**TGGGA**CACAAC**GCACA**GGGCTA	< 0.0005	> 2000	> 4.46
G8.50 (dNTP)	GCGAGG**TGTGA**CACAGG**TCACA**GGGCAT	0.042 ± 0.005	24	1.86

Figure 6. Relative binding constants and binding free energy calculations for mutants (**A**) and dNTP containing clones (**B**). The sequences are aligned about the core consensus of the two half-sites: TGTGA-N_6-TCACA (bold letters). The data were obtained from competition assays and $K_{relative}$ and $\Delta\Delta G$ was calculated as described in the Materials and methods section. (A) The I-tracts was systematically changed in clone G8.05 to random sequence (shaded grey) in order to isolate the effect of flanking sequence. The mutants (75 bp) containing I–D competed against the dNTP containing ICAP fragment (276 bp). (B) The effect of incorporating dNTP instead of I and D into 3 clones was studied. In each experiment two DNA fragments of different sizes and nucleobase content were generated by PCR from the same plasmid. In each experiment, the I–D containing fragments (75 bp) derived from clone G8.06, G8.29 and G8.50 competed against their own dNTP containing sequence, respectively (180 bp). ICAP DNA was in these experiments not used as internal standard as the relative affinities for the dNTP containing DNA fragments were off-scale. Therefore, each clone competed against its own sequence. Nevertheless, it was not possible to calculate a $K_{relative}$ value for clone G8.29, as no CRP–DNA (dNTP) complex was visible even after long-term exposure using phosphor imager screens indicating a very low binding constant.

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 rise to markedly macroscopic curvature at room temperature (33,40–42,54), although it has been found that I-tracts may produce slight static DNA bending, which is enhanced at low temperature (33,34,40–42,54). Several natural CRP-binding sites in *E. coli* contain A/T-rich sequences in the flanking regions, and mutational analysis of the *Lac* P1 CRP-binding site has demonstrated that placing a pure A-tract, which is directionally bent into the minor groove, in the flanking regions, produce much stronger CRP than binding sites than analogous TA containing A/T-rich sequences (29,30). Consequently, and not surprisingly, the strongest CRP-binding sites contain already pre-bent flanking A-tracts. Less efficient but still strong binding sites are expected for helically phased anisotropically flexible sequences followed by isotropically flexible sequences (such as AT-tracts). Weak binding sites contain flanking regions of low bendability (flexibility). The enhancing effect of the asymmetrically positioned flanking I-tracts on CRP binding is therefore most likely due to anisotropic flexibility of I-tracts (rather than static bending).

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 are 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 ($\Delta\Delta G = 2.50$ 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 ($\Delta\Delta G = 1.86$ kcal/mol) and that of clone G8.29 > 2000-fold ($\Delta\Delta G > 4.4$ 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$_6$TCDCD consensus sequence, corresponding to TGTGAN$_6$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_6G_7 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 anisoptropically 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

1. Schleif,R. (1988) DNA binding by proteins. *Science*, **241**, 1182–1187.
2. Pabo,C.O. and Sauer,R.T. (1984) Protein-DNA recognition. *Annu. Rev. Biochem.*, **53**, 293–321.
3. Pabo,C.O. and Sauer,R.T. (1992) Transcriptions factors: structural families and principles of DNA recognition. *Annu. Rev. Biochem.*, **61**, 1053–1095.
4. Martin,A.M., Sam,M.D., Reich,N.O. and Perona,J.J. (1999) Structural and energetic origins of indirect readout in site-specific DNA cleavage by a restriction endonuclease. *Nat. Struct. Biol.*, **6**, 269–377.
5. Luscombe,N.M., Laskowski,R.A. and Thornton,J.M. (2001) Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. *Nucleic Acids Res.*, **29**, 2860–2874.
6. Hilchey,S.P. and Koudelka,G.B. (1997) DNA-based loss of specificity mutations. *J. Biol. Chem.*, **272**, 1646–1653.
7. Koudelka,G.B., Harrison,S.C. and Ptashne,M. (1987) Effect of non-contacted bases on the affinity of 434 operator for 434 repressor and Cro. *Nature*, **326**, 886–888.
