Comparing binding site information to binding affinity reveals that Crp/DNA complexes have several distinct binding conformers

Peter C. Holmquist[1]*, Gerald P. Holmquist[2], and Michael L. Summers[1]

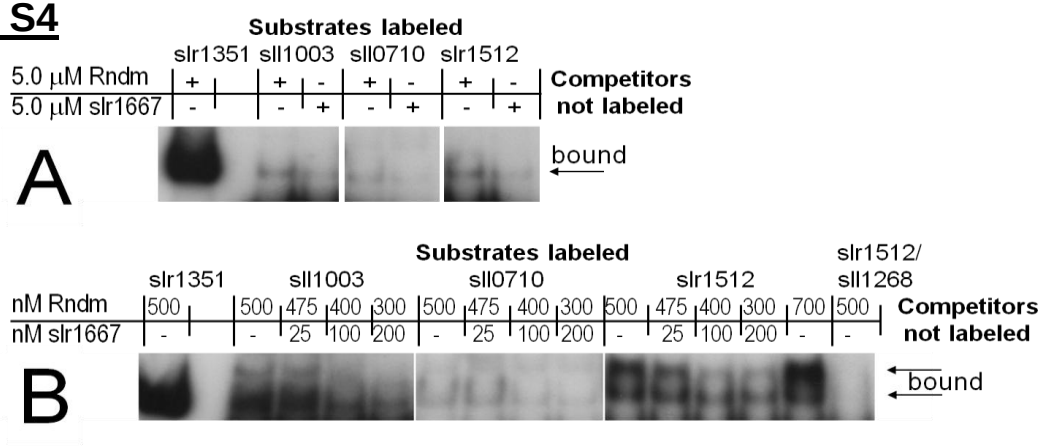
[1]California State University Northridge, Department of Biology, 18111 Nordhoff St. Northridge, CA 91330

[2]City of Hope, 1500 E. Duarte Blvd. Duarte, CA 91010

*corresponding author. Tel.: +1-818-677-7238; fax.: +1- 818 677-2034. E-mail address: peter_holmquist@yahoo.com

Supplementary Figures S4-S9

Figure S4

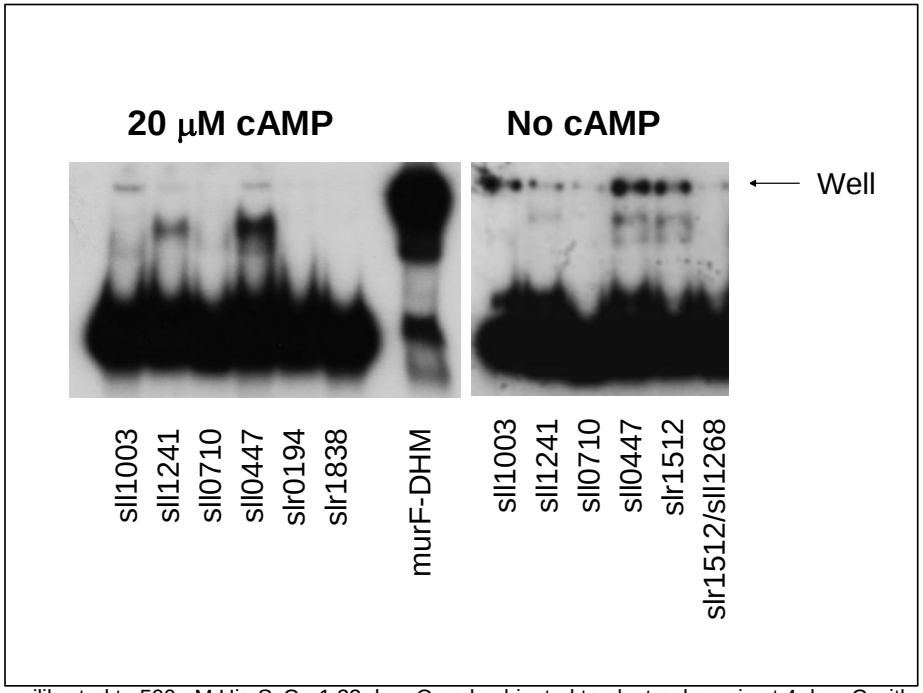


Sequence specificity for (+)monad (+)bend substrates. Competitive EMSA was performed using 1.0 nM radiolabeled substrates from Figure 1A against the non-specific (Rndm.) or specific (slr1667) unlabeled competitors as indicated. **A.** Competitive EMSA using either specific or non-specific competitor at high concentrations. **B.** Titration by increasing the specific competitor while maintaining a constant total 500 nM concentration of specific and non-specific unlabeled competitors combined. Final binding reaction equilibration and electrophoresis were performed at 4 °C. All reactions contained 500 nM His-SyCrp1, 20 μM cAMP, and reaction buffer. The specific activity of ^{32}P in the 1.0 nM slr1351 substrate was reduced to visualize the shifted band.

For all EMSA, the binding reaction buffer contained 20 μM cAMP where indicated, 500 nM His-SyCrp1 or as indicated, 50mM Tris-HCl pH 7.5, 60 mM NaCl, 1.0 mM EDTA, 8.3 % glycerol, and 0.1 mg/ml acetylated bovine serum albumin (BSA). His-SyCrp1 titration experiments to determine K_d contained 0.1 nM labeled dsDNA substrate. All other reactions contained 1.0 nM labeled dsDNA substrate. Unlabeled competitor substrates Rndm. (randomized sequence from a 50% G/C pool) or slr1667 described by Hedger et al., (2009) (2) and His-SyCrp1 were added to the reaction buffer prior to adding labeled dsDNA substrates. Equilibrium binding reactions and electrophoretic conditions consisted of either 4 or 22 °C. For 4 °C assays, reactions were incubated at 22º C for 25 min, then iced for 15 min. For 22 °C assays, reactions were incubated at 22º C for 25 min. All 4 °C assays contained 500 nM Rndm. substrate to facilitate migration into the gel whereas 22º C assays did not. The reactions were directly loaded on the gel without loading buffer or dye. 90 V was immediately applied for 10 min (4 °C assays only), and then 200 V for an additional 35 to 45 min. The gel apparatus and buffers were equilibrated to the indicated temperatures prior to loading. 90 V was applied for at least 30 min prior to loading to remove mobile charged molecules and the negative electrode tank was replaced with fresh buffer immediately prior to loading. Running buffers for 10 % acrylamide gels was 0.25 × TBE pH 8.0 at 4 °C or pH 7.5 at 22 °C titrated to match the reaction buffer at the indicated running temperature, 20 μM cAMP (or none where noted), and a 50:1 (w/w) acrylamide to bis-acrylamide ratio. All gels, running buffers, and reaction buffers contained 20 μM cAMP (except where noted). Reagents were not filtered following addition of BSA or cAMP to maintain the indicated concentrations.

