Published online 18 December 2006 *Nucleic Acids Research, 2007, Vol. 35, No. 2* ***623–633***
doi:10.1093/nar/gkl1095

Analysis of the DNA-binding sequence specificity of the archaeal transcriptional regulator Ss-LrpB from *Sulfolobus solfataricus* by systematic mutagenesis and high resolution contact probing

Eveline Peeters, Carine Wartel, Dominique Maes[1] and Daniel Charlier*

Erfelijkheidsleer en Microbiologie and [1]Laboratorium voor Ultrastructuur, Vrije Universiteit Brussel and Vlaams interuniversitair Instituut voor Biotechnologie (VIB), Pleinlaan 2, B-1050 Brussel, Belgium

Received October 20, 2006; Revised November 22, 2006; Accepted November 24, 2006

ABSTRACT

To determine the sequence specificity of dimeric Ss-LrpB, a high resolution contact map was constructed and a saturation mutagenesis conducted on one half of the palindromic consensus box. Premodification binding interference indicates that Ss-LrpB establishes most of its tightest contacts with a single strand of two major groove segments and interacts with the minor groove at the center of the box. The requirement for bending is reflected in the preference for an A+T rich center and confirmed with C·G and C·I substitutions. The saturation mutagenesis indicates that major groove contacts with C·G at position 5 and its symmetrical counterpart are most critical for the specificity and strength of the interaction. Conservation at the remaining positions improved the binding. Hydrogen bonding to the O^6 and N^7 acceptor atoms of the $G_{5'}$ residue play a major role in complex formation. Unlike many other DNA-binding proteins Ss-LrpB does not establish hydrophobic interactions with the methyls of thymine residues. The binding energies determined from the saturation mutagenesis were used to construct a sequence logo, which pin-points the overwhelming importance of C·G at position 5. The knowledge of the DNA-binding specificity will constitute a precious tool for the search of new physiologically relevant binding sites for Ss-LrpB in the genome.

INTRODUCTION

Understanding the sequence-specific binding of a transcription regulator is of central importance in order to unravel the functioning of this protein in the establishment of the regulatory process. *In vivo* any sequence-specific DNA-binding protein will also bind to pseudo-sites with a reduced affinity. These pseudo-sites nevertheless play an important role both thermodynamically, by setting the concentration of freely available regulatory protein and kinetically, by regulating the rate of location of the specific targets. Furthermore, low affinity sites may have a considerable impact on the outcome of the regulatory response if they occur in proper juxtaposition with a high affinity site and are bound in a cooperative fashion. A detailed knowledge of the sequence specificity is therefore required to search for new binding sites and to distinguish between physiologically relevant regulatory sites and pseudo-sites.

At the present time, archaeal regulation is still poorly documented. The lack of efficient tools in genetics and molecular biology, especially for hyperthermophilic archaea, constitutes a severe limitation for the analysis of regulator function and the identification of regulons, modulons and stimulons. Nevertheless, the (potential) binding sites of a handful archaeal regulators have been identified in upstream promotor/operator regions of their respective target genes through *in silico* or experimental approaches (*in vitro* or more rarely *in vivo*) [for a review on archaeal transcription regulation, see Ref. (1)]. Most of these sites are semi-palindromic.

Several characterized archaeal regulators belong to the archaeal/bacterial Lrp/AsnC family (2). Lrp-like regulators are widely distributed among archaea, but with the exception of LysM from *Sulfolobus solfataricus*, the physiological role of these potential regulators in archaea remains mainly elusive (3). Lrp-like proteins show a variable degree of amino acid sequence identity, but share the same fundamental architecture. Crystal structure determination has been done for the archaeal members LrpA (4) and FL11 (5), both from *Pyrococcus* species and for the bacterial members LrpC from *Bacillus subtilis* and AsnC from *Escherichia coli* (6). The N-terminal helix–turn–helix DNA-binding

*To whom correspondence should be addressed. Tel: +32 2 629 13 42; Fax: +32 2 629 13 45; Email: dcharlie@vub.ac.be

domain is connected with a flexible linker to the C-terminal domain that shows a typical αβ sandwich fold, also called RAM domain (regulation of amino acid metabolism) (7). The latter is involved in effector binding and oligomerization. In solution, Lrp proteins exist as dimers or oligomers of dimers. They usually bind cooperatively to operators carrying an array of degenerate semi-palindromic targets (2).

Ss-LrpB is an Lrp-like protein from *S.solfataricus*. This regulator binds its own operator region at three similar, regularly spaced 15-bp binding sites (8). The deduced palindromic consensus sequence, 5′-**TTG**CAAAATT**TG**CAA-3′, has four highly conserved base pairs (in bold) in each half-site and a 5-bp long central region exclusively composed of weak base pair. A recent AFM (atomic force microscopy) study of Ss-LrpB:operator complexes indicates that each binding site is contacted by an Ss-LrpB dimer (9). Furthermore, occupation of the three binding sites results in the formation of a globular complex in which ~100 bp of the operator DNA are wrapped around the interacting regulator molecules.

The DNA-binding sequence specificity of a regulator can be determined by using the SELEX strategy (systematic evolution of ligands by exponential enrichment) (10). This technique will select a set of high affinity DNA sites that define a consensus-binding sequence for the protein. It has been applied for the Lrp-like regulators *E.coli* Lrp and Ptr1 and Ptr2 from *Methanocaldococcus jannaschii* (11,12). In the present study we have followed another strategy. We determined the DNA-binding sequence specificity of Ss-LrpB by measuring the *in vitro* binding to a set of mutated variants of the idealized symmetrical consensus box. This was done systematically for all possible single base pair substitutions of one half of the binding site (saturation mutagenesis) and also for targets containing an abasic position or a non-canonical base: inosine, uracil, 5-methyl cytosine or 2-aminopurine. The importance of the spacing between the two consensus-half-sites was assessed by including single and double base pair deletions and insertions. Binding profiles were constructed based on electrophoretic mobility shift assay (EMSA) experiments, allowing an estimation of the apparent binding equilibrium dissociation constants (K_D). The quantitative binding data from the saturation mutants were represented in an energy normalized sequence logo (13,14). Combined with the results of a high resolution contact probing analysis of the Ss-LrpB:consensus box interaction, this provides a detailed view of how each base pair energetically contributes to the specific binding. Such an extended and detailed experimental analysis had not yet been performed for an archaeal regulator or a bacterial/archaeal Lrp-like regulator.

MATERIALS AND METHODS

Protein purification

Recombinant Ss-LrpB protein was produced in *E.coli* and purified by a combination of heat treatment and ion exchange chromatography as described previously (9). The protein concentration was determined by a MicroBCA assay (Pierce). The purified protein was divided into small aliquots, which were frozen and thawed individually before each EMSA experiment.

Footprinting and binding interference analysis

Footprinting and binding interference analysis were performed with a 150-bp DNA fragment with the consensus box near the center. This fragment was generated by PCR using the plasmid pBendCon as template and the oligonucleotides EP9 and EP10 as primers (8). PCRs were performed by using ReadyMix *Taq* PCR Mix (Sigma–Aldrich). One of the oligonucleotides was 5′ end labeled with [γ-^{32}P]ATP (GE Healthcare Bio-Sciences) and T4 polynucleotide kinase (Roche). The resulting DNA fragments (either top strand or bottom strand labeled) were purified by PAGE prior to analysis.

DNase I footprinting (15), in-gel Cu-OP footprinting (8) and premethylation, deoxyuridine substitution, depurination and depyrimidation binding interference experiments (16) were all performed as described previously. All binding reactions were done in LrpB binding buffer [20 mM Tris–HCl (pH 8.0), 1 mM $MgCl_2$, 0.1 mM dithiothreitol, 12.5% glycerol, 50 mM NaCl and 0.4 mM EDTA]. Using chemical sequencing reference ladders were generated (17).

