Chapter 9. Weaver, 2/e. Mol Biol X107A.

Ch 9. DNA-protein interactions in prokaryotes.

A. Introduction 1
B. Methodology 2
C. Base-specific DNA-binding 2
D. DNA structure -- variability and flexibility 2
E. DNA-binding motifs 4
F. Further reading 5
G. Errata 6
H. Homework 6
I. Partial answers 9

Reading note. Skip Box 9.1. (X-ray crystallography. If you browse it, and get some idea of the method, fine. But you certainly are not responsible for it.)

A. Introduction

Our goal in Ch 9 is to look at some general aspects of DNA-protein interactions.

Weaver starts Ch 9 with some specific examples of how DNA-binding proteins work, then develops general ideas. I will take the opposite approach here, with the goal of complementing his approach. Different people may quite reasonably read the parts of Ch 9 in different order.

Highlights

Principles of DNA-binding proteins:

- Proteins bind to DNA and modulate the function of that DNA.
- Binding is specific for a DNA sequence. (Of course, the degree of specificity varies.)
- Proteins can recognize DS DNA sequences via the DNA grooves, especially the major groove.
- The structural variability and flexibility of DNA are important.
  - DNA bending and looping are examples.
  - Indirect readout of DNA sequence, through effect on backbone.

– Protein flexibility is also important.

- The helix-turn-helix (HTH) motif is an example of a recurrent DNA-binding motif; a specific key part of this is the recognition helix.

B. Methodology

We learn much about protein-DNA interactions from classical methods, many discussed in Ch 6 and 7. These include genetic analysis and biochemical experiments such as footprinting.

The major breakthrough of recent years has been the explosion in measurements of actual structures. This is done largely by X-ray crystallography, and also some by NMR. (X-ray crystallography is an old method, but technical advances have resulted in greatly increased throughput in recent years.) We will not discuss the methodology per se, but will accept the results. See Smutzer (2001) for more about X-ray crystallography.

C. Base-specific DNA-binding

The top level idea, which already emerged in Ch 6 & 7, is that proteins bind to DNA and modulate the function of that DNA. We know that the binding can be specific for a DNA sequence.

We now learn that proteins can recognize DS DNA sequences via the DNA grooves, especially the major groove; Fig 9.19. This idea was briefly noted in the Ch 2 handout (Sect G and hw #3).

> Kielkopf et al (1998) discuss a way to distinguish the DNA sequence by reading the minor groove.
>
> Weaver makes minor groove recognition seem easier than I think it is. He was aware of the work achieving such recognition, so his facts are fine. But minor groove recognition, especially distinguishing AT and TA, is subtle.

We also learn that many protein-DNA contacts are with the DNA backbone. These would appear to not be sequence-specific. However, they may be sequence-specific if the backbone shape depends on the base sequence. Weaver refers to such sequence-specific binding to the backbone as "indirect readout" (pp 246-7). Further, non-specific contacts contribute to the overall binding energy, even when they do not contribute to specificity.

> The Figure near the top of my X107A web page is a simple ribbon diagram showing the phage λ Cro protein bound to DNA.

D. DNA structure -- variability and flexibility

Distinguish…

- The “natural” structural variability of DNA, based on sequence.

- The flexibility of DNA, or its ability to change structure in response to some event, such as protein-binding.

These structural variations of DNA (and of proteins) are relevant, in several ways:

- Indirect readout, discussed in Sect C, shows how the natural variability of DNA structure can be exploited.

- The DNA structure may change when a protein is bound, to produce a proper fit.

  We introduced this in the Ch 7 handout Sect E, with the CAP. Fig 9.20 now provides more detail. The bend is not gentle or a general distortion of the DNA. Instead, the bend involves specific major local distortion, as the protein interacts at a particular DNA region.

  Figs 9.13 & 14 show an example of a smaller, but significant, distortion of the DNA by a bound protein.

  How can this happen? The simple answer is that the energy of binding the distorted DNA is greater than the energy of distortion. That is, the low energy state is the protein bound to distorted DNA. Recall Ch 7 hw #6.

- In “extreme” cases, the DNA may loop, allowing a protein (or a protein complex) to bind to “distant” sites on the same DNA molecule.

  Fig 9.21 & 22 show this. (Also recall Fig 7.18.) Fig 9.22b also shows how such bending affects the DNA structure, in a “reasonable” way.

  For “tight” turns, it may be important that the two protein-binding sites on the DNA are on the correct face of the DNA helix. Fig 9.24a illustrates this. Also recall Fig 7.24.

  Fig 9.28 shows that the two sites may be “independent”. That is, the two proteins must bind to the two sites, but the looping per se is not an issue. Note that this may apply to some cases, not all.

  Such “action at a distance” is a minor issue in bacteria, with small and compact genomes. It becomes a much bigger issue with more complex eukaryotic genomes.