Biomax Light film from Kodak were exposed to EMSA gels at –80 °C using an intensifying screen and then developed by hand. For quantification purposes, EMSA gels were dried and exposed to a Fujifilm BAS-MS phosphor screen at room temperature without an intensifying screen. Exposure levels of these phosphor screens were quantified using a Biorad Personal FX phospho imager and Quantity One software.

Figure S5



Xu and Su, (2009) (1) were able to find only one proposed Crp target in common with multiple cyanabacteria, so it was tested. His-Sycrp1 did not bind this target for $ccmK_{III}$ (slr1838). This result does not nullify the prediction that *ccmK* is targeted in multiple cyanobacteria because computational predictions presented here and by Xu and Su, (2009) (1) implicates SyCrp1 targeting to a region adjacent to the primary *ccmK* - sll1029. Xu and Su, (2009) (1) predicted a slr1109 target that is divergently transcribed from sll1027 followed by sll1028-1032 encoding *GltD*, $ccmK_{II}$, $ccmK_{I}$, *ccmM*, and *ccmN* respectively. The supplemental data presented here identifies the same proposed substrate sequence in addition to a high scoring (+)monad (+13)bend for $ccmK_{II}$. Consequently, *ccmK* targeting by SyCrp1 can not be discounted yet.

EMSA equilibrated to 500 nM His-SyCrp1 22 deg. C and subjected to electrophoresis at 4 deg. C with and without cAMP.
The exact same labeling prep. was used here as was used in the manuscript.
Coding strand sequences of dsDNA substrates not shown in Figure 1 of the main text are listed below:

	Substrate sequence	**Substrate name**
slr0194 *rpiA* is a negative control from Hedger et al., (2009) (1),	AACCGGAACTGTTCTGATAATGTTCGCACTGTAGAGATTT	slr0194
slr1838 *ccmK3* ; Crp site common to many cyanobacteria as predicted by Xu and Su, (2009)(2),	GCAACTAATTACCGTGGTCAACAGCACTAGGACCTTGCAG	slr1838
murF-DHM is simlar but longer than "object for bioinformatics" in Figure 1,	GTCAACGCAATTTTTATCTGTGATCTAGATCACAGATAAAATAGGCACCCTG	murF-DHM

murF-DHM Contains G/C rich flanking regions similar to those in ICAP to aid in base pairing during annealing.

Figures S4 and S5 Results

His-SyCrp1 binding to dsDNA containing (+)monad (+13)AAAA elements

DNA substrates sll1003, sll0710, and slr1512 were tested for sequence-specific binding to SyCrp1 (Figure S4) using a competitive electromobility gel shift assay (EMSA). Binding was sequence-specific because addition of the unlabeled specific competitor slr1667 decreased the amount of labeled DNA/His-SyCrp1 complex whereas addition of the unlabeled non-specific competitor Rndm did not. Binding to sll0447 and sll1241 was cAMP independent (Figure S5) and sequence specific (data not shown).

The doublet bands from sll1003, sll0710, and slr1512 (+)monad (+13)AAAA substrates in Figure S4 were reproducible in independent preparations that included annealing, gel purification, and labeling (data not shown). Omission of cAMP decreased binding to these but not sll0447 and sll1241 substrates (Figure S5).