EMSA

EMSA experiments were performed either with the labeled 150-bp fragment, as described above, or with 45-bp oligonucleotide duplexes (annealing of complementary oligonucleotides) of which one oligonucleotide was 5′ end labeled with [γ-^{32}P]-ATP and T4 polynucleotide kinase (Roche). The latter applies for all mutated variants of the consensus box. These 45-bp oligonucleotides contain the wild-type (TAAAAAGGCATTATCTTGCAAAATTTGCAATAATC-CTTTTATGTT) or mutated consensus sequence starting at position 16, preceded and followed respectively by 15-bp of the sequences upstream and downstream of the naturally occuring strong Box1 in the *Ss-lrpB* promotor/operator region. All oligonucleotides have been purchased from Sigma-Aldrich, except the abasic variants (Eurogentec) and the 2-aminopurine variant (VBC Biotech Services GmbH).

EMSAs were performed as described previously (15). All binding reactions proceeded at 37°C in LrpB binding buffer (see above) and in the presence of a large excess of non-specific competitor DNA (25 μg/ml sonicated herring sperm DNA).

Data analysis

All EMSA autoradiographs were scanned and the integrated band intensities were quantified using the Intelligent Quantifier software (Bio Image). For each lane, the background intensity was also measured and subtracted from the band intensities. Only the integrated intensities (I.I.) from the unbound DNA bands were considered for further analysis because of the 'smearing' effect. All I.I. values where divided by the average of two I.I. measurements of free DNA (without addition of Ss-LrpB). This corresponds to the fraction of unbound DNA. The fraction of bound DNA in each lane was then calculated to be: fraction bound DNA = 1 − fraction unbound DNA. These quantitated data were plotted versus Ss-LrpB concentration (binding profile) and fitted using the Hill equation (Origin, non-linear least squares method):

$$\text{Fraction bound DNA} = v_{max} \cdot [\text{Ss-LrpB}]^n / (k^n\ [\text{Ss-LrpB}]^n) \quad \mathbf{1}$$

In this equation, n corresponds to the Hill coefficient, which is a measure of binding cooperativity. In several cases, v_{max} was <1 and therefore k did not equal the apparent K_D. The apparent K_D was instead determined to be the protein concentration at which the fraction of bound DNA equals 0.5. EMSAs were repeated several times (at least twice, mostly four times) for each oligonucleotide duplex variant. Visible outliers, due to experimental errors such as pipetting errors, were omitted prior to plotting. Typically, K_D values determined from EMSA experiments can vary by as much as 2-fold (18). The largest source of errors in determining apparent K_D values from EMSAs is the size of increments in the used protein concentrations. We have been particularly careful to keep these increments low, generally 1.5-fold, with a maximum of 2-fold. This led to variations comprised between 20 and 50% for good binding sites and to higher values for very low affinity binding sites.

Sequence logos were constructed with the web-based tool enoLOGOS, using relative entropy [(14); available at http://biodev.hgen.pitt.edu/enologos]. This tool was used to create a logo from the aligned forward and reverse sequences from the naturally occuring Ss-LrpB binding sites. A logo was also created based on the binding affinity data. A position weight matrix was constructed with binding energy data from the saturation mutagenesis analysis. Binding energies (expressed in kT units) were calculated to be:

$$E = \ln K_D. \quad \mathbf{2}$$

RESULTS AND DISCUSSION

High resolution contact probing

Previously it has been shown that Ss-LrpB binds a 150-bp fragment bearing the consensus-binding site with an apparent K_D of ~20 nM (8). Here, the study of this interaction is extended by enzymatic and chemical footprinting and by premodification binding interference assays (Figure 1). Ss-LrpB protected on both strands a 22 nt long stretch against DNase I cleavage. This stretch covers the 15-bp consensus box and extends 5 nt towards the 3′ end and 2 nt towards the 5′ end (Figure 1a and f). Protection on both strands is therefore offset by 3 nt towards the 3′ end; this reflects how DNase I is positioned in the minor groove and cleaves phosphodiester bonds on opposing strands across the minor groove. The presence of hyperreactive sites at the 3′-boundaries of the consensus box are indicative of Ss-LrpB-induced DNA deformations resulting in local minor groove widening. The limits of protection areas and the sites of reactivity and hyperreactivity are fully symmetrical (Figure 1f). This indicates that specific binding results in the alignment of the 2-fold symmetry of the Ss-LrpB dimer to the dyad axis of symmetry of the palindromic binding site.

In order to obtain a higher resolution footprint, the nuclease activity of the smaller 1, 10-phenanthroline-copper ion [$(OP)_2$-Cu^+] (19) was used for in-gel footprinting of Ss-LrpB: consensus box DNA complexes separated from bare DNA by gel electrophoresis (Figure 1b). This resulted in a protection area that is on both strands limited to the 15-bp consensus sequence (Figure 1b and f). Remarkably, in the DNase I and in-gel footprinting experiments the bare DNA showed