8. Mendieta,J., Pèrez-Lago,L., Salas,M. and Camacho,A. (2007) DNA sequence-specific recognition by a transcriptional regulator requires indirect readout of A-tracts. *Nucleic Acids Res.*, **35**, 3252–3261.
9. Lawson,C.L., Swigon,D., Murakami,K.S., Darst,S.A., Berman,H.M. and Ebright,R.H. (2004) Catabolite activator protein: DNA binding and transcription activation. *Curr. Opin. Struct. Biol.*, **14**, 10–20.
10. Otwinovski,Z., Schevitz,R.W., Zhang,R.G., Lawson,C.L., Joachimiak,A., Marmorstein,R.Q., Luisi,B.F. and Sigler,P.B. (1988) Crystal structure of trp repressor/operator complex at atomic resolution. *Nature*, **335**, 321–329.
11. Bareket-Samish,A., Cohen,I. and Haran,T.E. (2000) Signals for TBP/TATA box recognition. *J. Mol. Biol.*, **299**, 965–977.
12. Hedge,R.S. (2002) The papillomavirus E2 proteins: structure, function, and biology. *Annu. Rev. Biophys. Biomol. Struct.*, **31**, 343–360.
13. Salgado,H., Gama-Castro,S., Peralti-Gil,M., Dias-Peredo,E., Sànches-Solano,F., Santos-Zavaleta,A., Martìnez-Flores,I., Jimènez-Jacinto,V., Bonavides-Martìnez,C., Segura-Salazar,J. *et al.* (2006) RegulonDB (version 5.0): *Escherichia coli* K-12 transcriptional regulatory network, operon organisation, and growth conditions. *Nucleic Acids Res.*, **34 (Database issue)**: D394–D397.
14. Hollands,K., Busby,S.J.W. and Lloyd,G.S. (2007) New targets for the cyclic AMP receptor protein in the *Escherichia coli* K12 genome. *FEMS Microbiol. Lett.*, **274**, 89–94.
15. Kolb,A., Busby,S., Buc,H., Garges,S. and Adhya,S. (1993) Transcriptional regulation by cAMP and its receptor protein. *Annu. Rev. Biochem.*, **62**, 749–795.
16. O'neill,M.C., Amass,K. and de Crombrugghe,B. (1981) Molecular model of the DNA interactions site for the cyclic AMP receptor protein. *Proc. Natl Acad. Sci. USA*, **78**, 2213–2217.
17. Berg,O.G. and von Hippel,P.H. (1988) Selection of DNA binding sites by regulatory proteins II. The binding of cyclic AMP receptor protein to recognition sites. *J. Mol. Biol.*, **200**, 709–723.
18. Ebright,R.H., Ebright,Y.W. and Gunasekera,A. (1989) Consensus DNA site for the Escherichia coli catabolite gene activator protein (CAP): CAP exhibits a 450-fold higher affinity for the consensus DNA site than for the E. coli lac DNA site. *Nucleic Acids Res.*, **17**, 10295–10305.
19. Parkinson,G., Wilson,C., Gunasekera,A., Ebright,Y.W., Ebright,R.H. and Berman,H.M. (1996) Structure of the CAP-DNA complex at 2.5 angstrom resolution: a complete picture of the protein-DNA interface. *J. Mol. Biol.*, **260**, 395–408.
20. Gunasekera,A., Ebright,Y.W. and Ebright,R.H. (1992) DNA sequence determinants for binding of the Escherichia coli catabolite gene activator protein. *J. Biol. Chem.*, **267**, 14713–14720.

21. Liu-Johnson,H.N., Gartenberger,M.R. and Crothers,D.M. (1986) The DNA binding domain and Bending angle of E. coli CAP protein. *Cell*, **47**, 995–1005.
22. Wu,H.M. and Crothers,D.M. (1984) The locus of sequence-directed and protein-induced DNA bending. *Nature*, **308**, 509–513.