The (+)monad (+13)AAAA determinant for His-SyCrp1 binding *in vitro*

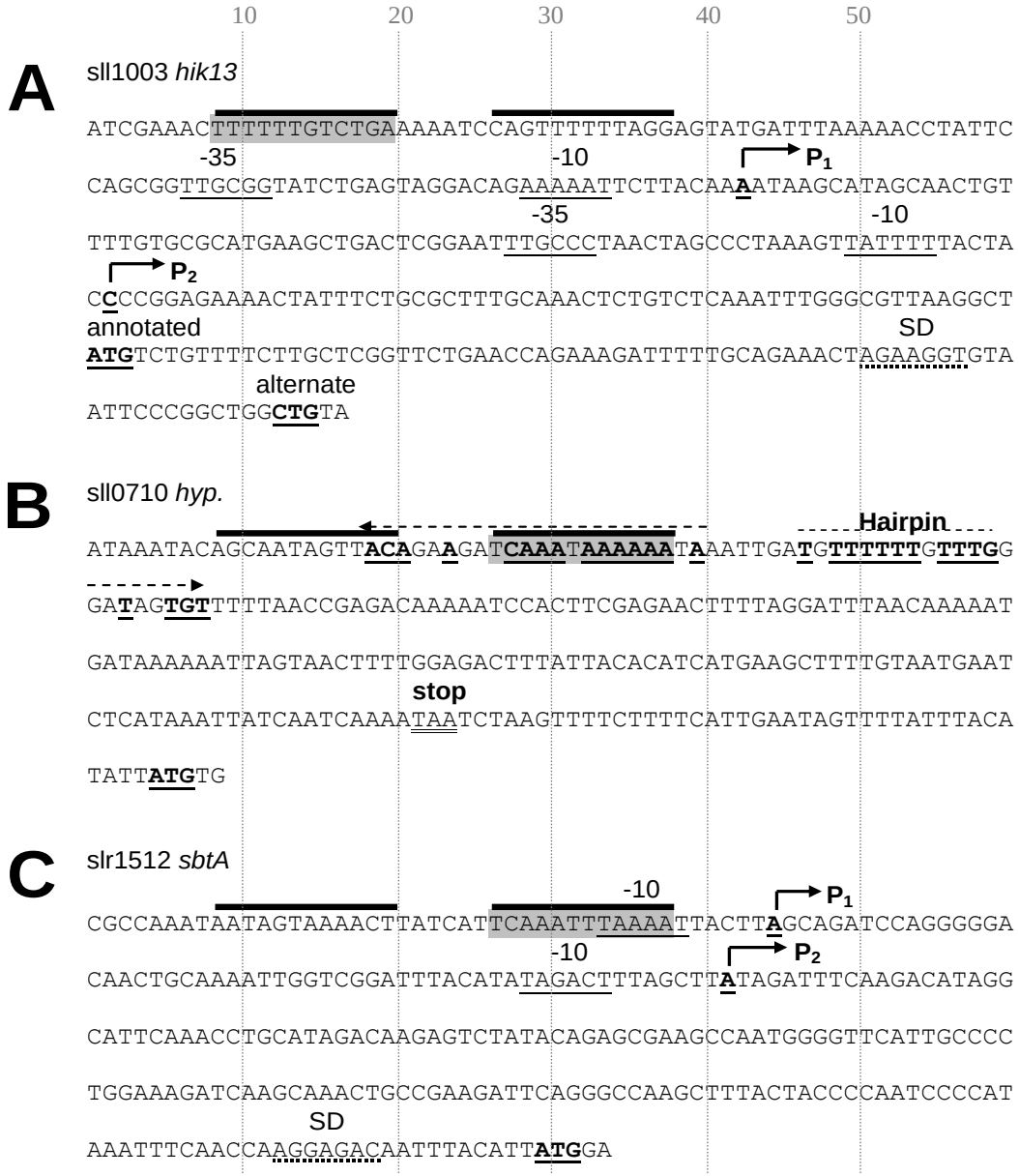
To determine if positions 8 – 19 downstream of the dyad core region are important for recognition by His-SyCrp1, the sequence of sll1268 in these positions was used to replace those in slr1512 as indicted in Figure 1 of the main text. Essentially, the flexible flanking bend [(+13)AAAA] was replaced with a less flexible sequence [(+13)GGCC] thereby abolishing detectable binding (Figure S4).

Reference List

1. Hedger,J., Holmquist,P.C., Leigh,K.A., Saraff,K., Pomykal,C. and Summers,M.L. (2009) Illumination stimulates cAMP receptor protein-dependent transcriptional activation from regulatory regions containing class I and class II promoter elements in *Synechocystis* sp. PCC 6803. *Microbiology*, **155**, 2994-3004.

2. Xu,M. and Su,Z. (2009) Computational prediction of cAMP receptor protein (CRP) binding sites in cyanobacterial genomes. *BMC. Genomics*, **10**, 23-39.

Figure S6



Promoter and 5' upstream regions containing (+)monad (±13)bend (grey box) SyCrp1 'targets' defined by sequence specific activator/cis-element binding *in vitro* and activator dependent transcription *in vivo*. (**A**) The histidine kinase, sll1003 (*hik13*), is located immediately downstream of a putative transposase. (**B**) The hypothetical protein, sll0710, is located downstream and inframe with a putative endonuclease sll0709. Putative hairpin terminator predicted RNA ΔG = -10.5, T_m = 68.7 °C (dashed arrow, complementary bases are **bold underlined**). (**C**) The sodium-dependent bicarbonate transporter, slr1512 (*sbtA*), is divergently transcribed from the photosystem II oxygen-evolving complex 23K protein PsbP homolog *psbP2*. Transcriptional start site results of RACE (Bolded and underlined with arrows and labeled $\mathbf{P}_n$), proposed –10 and –35 sigma factor binding regions (underlined), substrates in Figure 1 {regions underlined –D –C –B –A A B C D is indicated with thick overlines, while the (+)monad (+13)bend is boxed in grey}, annotated and alternate ATG translational start sites (**bold underlined**), proposed Shine-Dalgarno (dotted underline), and in frame stop codon (double underlined) are indicated as described.

Figure S6 Results and Methods

Open reading frames *sll1003, sll1241, sll0710,* and *sll0447* have met target criteria for SyCrp1 dependent transcription (1) and sequence specific binding (Figures S4 and S5).

Slr1512 also met the two required target criteria. Open reading frame *slr1512* transcript abundance measured by quantitative PCR was 10 times less abundant in WT cells than in *syrp1* null cells during inorganic carbon-replete growth conditions but 3 times more highly expressed in WT cells than in *syrp1* null cells during inorganic carbon-limiting conditions. Conversely, the *sll1732* negative-control transcript levels were identical between strains under these conditions (data to be presented elsewhere). Sequence specific binding was demonstrated in Figure S4.

Including SyCrp1 targets and -10 RNAP binding loci previously identified/suggested, the bp distance (termed aligned on-axis distance) from the dyad axis of symmetry to the middle of the -10 (denoted with an asterisk: TAT*AAT) for all known SyCrp1 target promoters are: sll1003; 68 bp and slr1512 (2); 68 bp (also determined here), slr1667; 152 bp (1), slr1351; 60 bp, slr0442; 32 bp, and though transcription criteria is lacking sll1268; 32 bp (3). A certain relationship among these aligned on-axis distances will be described elsewhere.