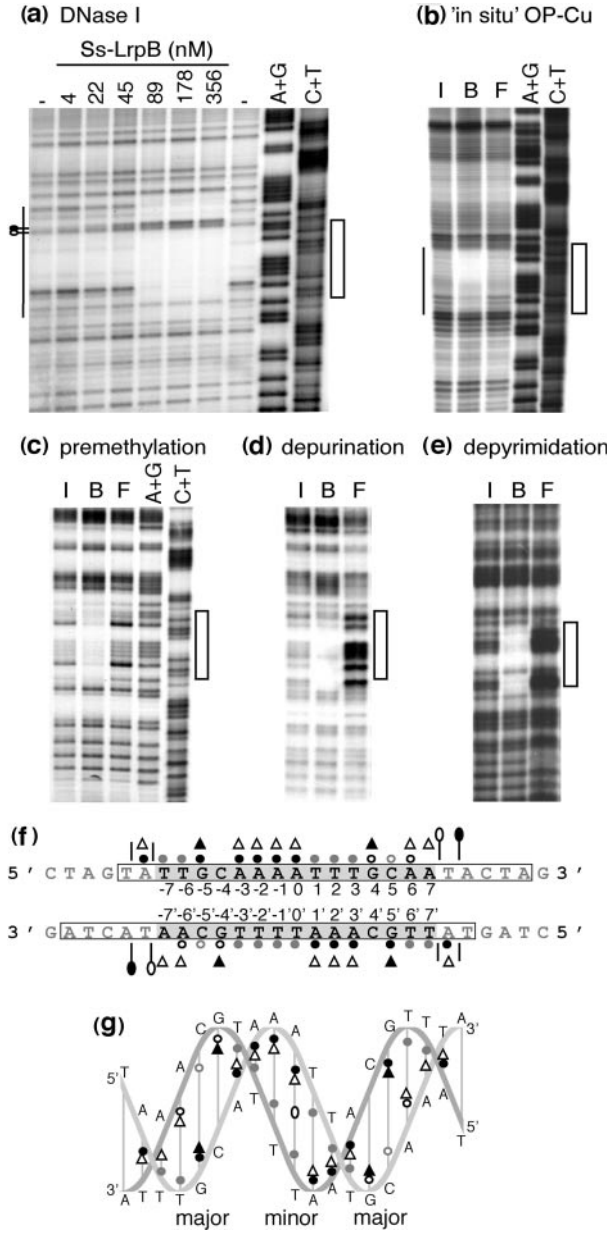

Figure 1. (a)–(e) Autoradiographs of various footprinting and binding interference experiments with the top strand labeled. The position of the consensus box is indicated on the right of each autoradiograph. The Ss-LrpB concentrations used in the DNase I footprinting experiment are indicated on top of the DNase I footprint (in nM, monomer equivalents). DNase I hypersensitivity is represented by an open circle (weak effect) or filled circle (strong effect) on the left side of the DNase I autoradiograph. The bar on the left side of both footprints corresponds to the protected area. For the in-gel Cu-OP footprinting and all binding interference experiments, the lanes with input DNA (I), bound DNA (B) and free DNA (F) are indicated. **(f)** Linear sequence of the consensus box (bold) and flanking sequences (grey) with a summary of all observations made in the high-resolution contact probing analysis. Position numbering is also indicated. The DNase I protected regions are boxed, the Cu-OP protection is indicated in gray. Circles represent effects observed in the depurination (black) and depyrimidation (gray) binding interference experiments. Triangles represent effects observed in the premethylation binding interference experiments. Strong and weak effects are depicted by filled and open symbols, respectively. DNase I hypersensitivity is indicated by a vertical bar with a circle, DNase I reactivity by a vertical bar. **(g)** Helical representation of the observed effects, with indication of the major and minor groove segments that are contacted by Ss-LrpB. The same symbols have been used as in (f).

little cleavage activity in the consensus box whereas the flanking sequences were cleaved more efficiently. This heterogeneity in the sensitivity to both the enzymatic and the chemical nuclease reflects local conformational variations of the DNA and suggests a narrowing of the minor groove in the A+T rich (73.3%) consensus-binding site (20,21).

Critical base-specific contacts were identified by premodification binding interference experiments. These techniques are based on the creation of a pool of DNA molecules with on average one base-specific modification per molecule. Low and high affinity molecules are subsequently separated in an EMSA designed to result in ~50% binding. Free and bound DNA are recovered from the gel, cleaved at the site of modification and analyzed by gel electrophoresis in denaturing conditions to distinguish positions that are crucial for binding from sites that are irrelevant (see Materials and Methods). Experimental results are shown for the top strand only (Figure 1c–e); a summary of all the effects is given in Figure 1f and in a helical presentation in Figure 1g. Methylation of the N^7 atom (major groove) of any of four guanine residues (at the adjacent positions 4, 5′ and the symmetrical counterparts −4′, −5) strongly inhibited complex formation (Figure 1c and f) indicating that Ss-LrpB makes strong major groove contacts with these parts of the binding site (Figure 1g). The methylation of the N^3 atom (minor groove) of 12 adenine residues (positions 0, 1′, 2′, 3′, 6, 7 and the symmetrical counterparts) resulted in a weaker but significant interference (Figure 1c and f). A similar effect was observed at two symmetrically related adenine residues juxtaposing the 15-bp target. These results suggest local minor groove contacts, mainly in the central part of the binding site and near its extremities (Figure 1g), though it can not be excluded that some effects might be indirect and result from subtle local alterations of the helix induced by the methyl group.

Base removal binding interference effects (missing contact) are usually considered to represent direct effects, which imply that they reflect interactions established between these bases and the interacting protein molecule (22). Alternatively, indirect effects might occur owing to structural alterations or changes in the DNA conformability generated by the absence of a base (8,23). Based on the observed effects (Figure 1d–g) it can be concluded that the vast majority of the purine and pyrimidine residues of the consensus box contribute to Ss-LrpB binding. On the top strand the strongest effects were observed upon removal of G_{-5} and any adenine of the stretch A_{-3}–A_0 and of T_{-7}, T_{-6}, T_1, T_2 and T_3. Removal of the adenine 5′ to the consensus box also interfered with Ss-LrpB binding. Weaker effects were observed upon removal of G_4, C_5 and A_6. In contrast, the removal of C_{-4} and A_7 hardly affected complex formation. Therefore the inhibitory effect of premethylation of residue A_7 is likely indirect since it occurs in the minor groove, on the backside of the DNA molecule. The effects observed upon removal of residues of the lower strand provided essentially the symmetrical image of the results on the top strand (Figure 1f and g). It appears that Ss-LrpB establishes the vast majority of its tightest major groove contacts with one single strand segment in each half-site (Figure 1f and g). Generally, the results of the footprinting and missing contact probing assays performed with the consensus box are in good agreement with the results obtained previously with the array of three degenerate binding sites in the *Ss-lrpB* control region (8).

Finally, random deoxyuridine substitution binding interference experiments (24) were performed. Substitution of thymine by uracil, which corresponds to the removal of the C^5-methyl group in the major groove, did not impair complex formation (data not shown). This indicates that the methyl groups of thymine residues potentially contacted through the major groove (positions 3, 6′, 7′ and the symmetrical counterparts at −3′, −6 and −7) do not significantly contribute to Ss-LrpB binding. This is unlike what has been observed in many other protein–DNA complexes and is even more surprising in view of the elevated A+T content of the consensus box.

Binding affinity of Ss-LrpB to the consensus box depends on the length of the flanking sequences

The sequence preference of Ss-LrpB at all positions of one half-site of the binding site was studied with derivatives of a 45-bp duplex DNA (see Materials and Methods). The annealing of complementary oligonucleotides allowed to study in identical conditions all the substitution mutants, the abasic molecules and molecules carrying a non-canonical base. The apparent equilibrium dissociation constant (K_D) of the interaction was determined by applying the EMSA technique (Figure 2). This was done by quantitating the free DNA population by densitometry and plotting the

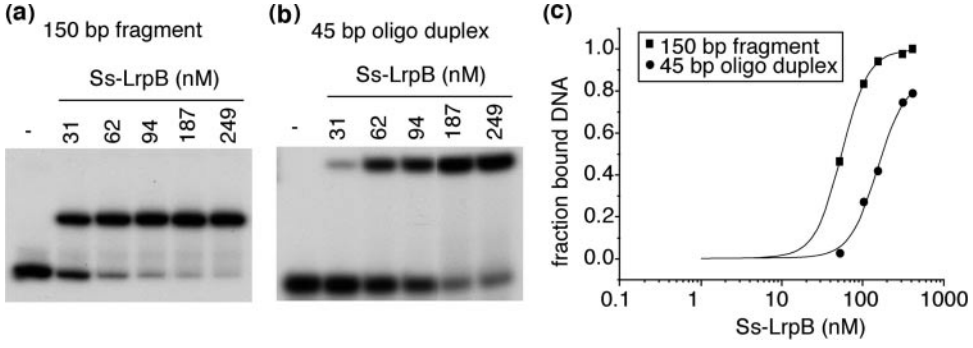

Figure 2. **(a)** Autoradiograph of an EMSA with a 150-bp fragment containing the consensus box. Ss-LrpB concentrations are indicated on top of the autoradiograph (in nM, monomer equivalents). **(b)** Autoradiograph of an EMSA with a 45-bp oligonucleotide duplex containing the consensus box. **(c)** Binding profiles of the two EMSAs shown in (a) and (b).

fraction of bound DNA versus the concentration of Ss-LrpB. As demonstrated previously, the EMSA can provide estimates of the macroscopic binding constants of protein–DNA interactions (25). The construction of a binding profile by fitting the Hill equation allowed us to determine the apparent K_D as the protein concentration required to shift a DNA fraction of 0.5. This also allowed us to determine the Hill coefficient, which is a measure of binding cooperativity. The affinity of binding to the 45-bp consensus duplex (K_D of 91 nM) was ~3-fold lower than binding to a 150-bp DNA fragment bearing the same consensus-binding site (K_D of 33 nM) (Figure 2 and Table 1). Similar differences in binding affinity depending on the length of the flanking sequence have been observed in other studies (26). Footprinting experiments (DNase I and in-gel Cu-OP) clearly indicate that the 150-bp fragment does not assemble >1 Ss-LrpB dimer. Therefore, the most plausible explanation of the size effect is that the flanking sequences provide 105 non-specific sites that contribute the equivalent of two specific sites. Furthermore, accelerated targeting by 1D diffusion along the DNA molecule and the possibility of having different dissociation rates from internal sites and the ends of the fragments may add to the observed differences in binding affinity to the 105- and 45-bp fragments (27).