23. Barber,A.M. and Zhurkin,V.B. (1990) CAP binding sites reveal pyrimidine–purine pattern characteristic of DNA bending. *J. Biomol. Struct. Dyn.*, **8**, 213–232.
24. Schultz,S.C., Shields,G.C. and Steitz,T.A. (1991) Crystal structure of a CAP-DNA complex: the DNA is bent by 90 degrees. *Science*, **253**, 1001–1007.
25. Chen,S., Vojtechovsky,J., Parkinson,G.N., Ebright,R.H. and Berman,H.M. (2001) Indirect readout of DNA sequence at the primary-kink site in the CAP-DNA complex: DNA binding specificity based on energetics of DNA kinking. *J. Mol. Biol.*, **314**, 63–74.
26. Chen,S., Gunasekera,A., Zhang,X., Kunkel,T.A., Ebright,R.H. and Berman,H.M. (2001) Indirect readout of DNA sequence at the primary-kink site in the CAP-DNA complex: alteration of DNA binding specificity through alteration of DNA kinking. *J. Mol. Biol.*, **314**, 75–82.
27. Napoli,A.A., Lawson,C.L., Ebright,R.H. and Berman,H.M. (2006) Indirect readout of DNA sequence at the primary-kink site in the CAP-DNA complex: recognition of pyrimidine-purine and purine-purine steps. *J. Mol. Biol.*, **357**, 173–183.
28. Kapanidis,A.N., Ebright,Y.W., Ludescher,R.D., Chan,S. and Ebright,R.H. (2001) Mean DNA bend angle and distribution of DNA bend angles in the CAP-DNA complex in solution. *J. Mol. Biol*, **312**, 453–468.
29. Gartenberg,M.R. and Crothers,D.M. (1988) DNA sequence determinants of CAP-induced bending and protein binding affinity. *Nature*, **333**, 824–829.
30. Dalma-Weiszhausz,D.D., Gartenberg,M.R. and Crothers,D.M. (1990) Sequence-dependent contribution of distal binding domains to CAP protein-DNA binding affinity. *Nucleic Acids Res.*, **19**, 611–616.
31. Gaston,K., Kolb,A. and Busby,S. (1989) Binding of the Escherichia coli cyclic AMP receptor protein to DNA fragments containing consensus nucleotide sequences. *Biochem. J.*, **261**, 649–653.
32. Robison,K., Mcguire,A.M. and Church,G.M. (1998) A comprehensive library of DNA binding site matrices for 55 protein applied to the complete *Escherichia coli* K-12 genome. *J. Mol. Biol.*, **284**, 241–254.
33. Møllegaard,N.E., Bailly,C., Waring,M.J. and Nielsen,P.E. (1997) Effects of diaminopurine and inosine substitutions on A-tract induced DNA curvature. Importance of the 3'-A-tract junction. *Nucleic Acids Res.*, **25**, 3497–3502.
34. Bailly,C., Møllegaard,N.E., Nielsen,P.E. and Waring,M.J. (1995) The influence of the 2-amino group of guanine on DNA conformation. Uranyl and DNase I probing of inosine/diaminopurine substituted DNA. *EMBO J.*, **14**, 2121–2131.
35. Bailly,C., Waring,M.J. and Travers,A.A. (1995) Effects of base substitutions on the binding of a DNA-bending protein. *J. Mol. Biol.*, **253**, 1–7.
36. Bailly,C., Payet,D., Travers,A.A. and Waring,M.J. (1996) PCR-based development of DNA substrates containing modified bases: an efficient system for investigating the role of the exocyclic groups in chemical and structural recognition by minor groove binding drugs and proteins. *Proc. Natl Acad. Sci. USA*, **93**, 13623–13628.
37. Bailly,C. and Waring,M.J. (1995) Transferring the purine 2-amino group from guanines to adenines in DNA changes the sequence-specific binding of antibiotics. *Nucleic Acids Res.*, **23**, 885–892.