Strains, culture conditions, and RNA extraction

The motile glucose-sensitive *Synechocystis* sp. PCC 6803 was obtained from the Pasteur Culture Collection of cyanobacteria (PCC). Wild-type and sy*crp1* mutants were cultured at pH 8.0 and collected for RNA extraction as described for low to high intracellular cAMP conditions (3). For high to low inorganic carbon conditions, cultures bubbled with air mixed to 10% (v/v) CO_2 and maintained for 2 weeks of logarithmic growth in BG-11 supplemented with 75 mM TES pH 8.0 were washed twice and grown for 16 h with the same media lacking bicarbonate and bubbled with air passed through a 1.0 m soda lime column, then 1.0 M sodium hydroxide, and followed by 10.0 mM TRIS pH 8.0. Growth during the transition to CO_2 limitation without washing (data not shown) closely resembled that shown previously (4). Essentially, cells were grown on an orbital shaker in TES buffered BG-11 (5) at 30 °C and illuminated with 30.0 µmol photons $m^{-2}s^{-1}$ from cool white fluorescent lamps in both cases.

Rapid amplification of cDNA ends (RACE)

The +1 start site of transcription was determined for selected genes using primary and nested intragenic primers, respectively, for: sll1003 (*hik13*) TCATCGCTATGCTGGACAAA and GCTATGCTGGACAAATCGGTA, sll0710 (*hyp.*) CGATCAAACTTCCTTAGTGTTCG and CACGGACTTTATTGAATTTGTCG, slr1512 (*sbtA*) CAAGGGCGGCAATAACCAT and CCCAATCAGAAAGGCTAGGG. RACE is a ligation mediated PCR reaction that ligates a known primer anchor to cDNA generated by reverse transcriptase (3). However, RACE is a qualitative technique that should not be confused with the quantitative 'primer extension' technique that can also be used to identify transcriptional start sites.

Reference List

1. Yoshimura,H., Yanagisawa,S., Kanehisa,M. and Ohmori,M. (2002) Screening for the target gene of cyanobacterial cAMP receptor protein SYCRP1. *Mol. Microbiol.*, **43**, 843-853.

2. Lieman-Hurwitz,J., Haimovich,M., Shalev-Malul,G., Ishii,A., Hihara,Y., Gaathon,A., Lebendiker,M. and Kaplan,A. (2008) A cyanobacterial AbrB-like protein affects the apparent photosynthetic affinity for CO by modulating low-CO-induced gene expression. *Environ. Microbiol.*, **11**, 927-936.

3. Hedger,J., Holmquist,P.C., Leigh,K.A., Saraff,K., Pomykal,C. and Summers,M.L. (2009) Illumination stimulates cAMP receptor protein-dependent transcriptional activation from regulatory regions containing class I and class II promoter elements in *Synechocystis* sp. PCC 6803. *Microbiology*, **155**, 2994-3004.

4. Wang,H.L., Postier,B.L. and Burnap,R.L. (2004) Alterations in global patterns of gene expression in *Synechocystis* sp. PCC 6803 in response to inorganic carbon limitation and the inactivation of *ndhR*, a *lysR* family regulator. *J. Biol Chem*, **279**, 5739-5751.

5. Stanier,R.Y., Kunisawa,R., Mandel,M. and Cohen-Bazire,G. (1971) Purification and properties of unicellular blue-green algae (order Chroococcales). *Bacteriol. Rev.*, **35**, 171-205.

Figure S7

Titration of murF-DHM (downstream-half-mirror) with increasing amounts of Sycrp1. The relative distance from the wells *x* for bound murF-DHM bands starting at the top wells are 0.8, 1.4, and 2.6 cm respectively. Setting *y* = 3, 2, and 1 respectively yields $y = 4.641e^{-0.5893x}$, and $R^2 = 0.9992$ indicative of 1, 2, and 3 bound SyCrp1 homodimer/DNA substrate complexes, The free dsDNA band migrated 4.5 cm. The dsDNA murF-DHM substrate comprises 38% of the SyCrp1/murF-DHM complex mass. A SyCrp1 homodimer is 52.2 kDa and dsDNA murF-DHM is 31.9 kDa. These gels were reproduced and quantified in triplicate.