The Hill coefficient was similar for binding to the 150-bp fragment and the 45-bp duplex, on average 2.4. This indicates positive cooperativity. Since Ss-LrpB binds to a single palindromic site, this cooperativity might indicate the existence of an equilibrium between monomeric and dimeric forms of Ss-LrpB at low protein concentration. Indeed, although gel filtration and crosslinking experiments indicate that the predominant oligomeric form of Ss-LrpB in solution is a dimer (9), this does not exclude that dissociation might occur in the nM range.

Table 1. Apparent K_D-values of Ss-LrpB binding to the consensus box and derivatives thereoff as determined by EMSA; WT consensus

	K_D (nM)	K_{Drel}[a]
45-bp duplex	91	1
150-bp fragment	33	0.36

[a]K_{Drel} is the relative K_D value compared with the K_D of binding to the WT oligonuclotide duplex, which is 91 nM.

Table 2. Single base pair substitution mutants

Position	WT bp	Substitution	K_D (nM)	K_{Drel}[a]
0	A•T	C•G	348	3.82
		G•C	260	2.86
		T•A	109	1.20
		ab•T	477	5.24
1	T•A	A•T	168	1.85
		C•G	218	2.40
		G•C	217	2.39
2	T•A	A•T	271	2.98
		C•G	284	3.12
		G•C	360	3.96
3	T•A	A•T	801	8.80
		C•G	664	7.30
		G•C	488	5.36
4	G•C	A•T	386	4.24
		C•G	467	5.13
		T•A	356	3.91
5	C•G	A•T	>16 000[b]	>176[b]
		G•C	7830	86.0
		T•A	811	8.91
		ab•G	399	4.38
		C•*ab*	1447	15.9
		ab•*ab*	>11 000[b]	>121[b]
		C•I	115	1.26
		A•U	645	7.09
		C^{Me}•G	184	2.02
		C•2AP	>11 000[b]	>121[b]
6	A•T	C•G	465	5.11
		G•C	450	4.95
		T•A	281	3.09
7	A•T	C•G	633	6.96
		G•C	172	1.89
		T•A	369	4.05
		ab•T	214	2.35
		A•*ab*	278	3.05
		ab•*ab*	416	4.57
		I•C	155	1.70

[a]K_{Drel} is the relative K_D value compared with the K_D of binding to the WT oligonuclotide duplex, which is 91 nM.
[b]Represents the minimal value.

Analysis of the binding specificity of Ss-LrpB

Complex formation was studied with a complete set of 24 single base pair substitution mutants of one half of the symmetrical consensus box (saturation mutagenesis of positions 0–7) as described above. This allowed a detailed analysis of the sequence-specific contribution of each base pair to the interaction. An example of such analysis is presented for position 5 (Figures 3a–e). Average apparent K_D values for all mutants are summarized in Table 2. In the presence of excess non-specific competitor DNA, binding to these annealed oligonucleotides rarely resulted in a complete depletion of the free DNA population, even at the highest protein concentrations used and showed a more intense smearing, indicative of the formation of less stable complexes. As a consequence, the maximal fraction of bound DNA (corresponding to v_{max}) in the Hill fittings is slightly <1 in most cases and significantly <1 in a few instances (very low affinity binders).

None of the 24 single base pair substitution mutations resulted in a better binding of Ss-LrpB. Therefore, the consensus box appears to be optimized. In contrast, all possible changes except the A·T→T·A substitution at position 0 resulted in a nearly 2- to >100-fold reduction of the binding affinity (Table 2). A survey of the relative binding affinities (Figure 4) indicates that all base pairs of the consensus contribute in a sequence-specific manner and to a variable degree to complex formation, but that position 5 is crucial.

Major groove contacts: hydrogen bonding at C_5·$G_{5'}$ is crucial for Ss-LrpB binding

The saturation mutagenesis study (Table 2, Figure 4) clearly indicates that the C·G at position 5 is the most discriminating base pair of the Ss-LrpB binding site. Representative autoradiograms of binding to mutants of position 5 are shown in Figure 3a–d. Remarkably, substituting A·T for C·G almost completely destroyed binding. Even at the highest Ss-LrpB concentration used (16 220 nM), only

some unstable binding was observed. The C·G→G·C mutant showed detectable specific binding, although with a strongly reduced affinity (86-fold increase in K_D). The smallest effect was observed with the T·A substitution (8.9-fold increase in K_D).

Hydrogen bonds are by far the most important sequence-specific contributors to complex formation. Given the more extensive reorganization of potential hydrogen bonding groups when substituting T·A as compared with A·T for C·G (Figure 3a–d) one might have expected the latter to be