- Similarly, protein structures change. An important class of such changes is when the protein changes between an “active” and an “inactive” form by binding a small molecule. The example of the effect of Trp on the shape of the Trp repressor is spectacular (p 245, start of sub-section; also see Fig 9.15). However, some shape change presumably occurs in any case of a metabolite modulating a regulatory protein.

FR that deal with DNA structural issues related to DNA binding include Cheema et al (1999); Azam et al (1999). Foster et al (1999) explore drugs that enhance DNA-binding of a critical protein. Viadiu and Aggarwal (2000) examine specific vs non-specific binding and the role of a protein conformational change.

E. DNA-binding motifs

Certain protein structural motifs recur in many DNA-binding proteins. The helix-turn-helix (HTH) motif is one; a specific key part of this is the recognition helix.

Figs 9.1 & 2 show generally how the recognition helix fits into the major groove of the DNA (for one common class of regulatory proteins). Fig 9.3 & 4 show how specific amino acids on one face of the helix are responsible. In particular, changing these specific amino acids from one set to another changes the binding specificity appropriately. (Fig 9.4 would be even more useful if it included the controls. Weaver notes that the modified repressor no longer binds to its original operator. Further, the original 434 repressor does not footprint with this piece of P22 operator DNA, even at $4x10^{-6}$ M. These controls show that the binding sites are distinct.) Fig 9.8 shows the specific hydrogen bonding interactions.

Fig 9.9 shows interactions that involve the backbones of the DNA and/or protein. Also see Fig 9.8c, extreme left. At first thought, it might seem that these could contribute to overall binding energy, but not to specificity. However, since the microstructure of DNA and protein is sequence-dependent, it is possible that there is also some sequence-dependence of these backbone interactions. Again, recall earlier discussions of indirect readout, above.

Lest one overemphasize the importance of the HTH motif…

- There are many other DNA-binding motifs. Sect 12.1 lists some of them. Sect 12.2 elaborates, but we will not discuss this chapter.

- It is common that proteins with similar functions have similar structures. That is, the finding of DNA-binding motifs is simply one more example of a family of proteins with similar structure and function.

- We can analyze how any particular HTH protein recognizes DNA. However, it is hard to generalize. Not all HTH proteins face DNA exactly the same way, and there is no simple one-to-one code for recognition. Weaver uses the Trp repressor, Fig 9.17, to illustrate that HTH proteins may use different strategies for facing the DNA.

  The discussion of the Trp repressor also illustrates the importance of backbone contacts and of bridging water molecules, Fig 9.18. Recall Ch 2 hw #10. Weaver refers to contacts involving bridging water molecules as “indirect contact” (p 247).

Lewis et al (2000) examine an RNA-binding motif.

F. Further reading

C L Kielkopf et al, A structural basis for recognition of A·T and T·A base pairs in the minor groove of B-DNA. Science 282:111, 10/2/98. They have developed synthetic polymers (polyamides) that can distinguish all four possible base pairs of DS DNA -- through the minor groove. What makes this particularly interesting is that such discrimination would seem unlikely. The H-bonding ability of AT and TA base pairs in the minor groove appear to be the same, but are very subtly different.

E M Marcotte et al, Detecting protein function and protein-protein interactions from genome sequences. Science 285:751, 7/30/99. This paper raises some interesting issues, only vaguely related to the current chapter. Their central theme is determining protein-protein interactions. Their tool is using the ever-increasing databases of genome (hence protein) sequences. Their logical approach is to note that some protein-protein interactions are mimicked as single multi-domain proteins in other organisms. Thus they search genome databases for cases where domains of one protein seem to be in separate proteins in other organisms. Much of the paper discusses the logic of the method, and its limitations.

A K Cheema et al, A- and T-tract-mediated intrinsic curvature in native DNA between the binding site of the upstream activator NtrC and the nifLA promoter of Klebsiella pneumoniae facilitates transcription. J Bacteriol 181(17):5296-302, 9/99. Recall that activation of a σ^{54} promoter involves a distant site, and hence requires that the DNA bend. Here they show that the region between the promoter and activator sites is “intrinsically curved” -- a feature of A-rich sequences. (In other cases, DNA-bending proteins are involved in such activations.)

T A Azam et al, Growth phase-dependent variation in protein composition of the Escherichia coli nucleoid. J Bact 181:6361, 10/99. They measure the amounts of 12 proteins that are part of the E coli nucleoid. Some are DNA-bending proteins. They find major changes in the protein composition of the nucleoid as the bacteria go into stationary phase.

L H Hartwell et al, From molecular to modular cell biology. Nature 402 Suppl:C47, 12/2/99. This article is part of a special Nature supplement, Impacts of Foreseeable Science. It attempts to organize what will be learned over the coming years. In this case, the idea is to look at the level above simple interactions of, say, one protein with one DNA molecule. Nice reading!