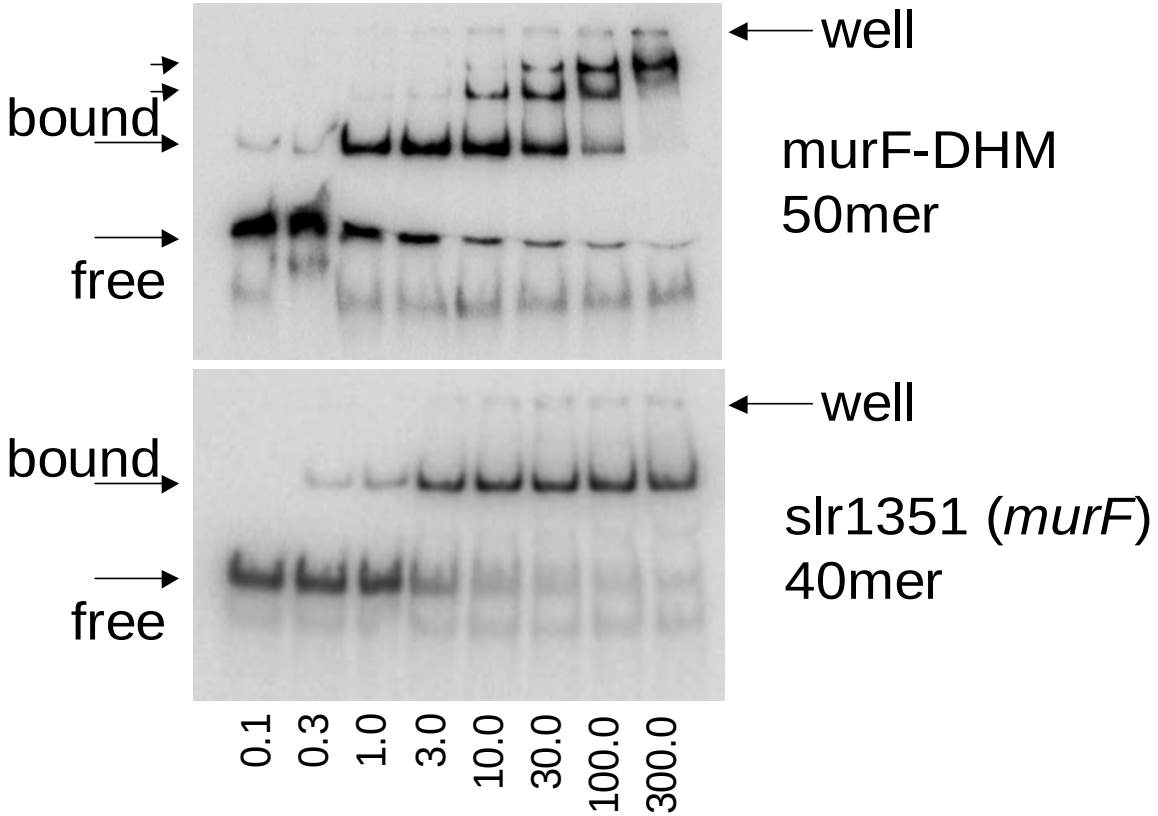


Figure S8

Lindemose *et. al*., (2008) EcCrp/XD-DNA

A Major mode PSSM #A from all 49 substrates

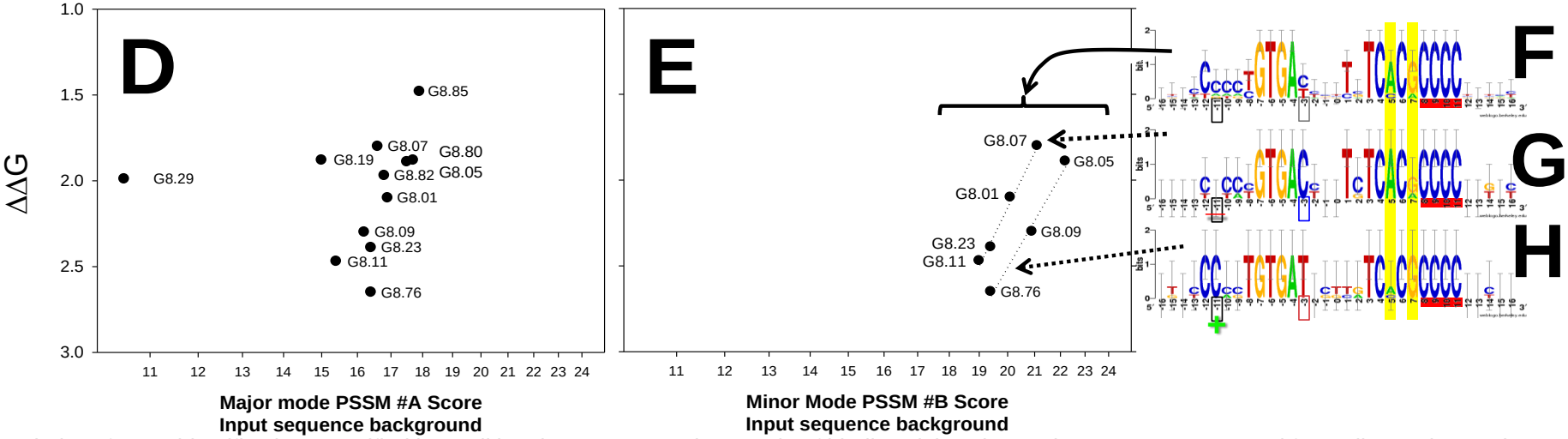
A |	13	4	11	9	6	11	5	8	0	0	0	0	48	2	8	1	5	2	5	0	0	41	0	26	0	1	0	2	1	11	10	8	17
C |	8	25	15	17	25	26	33	29	16	0	2	0	1	24	23	32	11	10	16	0	49	5	49	1	32	39	33	26	22	16	13	12	13
G |	12	11	12	6	12	8	2	4	0	49	0	49	0	1	2	8	20	5	28	1	0	0	0	20	4	4	0	2	10	6	9	10	6
T |	16	9	11	17	6	4	9	8	33	0	47	0	0	22	16	8	13	32	0	48	0	3	0	2	13	5	16	19	16	16	17	19	13

B Minor Mode PSSM #B from these substrates listed here in C.

A |	2	0	1	1	0	1	1	1	0	0	0	0	7	0	1	0	0	0	1	0	0	6	0	1	0	0	0	0	0	2	0	2	1
C |	2	2	2	4	6	5	5	5	2	0	0	0	0	4	3	3	2	1	3	0	7	1	7	0	7	7	7	7	2	1	2	3	3
G |	1	2	2	0	0	1	0	1	0	7	0	7	0	0	0	2	2	0	3	0	0	0	0	6	0	0	0	0	2	2	2	2	0
T |	2	3	2	2	1	0	1	0	5	0	7	0	0	3	3	2	3	6	0	7	0	0	0	0	0	0	0	0	3	2	3	0	3

C

Substrate sequence	PSSM #A Score	PSSM #B Score	ΔΔG	Substrate
CGTCTCCCCGTGACTCCTGTCACGCCCCTATAT	16.6	21.90	1.80	G8.07
TCCTCCCCTGTGACCTGTCTCACGCCCCGTGCC	16.9	25.50	2.10	G8.01
ACGCCATATGTGACTGCTCTCACACCCCTGTCC	16.4	20.00	2.39	G8.23
TTGACGCCCGTGACCTTTCTCACGCCCCCCGGT	15.4	21.60	2.47	G8.11
CTCCCCCCTGTGATCCTTGTCACGCCCCGTCAT	17.5	24.50	1.89	G8.05
GGTCCCCGTGTGATTGTTGTCCCGCCCCTGTGA	16.2	19.50	2.30	G8.09
ATATCCACTGTGATACGCATCACGCCCCCACCC	16.4	19.70	2.65	G8.76