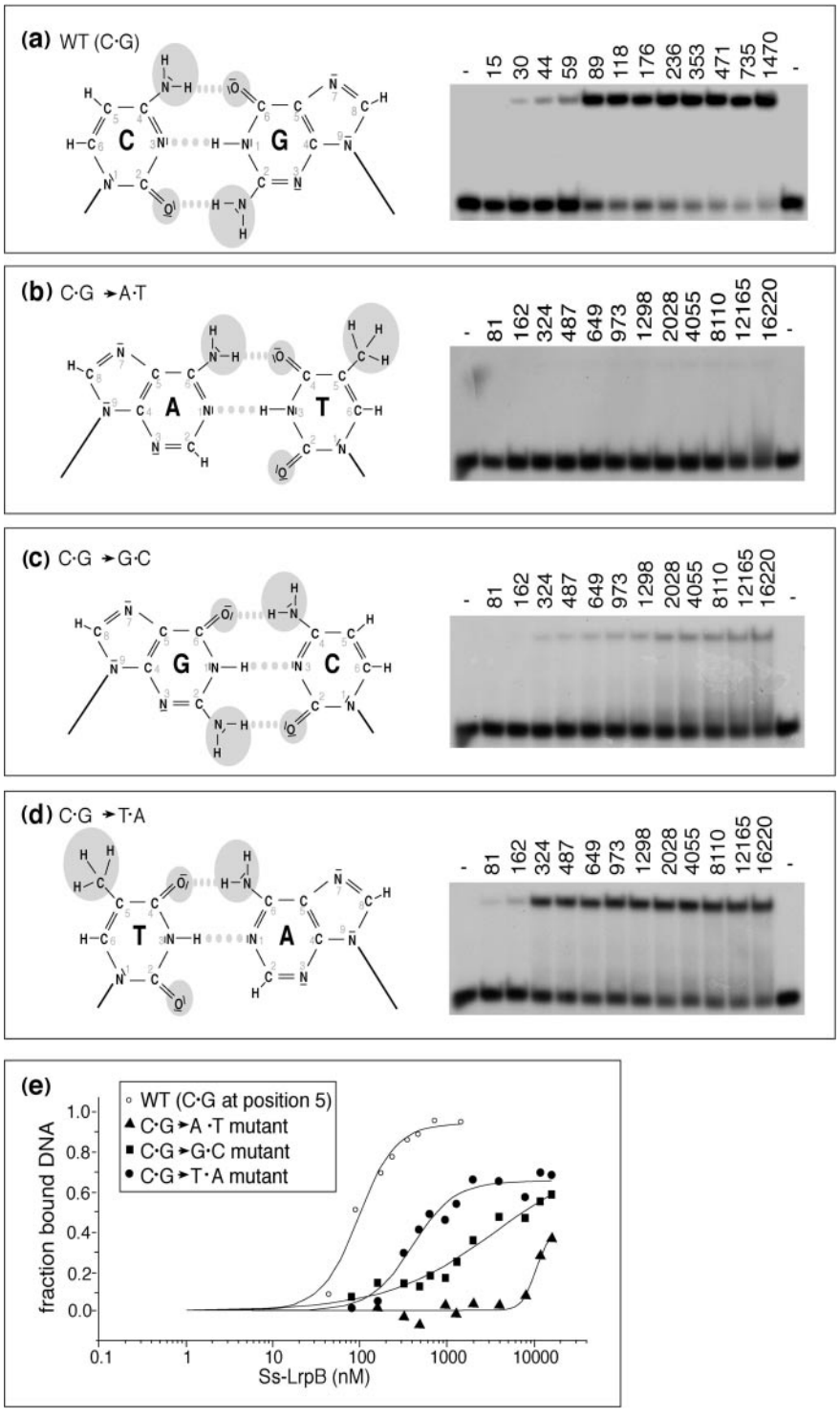

Figure 3. Examples of some EMSA analyses that have been used for the determination of the apparent K_D's. All the variants shown here are mutated at position 5. Ss-LrpB concentrations are indicated in nM (monomer equivalents). If appropriate, the molecular structure of the base pair is also shown. **(a)** Autoradiograph of an EMSA with the WT fragment (45-bp oligonucleotide duplex). **(b)–(d)** Autoradiographs of EMSAs with saturation mutagenesis variants. **(e)** Binding profiles of the EMSAs shown in (a)–(d).

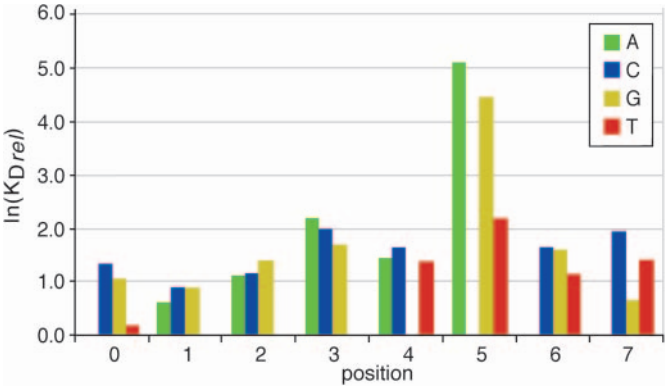

Figure 4. Histogram of $\ln(K_{Drel})$ values based on the results of the saturation mutagenesis analysis (Table 1). The value of the A_5 mutant corresponds to a minimal value.

the less detrimental change. The opposite was observed: this appears to be due to a position-dependent inhibitory effect of the hydrophobic methyl group of thymine, as demonstrated with the use of uracil and 5-methyl cytosine substitutions (Figure 5a and b; Table 2). Uracil is equivalent to thymine but lacks the C^5 methyl group on the major groove side of the base. Binding to the A·U mutant [7.1-fold increase in K_D as compared with wild type (WT)] was much better than to the A·T mutant (>170-fold increase in K_D) and comparable with the T·A mutant (8.9-fold increase). The substitution of C_5 by 5-methyl cytosine (C^{Me}), keeping the complementary $G_{5'}$ residue intact, resulted only in a 2-fold increase in the K_D. Combined, these results clearly indicate the strong negative position-dependent effect of a methyl group at position 5′ of the bottom strand.

The importance of the C·G base pair is also reflected in the results of the high resolution contact probing analysis. Both depyrimidation of C_5 and depurination of $G_{5'}$ strongly interfered with complex formation (see above). The respective contribution of each base to complex formation was further quantified with target molecules abasic for either one of these complementary bases or both (Figure 5e–g; Table 2). The removal of $G_{5'}$ resulted in a 3- to 4-fold higher K_D than the removal of its partner C_5. Binding to the double abasic consensus site was hardly detectable. This result emphasizes the overwhelming importance of $G_{5'}$ for the strength and specificity of complex formation.

The strong inhibitory effects of depurination and of premethylation of the N^7 atom of $G_{5'}$ suggest direct hydrogen bonding of Ss-LrpB to the N^7 and/or O^6 hydrogen bond acceptors on the major groove side of the guanine ring. To evaluate the importance of hydrogen bonding with the C^6 carbonyl of $G_{5'}$ we compared complex formation with targets carrying either 2-aminopurine (2-AP) or inosine instead of guanine (Figure 5c and d; Table 2). 2-AP lacks the carbonyl group on the major groove side of the base, whereas inosine lacks the exocyclic amino group on the minor groove side. 2-AP was highly detrimental for complex formation (>120-fold increase in K_D as compared with the WT) whereas, as expected, inosine had nearly no effect (1.3-fold increase in K_D). Combined, our results pin-point the capital importance of hydrogen bonding with the C^6 carbonyl of guanine at position 5 (and its symmetrical counterpart $G_{-5'}$) in the site selectivity of Ss-LrpB.

Arginine:guanine is the most common specific contact found in crystal structures of protein–DNA complexes, followed by arginine:cytosine (28). Furthermore, in 87% of the arginine:guanine pairs, hydrogen bonds are formed with the amide group of arginine and the N^7 and O^6 acceptor atoms of the guanine. Therefore, the C·G base pair at position 5 and its symmetrical counterpart might very well be contacted by the side chain of one or more arginine residues of the recognition helix of Ss-LrpB. Ss-LrpB bears three arginine residues in its recognition helix, at positions 42, 44 and 47. Of these, R_{44} is highly conserved among Lrp-like regulators. Alanine substitution of these residues might provide further molecular details on the interaction of Ss-LrpB with position 5 of the binding site. It is worth noting that HTH motives generally do not interact in a one-to-one mode with their DNA target but rather establish complex patterns of interactions in which one base is contacted by different amino acids and one amino acid contacts several bases.

Major groove contacts: positions 3, 4, 6 and 7

At position 3, contacted through the major groove, the T·A→G·C mutant showed the smallest reduction in binding affinity (5.4-fold increase in K_D), followed by the C·G (7.3-fold) and A·T (8.8-fold increase) mutants. This particular order in the base pair preference might in part be explained by the fact that the hydrogen bond acceptor groups of T and G are only slightly shifted in a G·C base pair as compared with a T·A pair, whereas they are completely rearranged in an A·T pair (see Figure 3a–d). Here, unlike what we observed at position 5, the hierarchy of relative binding affinities correlates with the degree of reorganization of major groove constituents in the different substitution mutants. Otherwise, steric hindrance on neighboring contacts and differences in the local groove geometry might also contribute to the observed differences in binding affinity.

Replacing the G·C base pair at position 4 of the consensus box by any of the three other possible combinations resulted in a similar modest effect (4- to 5-fold increases in K_D; Table 2). This observation is compatible with the higher variability observed at position 4 in the three Ss-LrpB binding sites in the control region of its own gene.