38. Bailly,C. and Waring,M.J. (1998) The use of diaminopurine to investigate structural properties of nucleic acids and molecular recognition between ligands and DNA. *Nucleic Acids Res.*, **26**, 4309–4314.
39. Bailly,C. and Waring,M.J (2001) Use of DNA molecules substituted with unnatural nucleotides to probe specific drug-DNA interactions. *Methods Enzymol.*, **340**, 485–502.
40. Diekmann,S., von Kitzing,E., McLaughlin,L.W., Ott,J. and Eckstein,F. (1987) The influence of exocyclic substituents of purine bases on DNA curvature. *Proc. Natl Acad. Sci. USA*, **84**, 8257–8261.
41. Diekmann,S., Mazzarelli,J.M., McLaughlin,L.W., von Kitzing,E. and Travers,A.A. (1992) DNA curvature does not require bifurcated hydrogen bonds or pyrimidine methyl groups. *J. Mol. Biol.*, **225**, 729–738.
42. Koo,H.S. and Crothers,D.M. (1987) Chemical determinants of DNA bending at adenine-thymine tracts. *Biochemistry*, **26**, 3745–3748.
43. Buttinelli,M., Minnock,A., Panetta,G., Waring,M. and Travers,A.A. (1998) The exocyclic groups of DNA modulate the affinity and position of the histone octamer. *Proc. Natl. Acad. Sci. USA*, **95**, 8544–8549.
44. Virstedt,J., Berge,T., Henderson,R.M., Waring,M.J. and Travers,A.A. (2004) The influence of DNA stiffness upon nucleosome formation. *J. Struct. Biol.*, **148**, 66–85.
45. Ghosaini,L.R., Brown,A.M. and Sturtevant,J.M. (1988) Scanning calorimetric study of the thermal unfolding of catabolite activator protein from *E. coli* in the absence and presence of cyclic mononucleotides. *Biochemistry*, **27**, 5257–5261.
46. Sambrook,J., Fritsch,E. and Maniatis,T. (1989) *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbour Press, Cold Spring Harbour, NY, USA.
47. Pedersen,H. and Valentin-Hansen,P. (1997) Protein-induced fit: CRP activator protein changes sequence-specific DNA recognition by the CytR repressor, a highly flexible LacI member. *EMBO J.*, **16**, 2108–2118.
48. Blackwell,T.K. and Weintraub,H. (1990) Differences and similarities in DNA-binding preferences of MyoD and E2A protein complexes revealed by binding site selection. *Science*, **250**, 1104–1110.
49. Pollock,R. and Treisman,R. (1990) A sensitive method for the determination of protein-DNA binding specificities. *Nucleic Acids Res.*, **18**, 6197–6204.
50. Lindemose,S., Nielsen,P.E. and Møllegaard,N.E. (2005) Polyamines preferentially interact with bent adenine tracts in double-stranded DNA. *Nucleic Acids Res.*, **33**, 1790–1803.
51. Nielsen,P.E., Møllegaard,N.E. and Jeppesen,C. (1990) DNA conformational analysis in solution by uranyl mediated photocleavage. *Nucleic Acids Res.*, **18**, 3847–3851.
52. Kolb,S., Busby,S., Herbert,M., Kotlarz,D. and Buc,H. (1983) Comparison of the binding sites for the *Escherichia coli* cAMP receptor protein at the lactose and galactose promoters. *EMBO J.*, **2**, 217–222.
53. Møllegard,N.E., Lindemose,S. and Nielsen,P.E. (2005) Uranyl photoprobing of nonbent A/T- and bent A-tracts. A difference of flexibility? *Biochemistry*, **44**, 7855–7863.
54. Shatzky-Schwarts,M., Arbuckle,N.D., Eisenstein,M., Rabinovich,D., Bareket-Samish,A., Haran,T.E., Luisi,B.F. and Shakked,Z. (1997) X-ray and solution studies of DNA oligomers and implications for the structural basis of A-tract-dependent curvature. *J. Mol. Biol.*, **267**, 595–623.