A priori conformer identification quantified by traditional means as a minor mode of binding. (**A**) Major mode PSSM #A generated from all 49 unique substrate sequences (clone G#'s labeled) identified by Lindemose *et al*., (2008). (**B**) Minor mode PSSM #B generated from sequences in panel C. (**C**) The 7 substrates containing the sequence characteristic (+10)CCCC but not (-10)CCCC. (**D**) Dyad (+10)bend conformer. The 12 substrates containing (+10)CCCC scored with PSSM #A. (**E**) The 7 substrates sequences in panel C scored with PSSM #B. XD-DNA substrate sequences (represented as X = G, and D = A) containing a canonical flexible (+10)bend, but lacking a (-10)bend. This conformer is the minor binding mode defined by PSSM #B (all sequences shown) and minor sub-modes (dotted lines) contained therein. (**F**) Minor mode PSSM #B. This group of sequences contains the 7 minor mode substrates shown in panels C and E. (**G**) Minor sub-mode (dotted line, left in panel E) conservation at positions -3 and -11 (boxed) is negatively (red -) correlated or (**H**) (dotted line, right in panel E) positively (green **+**) correlated. Dependent positions -3 and -11 are boxed. Small sample corrected logos [includes e(*n*)] and error bars are shown. Yellow column shading in logos indicates primary kink associated positions. The base frequency background used for PSSM scoring is unique to each plot and consisted of the percent base composition of the total substrates shown for that plot (input seq. background). Note logarithmic scaling of the abscissa.

Figure S9

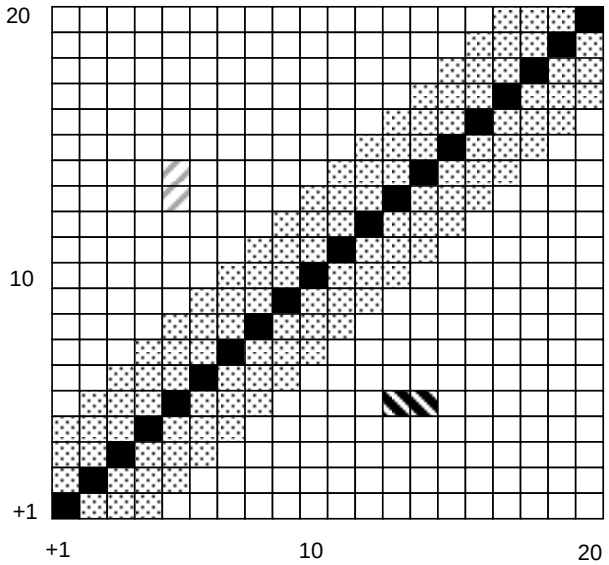


Figure S9 shows a standard interaction tensor matrix curtailed to two dimensions and to positions +1 to +20. More involved tensor matrices are available online (1). The assumption that interactions between positions is negligible, that the information used to generate a PSSM score is position independent and involves only the diagonal elements (2,3), is indicated by the black boxes. Our study indicates that this assumption, when applied to protein/DNA binding that results in curved DNA, is insufficient to yield a tight linear relation between $\Delta\Delta G$ and log calculated W_S. Were this an amino acid-amino acid interaction matrix for protein folding, so many off axis positions would be important that the mathematics would become insurmountable. Even a hidden Markov chain window size commonly 4 bp, shown with dotted boxes, is computationally expensive and excludes longer range interposition-dependencies. However, we show that very few interposition interactions, such as the hatched square indicating how the bases at positions +13 and +14 influences the information at position +5, can lead to a quite linear relation between $\Delta\Delta G$ and $W_{\psi S}$ (see main text and Figure S8). Our data suggests that positions above the black diagonal, i.e., lightly hatched position indicating how position +5 influences information at position +13, are probably negligible. The logos fitted to sine waves (see main text) also indicate that influential couplings are likely to occur every 8.5 or ≈ 10.6 base pairs depending on the conformer. Thus limited, we simplify the testing of off-diagonal couplings especially when a novel conformer is tested. Finally, the off-axis couplings increase the W_S of the binding sequences by a factor unaccounted for by the diagonal. The HS-algorithm artificially and conditionally transfers most of this lost information back to the diagonal so that traditional PSSM scores using only the black diagonal suffice by using HS-generated ψ-sequences.

1. Bindewald,E., Schneider,T.D. and Shapiro,B.A. (2006) CorreLogo: an online server for 3D sequence logos of RNA and DNA alignments. *Nucleic Acids Res.*, **34**, W405-W411.

2. Hertz,G.Z. and Stormo,G.D. (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics.*, **15**, 563-577.

3. Fields,D.S., He,Y., Al Uzri,A.Y. and Stormo,G.D. (1997) Quantitative specificity of the Mnt repressor. *J. Mol. Biol.*, **271**, 178-194.

Figures S8 and S9 Results and Discussion

Hidden state and PSSM #3 validation

For regular gel shift experiments having 1 protein complex bound to 1 dsDNA segment, the information of any given DNA sequence and affinity for binding to a transcription factor are quantifiably related by the laws of equilibrium thermodynamics (within the equilibrium of heterogeneous substances) by the change in binding affinity (ΔG) from a reference complex ($\Delta\Delta G$) such as the Crp/ICAP complex (1,2).

The hidden state (HS) model is inherent to the above proposal and should be testable by comparing $\Delta\Delta G$ to the DNA sequence information determined by the PSSM score. Experimentally determined Crp/DNA binding affinities have been reported for a set of DNA substrates derived from intergenic DNA sequences in *Synechocystis* (3). Consequently, affinity measurements ($\Delta\Delta G$) and predicted DNA sequence information (PSSM weight score W_s) can be tested for relatedness. However, both PSSM #3 (see main text) and the hidden state (HS) HS-model are both based on predictions. They are hypothesis that require testing. To first validate the HS-model, the experimentally determined Crp/DNA binding affinities following EcCrp SELEX with XD-DNA (see results) were tested for their relation to log W_s by using the PSSM #A (Figure S8A) resulting from all DNA sequences obtained with SELEX (4). Using the same methods, SyCrp1 binding affinities for DNA substrates encoded in the *Synechocystis* genome (3) were also tested for their relation to information content by using PSSM #3. In the EcCrp/XD-DNA system, the global major mode PSSM #A is known as all substrates collected by SELEX whereas the HS-algorithm requires validation. Once tested for PSSM score *vs*. $\Delta\Delta G$ correlation improvement, the HS-model is then validated and can be applied to the SyCrp1/DNA system where PSSM #3 required validation.