Together with position 4, position 6 appears to contribute the least to the specific binding of all major groove-contacted positions. The A·T→T·A mutant showed the highest binding affinity (3.1-fold increase in K_D). Both the G·C and C·G substitutions resulted in a similar, ∼5-fold increase in K_D. The A·T base pair at position 7 is also contacted through the major groove, as indicated by the similar relative binding affinities of G·C and I·C substitution mutants (K_D of 172 and 155 nM, respectively; Table 2). Nevertheless, the nature and the specificity of these interactions appear to be completely different from position 6 as indicated by the hierarchy in the binding specificities. At position 7, the G·C transition mutant had the smallest effect (1.9-fold increase in K_D), followed by the T·A mutant (∼4-fold increase) and the C·G mutant (∼7-fold increase). The surprisingly small

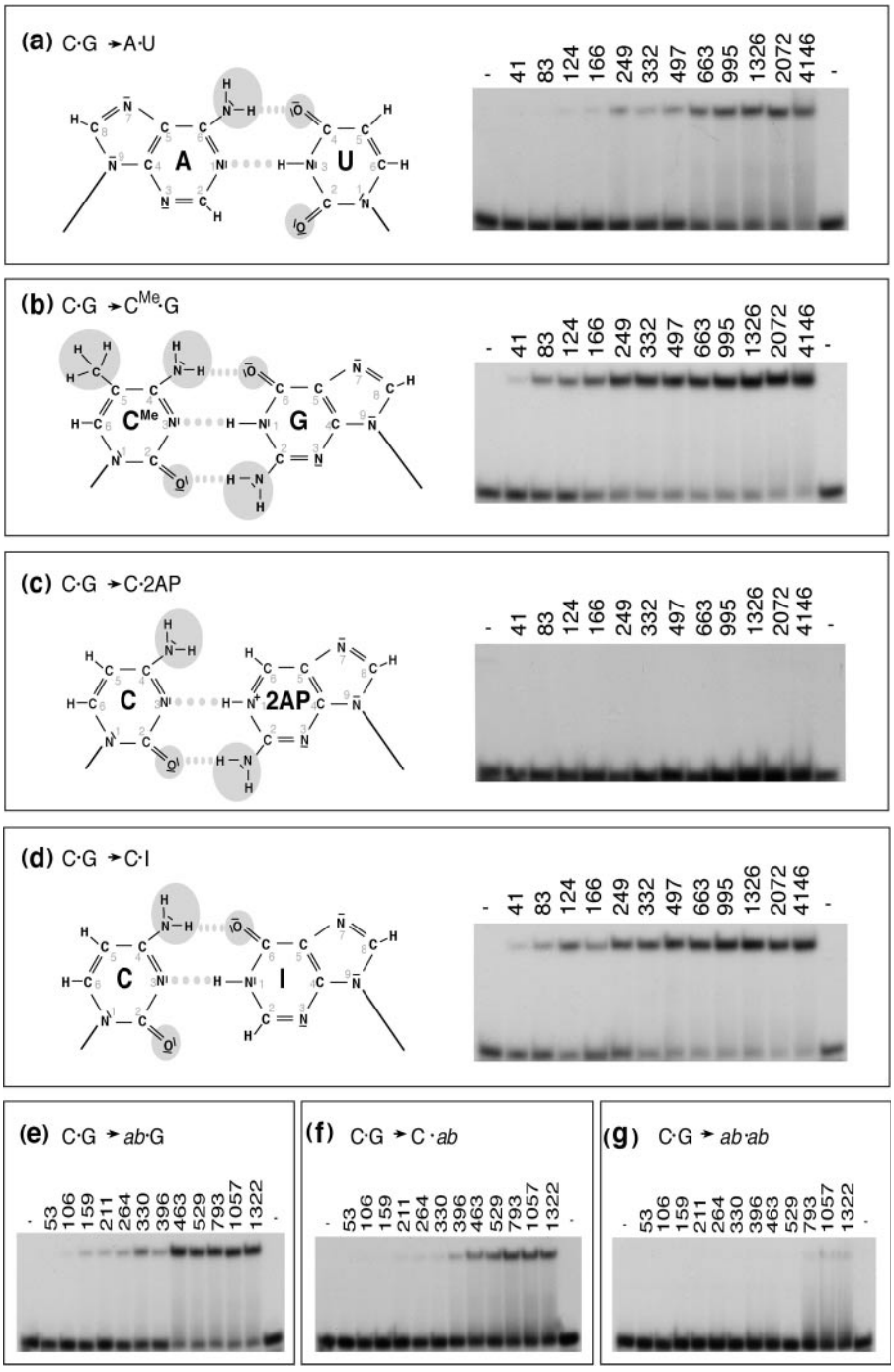

Figure 5. Examples of some EMSA analyses that have been used for the determination of the apparent K_D's. **(a)–(d)** Autoradiographs of EMSAs with an A·U bp, a C^{Me}·G bp, a C·2AP base pair or a C·I bp, respectively, at position 5 of the consensus box. **(e)–(g)** Autoradiographs of EMSAs with abasic molecules.

effect of the G·C substitution suggests that alternative contacts may be established upon the profound reorganization of the bottom of the major groove accompanying this substitution.

The mild effect (2.4-fold) of a double substitution mutant with uracil instead of $T_{6'}$ and $T_{7'}$ indicates that the methyl group of these residues does not contribute much to the binding energy and sequence specificity of the interaction (Table 3). This result is in full agreement with the random deoxyuridine substitution binding interference experiments (see above). The comparable mild (2- to 3-fold) reduction in relative binding affinity to the single and double (4.6-fold) abasic mutants (Table 2) indicates that the A and T residues of position 7 contribute significantly less to the binding energy and specificity than the G residue at position 5′.

Minor groove contacts: positions 0, 1 and 2

The positions that are supposed to be contacted through the minor groove generally had a higher tolerance to base pair

substitution than the half-site positions contacted through the major groove (Table 2, Figure 4). At these positions, the highest tolerance we observed is to mutation in the opposite weak base pair (T·A to A·T or the inverse). This reflects the fact that a protein can hardly distinguish a T·A base pair from an A·T base pair on the minor groove side since their hydrogen bond acceptor groups (C^2 carbonyl from T and N^3 of A) will roughly switch positions. The effect of the A·T to T·A transversion was smallest at the central position 0 (1.2-fold increase in K_D). In fact, this base pair substitution leaves the 15-bp consensus box unchanged; it simply corresponds to reading the opposite strand. Therefore, the observed 1.2-fold difference in binding affinities reflects the experimental error and is within the error range observed in the EMSAs (see Materials and Methods). Both bases of this complementary pair contribute significantly to complex formation as indicated by the missing contact probing assays (see above). This was confirmed and better quantified for A_0; binding to the abasic 45-bp duplex occured with an ~5.2-fold reduced affinity (Table 2).

None of the three Ss-LrpB binding sites in the own control region has a strong base pair in the central part of the box. The apparent prohibition of strong base pair might be related to the exocyclic C^2-amino group of guanine, which plays a dual role in DNA structure and recognition. The introduction of an NH_2 group in the minor groove results in groove widening and consequently in a decrease of the electronegative potential as compared with weak base pairs. It also constitutes a steric hindrance for the compression of the minor groove upon bending of the operator DNA towards the interacting Ss-LrpB molecule (8,9). The need for bending of an operator site is generally reflected in the fact that there is a preference for A+T rich sequences at the midpoint of the target, where the minor groove has to be narrow. To evaluate the influence of the C^2-amino group of guanine upon substitution of strong base pairs for A·T and T·A in the minor groove we measured complex formation with a pair of double base pair substitutions carrying G·C and C·G (positions −1 and 1, respectively) and the equivalent construct carrying inosine instead of guanine (Table 3). Inosine lacks the C^2-amino group of guanine on the minor groove side of the base and is also equivalent to C^6-deaminated adenine. Therefore an I·C pair resembles an A·T pair in the minor groove and a G·C pair in the major groove. The double guanine-bearing mutant exhibited an ~7-fold reduction in the relative binding affinity compared with the WT; a similar effect was observed in a double G·C for A·T substitution at positions 0 and 1 (Table 3). In contrast, the inosine bearing double mutant showed a 1.6-fold increase in K_D, only. Therefore, inosine interferes significantly less with complex formation than guanine. This result emphasizes the importance of minor groove geometry in the central part of the Ss-LrpB binding site.