> *** On Monday Oct 8, it was announced that Lee Hartwell, Univ Washington, is co-recipient of this year’s “biology” Nobel prize, for his pioneering work on cell cycle regulation. Co-recipients are Tim Hunt and Paul Nurse, of England. Long overdue.

B A Foster et al, Pharmacological rescue of mutant p53 conformation and function. Science 286:2507, 12/24/99. (+ News, Pennisi, p 2431.) p53 is a key DNA-binding regulatory protein, which is defective in a very high percentage of human cancers. In this work, they identify small molecules (“drugs”) that bind and stabilize the DNA-binding domain of some mutant p53. In a model system, they show that the drug reduces tumor growth. This work relates to protein folding, DNA binding, eukaryotic gene regulation, and cancer.

H A Lewis et al, Sequence-specific RNA binding by a Nova KH domain: Implications for paraneoplastic diseases and the fragile X syndrome. Cell 100:323, 2/4/00. Studying how proteins bind to RNA is also an important issue. Most such binding is to SS RNA, thus allowing more variation. The KH domain discussed here is found in many RNA-binding proteins.

H Viadiu & A K Aggarwal, Structure of BamHI bound to nonspecific DNA: A model for DNA sliding. Molecular Cell 5(5):889, 5/00. BamH1 is a restriction endonuclease, a type of enzyme that has extremely high discrimination of target vs non-target DNA. Here, they determine the crystal structure of a complex of the enzyme with a DNA that differs in one base pair from the target sequence. They show that the protein is in a distinct conformation, between that of the free protein and the properly bound protein. This conformation may allow good sliding along the DNA (so-called one-dimensional diffusion), which may be important for many DNA-binding proteins.

C C Sze et al, In vivo and in vitro effects of integration host factor at the DmpR-regulated σ^{54}-dependent Po promoter. J Bacteriol 183:2842-2851, 5/01. See hw.

G Smutzer, X-ray vision in structural genomics. The Scientist, 6/11/01, p 28. Nice review of the history of using X-ray analysis of structures, up through modern development of high-throughput methods. (The Scientist is available online at http://www.the-scientist.com.)

G. Errata

p 239, Fig 9.8c. For base pair 6, the Asn55 side chain has an extra O. Should be $-NH_2$ on the carbonyl O, not $-ONH_2$.

H. Homework

The following questions from Weaver are assigned, and we will discuss them in class:

W3. Note that you are not expected to know anything in particular about the underlying biology, only that the stated effect is true.

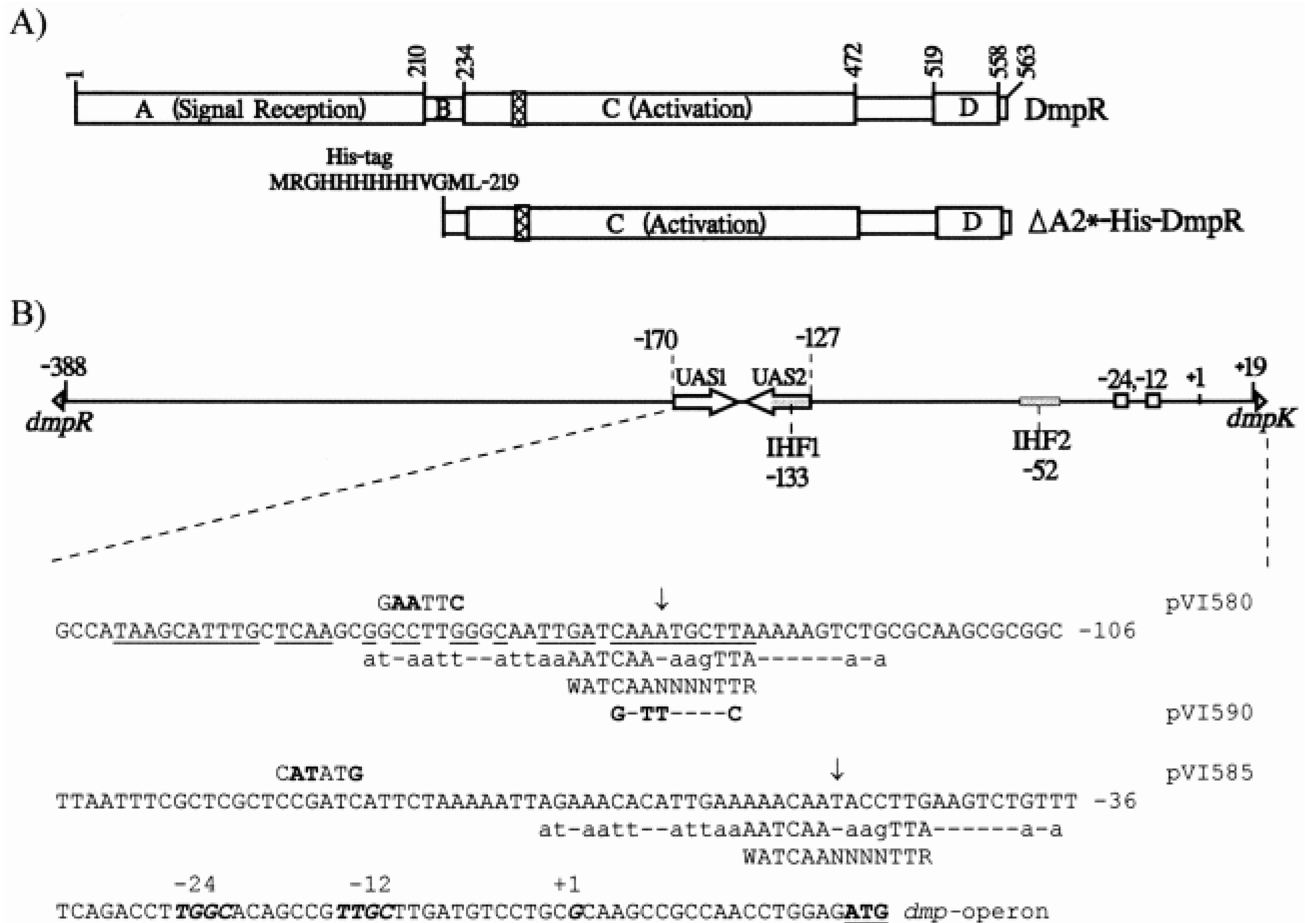
W4.