Figures S8 and S9 Results and Discussion

Sequence space and attractors

If there is one binding mode, then there will be one attractor (best binding sequence) in sequence space around which substrate sequences will cluster (5). Together, these sequences will generate a sequence of most frequently occurring base identities and a PSSM that serves to both approximate the attractors' location and provide a metric for how closely related each sequence is to this approximated attractor (1). A metric for these approximations is a PSSM weight score (W_s). Ideally, the larger the PSSM calculated W_s, the tighter should be the binding affinity (measured here as $\Delta\Delta G$) with the tightest binding being that of the attractor that can be estimated with a major mode PSSM. The sequence of the most frequently occurring base identities of major mode is an estimate of the attractor in much the same way that a population sample's mean is an estimate of the population mean. Within one minor binding mode, W_s and $\Delta\Delta$G should be highly correlated. However, if two minor binding modes exist, each with their own unique attractor, then, for the major mode of binding, a PSSM will describe some average point in sequence space between the two attractors. Then the correlation between W_s and $\Delta\Delta G$ will be diminished relative to that of a single mode containing only one attractor (e.g. ▲ in Figure 4B in main text). If two sequence groups could be *a priori* distinguished (e.g., either containing or lacking a bend) as being members of two distinct minor binding modes forming distinct conformers (Figure 4D, and E) within a major mode (Figure 4C), then generating a minor mode specific PSSM and subsequently scoring individual DNA substrate sequences of that minor mode with the minor mode specific PSSM should show a higher intra-group correlation of minor mode PSSM calculated W_s and $\Delta\Delta G$ (Figure S8E) than does the global correlation of all minor modes using a single major mode PSSM (Figure S8D).

Overlapping minor modes of binding

To elucidate EcCrp/XD-DNA binding modes, all 49 unique dsDNA substrate sequences captured by EcCrp/XD-DNA SELEX (comprising the major mode) were aligned (Figure 4C) and separated into two distinct sequence element groups (minor modes). One minor mode sequence group contained (+10)CCCC (Figures 4D and S8D), the other minor mode contained ≤ 2 X/C bp's total in both (+ and -10)NNNN tracts and essentially represented conservation of the dyad core alone (Figure 4E).

To first identify a distinct minor binding mode using traditional computational methods, the minor mode substrate group containing (+10)CCCC (Figures 4D and S8D) elements was scored with the major mode PSSM #A (Figuress 4C and S8A) and plotted relative to the empirically determined $\Delta\Delta G$ (Figure S8D). The (+10)CCCC minor mode sequence group/conformer [dyad (+10)bend in Figure 4D] was further divided into a distinct sub-group containing (+10)CCCC but not containing (-10)CCCC elements. This resultant minor mode group was then scored with the minor mode PSSM #B (Figure S8B) generated with the sequences of this group (Figure S8C) and plotted relative to the empirically determined $\Delta\Delta G$ (Figure S8E). Given that the sequences used to generate these PSSM's were those of the group analyzed, the relationship between $\Delta\Delta G$ and PSSM calculated weight scores in Figure S8D and S8E changed to yield the high correlation of $\Delta\Delta G$ and PSSM #B in Figure S8E. Two parallel minor sub-modes were suggested because there were two separate relationships that could be drawn (dotted lines). These minor sub-modes contain "C" at position -3 (left dotted line) "T" at position -3 (right dotted line) and showed a distinction between appropriate flexibility *vs.* appropriate DNA element phasing (C *vs*. T) respectively at position -3 in the spacer region (Figuress S8G and S8H). This result was consistent with PSSM #B identification of an average point in sequence space between two attractors for two separate minor sub-modes. However, these two parallel groupings were instead members of a single minor mode containing overlapping minor sub-modes as was demonstrated using the HS-algorithm (see Figure 4B).

Anisotropy and interposition-dependence during DNA bending

To allow a universal (major mode) PSSM to become applicable to all sequences, a thermodynamic landscape was created that allowed inter-conformer energetic comparisons within a ψ-sequence data set (Figures 3B, and 4B). The dyad (+10)bend conformer data fitting in the ψ-plot is precise because primary kink position +5 base identity is not conserved and is irrelevant to affinity. Conversely, position +5 is relevant in the dyad conformer because the most frequently occurring base identity is highly conserved, presumably due to primary kink formation that decreases binding affinity. Primary kink and primary bend conformers have been observed in crystal structures (6). The ψ -sequence thermodynamic landscape thus supports primary kink formation with XD-DNA at position +5 in the dyad, but smooth bending instead in the dyad (+10)bend conformer.

Figures S8 and S9 Results and Discussion

DNA bending enables long-range position-dependency. The entropic properties of dsDNA bound by DNA binding proteins have been generally described by computational modeling (8) and verified by experimentation (9). These studies used molecular dynamics modeling and nuclear magnetic resonance to demonstrate a wider dsDNA vibrational spectrum free dsDNA than for protein-bound dsDNA. For DNA binding molecules that do not bend dsDNA (10) such as the lambda (8) or Mnt repressors (11), the HS-model and hidden Markov model are not required because the information of a given dsDNA substrate sequence is for the most part interposition-independent, likely does not participate in quasi-harmonic entropy to any great extent, and is thus contained along the analysis diagonal of Stormo. However, when multiple *cis*-elements overlap (12), or the protein bends DNA (13), then information is no longer completely contained within the position specific diagonal of Stormo. The hidden Markov model suffices (12) when interposition-dependencies are mostly limited to adjacent or nearby positions because the complexed dsDNA is linear. When the protein bends dsDNA (13,14), the hidden Markov model is insufficient because interdependent positions are not necessarily adjacent or nearby. Thus, the HS-model is useful when the dsDNA vibrational spectrum is decreased by bending.