Binding to insertion and deletion mutants

To examine the importance of the alignment of the two half-sites of the consensus box contacted through the major groove we studied the binding to insertion and deletion mutants (Table 4). These mutants had an insertion or deletion of either 1 or 2 bp in the A+T rich central segment contacted through the minor groove. The insertion of one extra T·A base pair (ins1: in the stretch T_1–T_3) still allowed the formation of the specific complex, though with a 3.3-fold increase of the K_D as compared with the WT. In contrast, deleting a single A·T base pair (del1: in the stretch of A_{-3}–A_0) completely abolished Ss-LrpB binding (K_D > 11 000 nM; Table 4). A similar result was observed with a double base pair deletion mutant (A·T and T·A). A double base pair insertion mutant (consecutive A·T and T·A base pairs in the center of the box) resulted in a specific binding with a 11.8-fold increased K_D. These results indicate a limited conformational flexibility of the Ss-LrpB dimer. Increasing the separation between the two half-sites contacted through the major groove, thereby disturbing the helical alignment, is tolerated to a certain extent. In contrast, reducing their separation (thereby inducing a similar rotation of ~34° per base pair but in the opposite direction) is highly detrimental. These results suggest that steric hindrance between the two subunits of the Ss-LrpB dimer might exclude the simultaneous binding of their HTH motives to the unproperly aligned major groove segments of the deletion mutants.

Table 3. Double base pair substitution mutants

Position	Substitution	Position	Substitution	K_D (nM)	K_{Drel}[a]
−1	A•T → G•C	1	T•A → C•G	638	7.01
−1	A•T → I•C	1	T•A → C•I	145	1.59
0	A•T → G•C	−1	A•T → G•C	580	6.37
6	A•T → A•U	7	A•T → A•U	221	2.43

[a]K_{Drel} is the relative K_D value compared with the K_D of binding to the WT oligonuclotide duplex, which is 91 nM.

Table 4. Insertion and deletion mutants

Mutant	K_D (nM)	K_{Drel}[a]
ins1	299	3.29
ins2	1073	11.79
del1	>11 000[b]	>121[b]
del2	>11 000[b]	>121[b]

[a]K_{Drel} is the relative K_D value compared with the K_D of binding to the WT oligonuclotide duplex, which is 91 nM.
[b]Represents the minimal value.

Energy normalized sequence logo-modeling

DNA-binding sequence motifs can be graphically represented by sequence logos, based on the information theory (13). The height of the stack of letters corresponds to the sequence conservation, expressed in bits of information. The relative heights of the bases correspond to their relative frequencies. Therefore, a sequence logo provides more information than a consensus sequence. Here, we generated two kinds of logos with a web interface called enoLOGOS: (i) based on sequence comparisons, (ii) based on binding energies (Figure 6a and b; 14). A sequence logo was constructed based on the three binding sites for Ss-LrpB in the control region of its own gene (Figure 6a; 8). An alignment of both forward and reverse sequences was used since Ss-LrpB binds as a dimer. Two corrections were applied in the creation of this logo: (i) A small-sample correction, since the information content tends to be overestimated in the case of a small

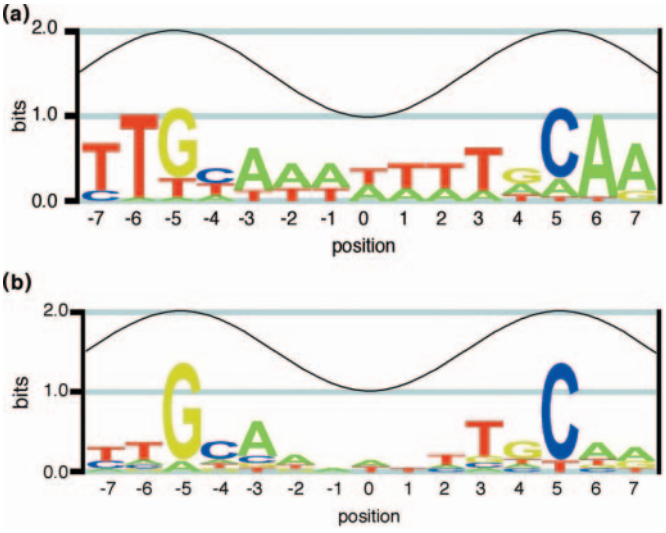

Figure 6. Sequence logo for Ss-LrpB binding based on the three natural binding sites. The height of the letters is expressed in information content (bit). The cosine wave represents the major and minor groove as contacted by Ss-LrpB. **(b)** Energy normalized sequence logo (enologo) for Ss-LrpB binding based on the saturation mutagenesis analysis.

dataset. (ii) A correction for background base frequencies in the case of genomes with a biased GC content such as *S.solfataricus* (GC content of 37%). The Ss-LrpB sequence logo confirms our previously deduced consensus sequence. Conservation is higher in the major groove contacted half-sites than in the A+T rich center. However, since only three binding sites have been used to create the logo, its significance is rather low.

In contrast, the saturation mutagenesis dataset used to create an energy normalized sequence logo is much larger and therefore more significant (Figure 6b). Full symmetry was assumed to construct the logo for the complete 15-bp binding site. The correction for the biased GC content was not applied since it is irrelevant in this case. Compared with the logo based on the natural targets, this 'enologo' has a higher resolution. The relative importance of a C at position 5 (and G at position −5) is highly emphasized. A cosine wave represents the twist of the DNA helix. Positions recognized in the minor groove have a maximal information content of 1 bit as opposed to 2 bits in the major groove (29). This is explained by the fact that in the minor groove a protein cannot discriminate weak base pairs from each other or strong base pairs from each other. Binding of the $C_5 \cdot G_{5'}$ base pair in the major groove is confirmed since the information content exceeds 1. On the contrary, based on our binding affinity data, the sequence in the minor groove is relatively less important. This is certainly the case for positions −1, 0 and 1, where little sequence preference is indicated in the logo. The total information content of this enologo is 7.2 bits as compared with 10.4 for the sequence logo based on the natural binding sites.

The logo constructed with the binding energy data from the saturation mutagenesis analysis can be used to search for new high affinity binding sites in the *S.solfataricus* P2 genome. Assuming additivity, which is usually a good approximation in order to find new sites (30), the binding affinity of each sequence can be predicted. Nevertheless, setting the threshold such that relevant high affinity sites are retrieved and the amount of false positives is minimalized, is a delicate and not so straightforward process. The algorithm will be designed to allow an extra 1 or 2 bp in the center of the box. The joint occurence of high and low affinity sites should also be considered. When correctly aligned the latter will be bound in a cooperative manner and consequently acquire a physiological role, as already demonstrated by the concentration-dependent formation of structurally very different complexes of Ss-LrpB with the control region of its own gene (9).

Although our binding assays indicate that the consensus box bears the optimal base at each position, this does not necessarily imply that the consensus box is the best possible binding site. Some positions might be functionally interdependent and therefore give rise to context-dependent effects. The tightest binding site could have been found by SELEX, but on the other hand this technique does not provide the energy landscape that was determined here. It is also worth noting that the consensus box does not occur in the *S.solfataricus* genome. This is frequently observed with regulatory sites. Tight binding and long half-lives of such complexes might be incompatible with the flexibility required to adapt to rapidly changing environmental conditions and microbial generation times.

ACKNOWLEDGEMENTS

This work was supported by a grant from the Research-Foundation-Flanders (FWO-Vlaanderen, G.0015.04), the Research Council of the Brussels University (OZR1184-VUB) and the Vlaamse Gemeenschapscommissie. E.P. is holder of a PhD grant of the Institute for the promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen). We also thank the reviewers for their helpful remarks and constructive comments. Funding to pay the Open Access publication charges for this article was provided by grant G.0015.04 of the Research Foundation-Flanders (FWO-Vlaanderen).