For proteins that do not bend DNA, the length L of DNA that contacts the protein is limited and constant. The binding sequences likely evolve one base at a time and the position-independent weight scores are quite amiable to W_s and binding affinity being highly correlated. For proteins like Crp that bend DNA, the DNA wraps around the protein and may add flanking sequences for further wrapping. This involves adding ancillary protein surfaces whilst relinquishing binding to traditionally involved surfaces. Such a mechanism not only allows for changes of L but likely involves saltation, a cause and effect process jumping over certain bp steps, in addition to individual steps through sequence space. Saltation makes traditional searches for a global binding matrix difficult while the added flanking sequences add a covariance to position pairs many bases distant from one another. This distance precludes using standard Markov analysis that detects covariance of positions close to one another. For such binding proteins that bend DNA, both novel methods for generating a global PSSM and then relating its W_s [or individual information (15)] to affinity are required.

Once a global PSSM model has been generated, it is a straight forward approach to finding conformers. Conformers fit the sine curves tracing sequence logos. Positions considered for HS-algorithms either deviate from that sine curve, or fit the sine curve and 'drag-along' other associated sequence elements. So once these sine curves are found, there is no longer a need to follow DNA evolution from the beginning of the path it took in sequence space to where it is now. Instead we start from where DNA evolution is now and work backwards from the global PSSM to the minor modes and minor sub-modes to relate all populations quantifiably no matter what laboratory performs the experiments as long as the same reference standard (e.g., ICAP) is used. It doesn't matter whether or not the DNA sequence 'remembers' the path it took to get to where it is now, currently, interacting with a DNA bending protein. The current conformer modes of binding and corresponding affinities of these populations can be found and calculated by sampling a small experimental data set to test a hypothetical major mode PSSM model.

Utilizing the anisotropic properties of some substance intrinsic to a system to identify distinct binding modes and thereby describe conformational selection involved in protein binding is not unique to the work here. One major mode and multiple minor modes of ubiquitin binding are described based on residual dipolar coupling measurements that can be obtained only by limiting certain axial rotations within a system during NMR using an anisotropic dilute crystalline or polyacrylamide medium instead of isotropic buffer alone (16). Though not as technically involved, the same basic principles have been applied in the work here. Here, the anisotropic substance was DNA. Its limited range of motion caused decreased conservation of positions 5, and 7 (TCA_5CA_7) that could then be measured using PSSM calculated weight scores and sequence logos. These measurements then lead to the description of multiple minor binding modes each composed of related substrate sequences coexisting within one major binding mode of conformational selection.

Figures S8 and S9 Results and Discussion

Reference List

1. Hertz,G.Z. and Stormo,G.D. (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics.*, **15**, 563-577.

2. Berg,O.G. and von Hippel,P.H. (1988) Selection of DNA binding sites by regulatory proteins. II. The binding specificity of cyclic AMP receptor protein to recognition sites. *J. Mol. Biol.*, **200**, 709-723.

3. Omagari,K., Yoshimura,H., Suzuki,T., Takano,M., Ohmori,M. and Sarai,A. (2008) DeltaG-based prediction and experimental confirmation of SYCRP1-binding sites on the *Synechocystis* genome. *FEBS J.*, **275**, 4786-4795.

4. Lindemose,S., Nielsen,P.E. and Mollegaard,N.E. (2008) Dissecting direct and indirect readout of cAMP receptor protein DNA binding using an inosine and 2,6-diaminopurine *in vitro* selection system. *Nucleic Acids Res.*, **36**, 4797-4807.

5. Eigen,M., Winkler-Oswatitsch,R. and Dress,A. (1988) Statistical geometry in sequence space: a method of quantitative comparative sequence analysis. *Proc. Natl. Acad. Sci. U. S. A*, **85**, 5913-5917.

6. 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.

7. Eddy,S.R. (2004) What is a hidden Markov model? *Nat. Biotechnol.*, **22**, 1315-1316.

8. Dixit,S.B., Andrews,D.Q. and Beveridge,D.L. (2005) Induced fit and the entropy of structural adaptation in the complexation of CAP and lambda-repressor with cognate DNA sequences. *Biophys. J.*, **88**, 3147-3157.

9. Tzeng,S.R. and Kalodimos,C.G. (2009) Dynamic activation of an allosteric regulatory protein. *Nature*, **462**, 368-372.

10. Carlson,C.D., Warren,C.L., Hauschild,K.E., Ozers,M.S., Qadir,N., Bhimsaria,D., Lee,Y., Cerrina,F. and Ansari,A.Z. (2010) Specificity landscapes of DNA binding molecules elucidate biological function. *Proc. Natl. Acad. Sci. U. S. A*, **107**, 4544-4549.

11. Fields,D.S., He,Y., Al Uzri,A.Y. and Stormo,G.D. (1997) Quantitative specificity of the Mnt repressor. *J. Mol. Biol.*, **271**, 178-194.

12. Drawid,A., Gupta,N., Nagaraj,V.H., Gelinas,C. and Sengupta,A.M. (2009) OHMM: a Hidden Markov Model accurately predicting the occupancy of a transcription factor with a self-overlapping binding motif. *BMC. Bioinformatics.*, **10**, 208.

13. Roulet,E., Busso,S., Camargo,A.A., Simpson,A.J., Mermod,N. and Bucher,P. (2002) High-throughput SELEX SAGE method for quantitative modeling of transcription-factor binding sites. *Nat. Biotechnol.*, **20**, 831-835.

14. Nagaraj,V.H., O'Flanagan,R.A. and Sengupta,A.M. (2008) Better estimation of protein-DNA interaction parameters improve prediction of functional sites. *BMC. Biotechnol.*, **8**, 94.

15. Schneider,T.D. (1997) Information content of individual genetic sequences. *J. Theor. Biol.*, **189**, 427-441.

16. Lange,O.F., Lakomek,N.A., Fares,C., Schroder,G.F., Walter,K.F., Becker,S., Meiler,J., Grubmuller,H., Griesinger,C. and de Groot,B.L. (2008) Recognition dynamics up to microseconds revealed from an RDC-derived ubiquitin ensemble in solution. *Science*, **320**, 1471-1475.