Conflict of interest statement. None declared.

REFERENCES

1. Ouhammouch,M. (2004) Transcriptional regulation in Archaea. *Curr. Opin. Genet. Dev.*, **14**, 133–138.
2. Brinkman,A.B., Ettema,T.J.G., de Vos,W.M. and van der Oost,J. (2003) The Lrp family of transcriptional regulators. *Mol. Microbiol.*, **48**, 287–294.
3. Brinkman,A.B., Bell,S.D., Lebbink,R.J., de Vos,W.M. and van der Oost,J. (2002) The *Sulfolobus solfataricus* Lrp-like protein LysM regulates lysine biosynthesis in response to lysine availability. *J. Biol. Chem.*, **277**, 29537–29549.
4. Leonard,P.M., Smits,S.H.J., Sedelnikova,S.E., Brinkman,A.B., de Vos,W.M., van der Oost,J., Rice,D.W. and Rafferty,J.B. (2001) Crystal structure of the Lrp-like transcriptional regulator from the archaeon *Pyrococcus furiosus*. *EMBO J.*, **20**, 990–997.
5. Koike,H., Ishijima,S.A., Clowney,L. and Suzuki,M. (2004) The archaeal feast/famine regulatory protein: potential roles of its assembly forms for regulating transcription. *Proc. Natl Acad. Sci. USA*, **101**, 2840–2845.
6. Thaw,P., Sedelnikova,S.E., Muranova,T., Wiese,S., Ayora,S., Alonso,J.C., Brinkman,A.B., Akerboom,J., van der Oost,J. and Rafferty,J.B. (2006) Structural insight into gene transcriptional regulation and effector binding by the Lrp/AsnC family. *Nucleic Acids Res.*, **34**, 1439–1449.

7. Ettema,T.J.G., Brinkman,A.B., Tani,T.H., Rafferty,J.B. and van der Oost,J. (2002) A novel ligand-binding domain involved in regulation of amino acid metabolism in prokaryotes. *J. Biol. Chem.*, **277**, 37464–37468.
8. Peeters,E., Thia-Toong,T.L., Gigot,D., Maes,D. and Charlier,D. (2004) Ss-LrpB, a novel Lrp-like regulator of *Sulfolobus solfataricus* P2, binds cooperatively to three conserved targets in its own control region. *Mol. Microbiol.*, **54**, 321–336.
9. Peeters,E., Willaert,R., Maes,D. and Charlier,D. (2006) Ss-LrpB from *Sulfolobus solfataricus* condenses about 100 base pairs of its own operator DNA into globular nucleoprotein complexes. *J. Biol. Chem.*, **281**, 11721–11728.
10. Tuerk,C. and Gold,L. (1990) Systematic evolution of ligands by exponentional enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science*, **249**, 505–510.
11. Cui,Y., Wang,Q., Stormo,G.D. and Calvo,J.M. (1995) A consensus sequence for binding of Lrp to DNA. *J. Bacteriol.*, **177**, 4872–4880.
12. Ouhammouch,M. and Geiduschek,E.P. (2001) A thermostable platform for transcriptional regulation: the DNA-binding properties of two Lrp homologs from the hyperthermophilic archaeon *Methanococcus jannaschii*. *EMBO J.*, **20**, 146–156.
13. Schneider,T.D. and Stephens,R.M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, **18**, 6097–6100.
14. Workman,C.T., Yin,Y., Corcoran,D.L., Ideker,T., Stormo,G.D. and Benos,P.V. (2005) enoLOGOS: a versitale web tool for energy normalized sequence logos. *Nucleic Acids Res.*, **33**, 389–392.
15. Enoru-Eta,J., Gigot,D., Thia-Toong,T.-L., Glansdorff,N. and Charlier,D. (2000) Purification and characterization of Sa-Lrp, a DNA-binding protein from the extreme thermoacidophilic archaeon *Sulfolobus acidocaldarius* homologous to the bacterial global transcriptional regulator Lrp. *J. Bacteriol.*, **182**, 3661–3672.
16. Wang,H., Glansdorff,N. and Charlier,D. (1998) The arginine repressor of *Escherichia coli* K-12 makes direct contacts to minor and major groove determinants of the operators. *J. Mol. Biol.*, **277**, 805–824.
17. Maxam,A.M. and Gilbert,W. (1980) Sequencing end-labeled DNA with base-specific chemical cleavages. *Methods Enzymol.*, **65**, 499–560.
18. Stormo,G.D. and Fields,D.S. (1998) Specificity, free energy and information content in protein–DNA interactions. *Trends Biochem. Sci.*, **23**, 109–113.
19. Kuwabara,M. and Sigman,D. (1987) Footprinting DNA–protein complexes *in situ* following gel retardation assay using 1, 10-phenantroline copper ion: *Escherichia coli* RNA polymerase-*lac* promoter complexes. *Biochem.*, **26**, 7234–7238.
20. Spassky,A. and Sigman,D. (1985) Nuclease activity of 1,10-phenantroline-copper ion. Conformational analysis and footprinting of the *lac* operon. *Biochem.*, **24**, 8050–8056.
21. Sigman,D.S., Spassky,A., Rimsky,S. and Buc,H. (1985) Conformational analysis of *lac* promoters using the nuclease activity of 1,10-phenantroline-copper ion. *Biopolymers*, **24**, 183–197.
22. Brunelle,A. and Schleif,R.F. (1987) Missing contact probing of DNA–protein interactions. *Proc. Natl Acad. Sci. USA*, **84**, 6673–6676.
23. Enoru-Eta,J., Gigot,D., Glansdorff,N. and Charlier,D. (2002) High resolution contact probing of the Lrp-like DNA-binding protein Ss-Lrp from the hyperthermoacidophilic crenarchaeote *Sulfolobus solfataricus* P2. *Mol. Microbiol.*, **45**, 1541–1555.
24. Pu,W. and Struhl,K. (1992) Uracil interferenc, a rapid and general method for defining protein–DNA interactions involving the 5-methyl group of thymines: the GCN4–DNA complex. *Nucleic Acids Res.*, **20**, 771–775.
25. Senear,D.F. and Brenowitz,M. (1991) Determination of binding constants for cooperative site-specific protein–DNA interactions using gel mobility-shift assay. *J. Biol. Chem.*, **266**, 13661–13671.
26. Szwajkajzer,D., Dai,L., Fukayama,J.W., Abramczyk,B., Fairman,R. and Carey,J. (2001) Quantitative analysis of DNA-binding by the *Escherichia coli* arginine repressor. *J. Mol. Biol.*, **312**, 949–962.
27. Halford,S.E. and Marko,J.F. (2004) How do site-specific DNA-binding proteins find their targets? *Nucleic Acids Res.*, **32**, 3040–3052.
28. Lejeune,D., Delsaux,N., Charloteaux,B., Thomas,A. and Brasseur,R. (2005) Protein-nucleic acid recognition: statistical analysis of atomic interactions and influence of DNA structure. *Proteins: Struc. Funct. Bioinformatics*, **61**, 258–271.
29. Papp,P., Chattoraj,D.K. and Schneider,T.D. (1993) Information analysis of sequences that bind the replication initiator RepA. *J. Mol. Biol.*, **233**, 219–230.
30. Benos,P.V., Bulyk,M.L. and Stormo,G.D. (2002) Additivity in protein–DNA interactions: how good an approximation is it? *Nucleic Acids Res.*, **30**, 4442–4451.