W5.
– "Methylene" is a common name for $-CH_2-$ groups.
– Added part: Where in the DNA are such groups likely to interact?

W7.

Added question:

This question is intended to broadly cover the set of Chapters 6-9. The question is based on a recent article on a particular promoter. It happens to be a promoter for σ^{54} (Fig 8.15), but the issues are general. The following Fig lays out the background. Part A of the Fig sketches the gene-specific activator protein, called DmpR. IHF is also required for activation. Part A also includes a specially constructed mutant form of the protein. Part B shows the DNA region that is relevant. The lower part of B shows the exact DNA sequence for part of this region; with one exception, you won't need this detail.



⇒ There is another copy of this Fig on the back page of the handout; it may be easier to read, especially for the details of the DNA sequence, which you need for one part.

a. The protein seems to have four domains, called A, B, C & D. B is simply a small spacer. What is domain D for? (The "D" is just a coincidence -- a nice coincidence.)

b. How long (how many amino acids) is the mutant protein shown there? Explain how you figure it out.

c. Identify the promoter elements and transcription start site. (Remember that this is a σ^{54} event; that σ has a distinct promoter.)

d. What is the (name of the) first gene whose transcription is being studied here?

e. The DmpR binding site is an inverted repeat, which is clearly shown but not labeled as the DmpR binding site. Where is it?

IHF is a DNA-binding protein whose role is often to bend the DNA. Two possible IHF-binding sites in the presumed *dmp* regulatory region have been identified by computer analysis; they are shown on the map. The consensus IHF binding site is shown along with the expanded DNA sequence; it is the line beginning with W (W = A or T; R = puRine) that accompanies each site.

f. Which of these two possible IHF binding sites is closer to the consensus sequence shown.

g. Which of these IHF binding sites seems more likely to be functionally relevant in this case? Why? I think there are <u>two</u> clues available to you, both of which point to the same conclusion. (First… what <u>is</u> the function?)

h. How would you go about trying to determine which IHF binding site is functionally relevant? Try to describe two distinct approaches, based on material in this group of chapters.

I. Partial answers

W3. Both are similar types of interactions, with a HTH protein recognizing a small DNA region, by site-specific contacts. The three operator sites have slightly different sequences. They are similar enough that one protein can bind to all three; both the cI repressor and Cro protein do so. But they are different enough that proteins do see them differently, by having different key amino acids at the critical contact points. Although both proteins recognize all three sites, cI repressor and Cro have different "preferences". See Figs 8.26 & 32.

W4. Recall chart of amino acids, in Ch 3. These amino acids do not form ions, but they are polar. Thus they form hydrogen bonds.

W5. hydrophobic

W7. This is based on a case discussed on p 244. Your comments (skepticisms?) beyond that are encouraged.

Added question:

a. What key function of a gene regulatory protein is not shown? DNA-binding.

b. The original protein is 563 aa long. The mutant protein is missing the first 218 of those (the Fig says that it starts with #219), plus it has an added tail of 13 aa. Thus the mutant protein has 358 aa.

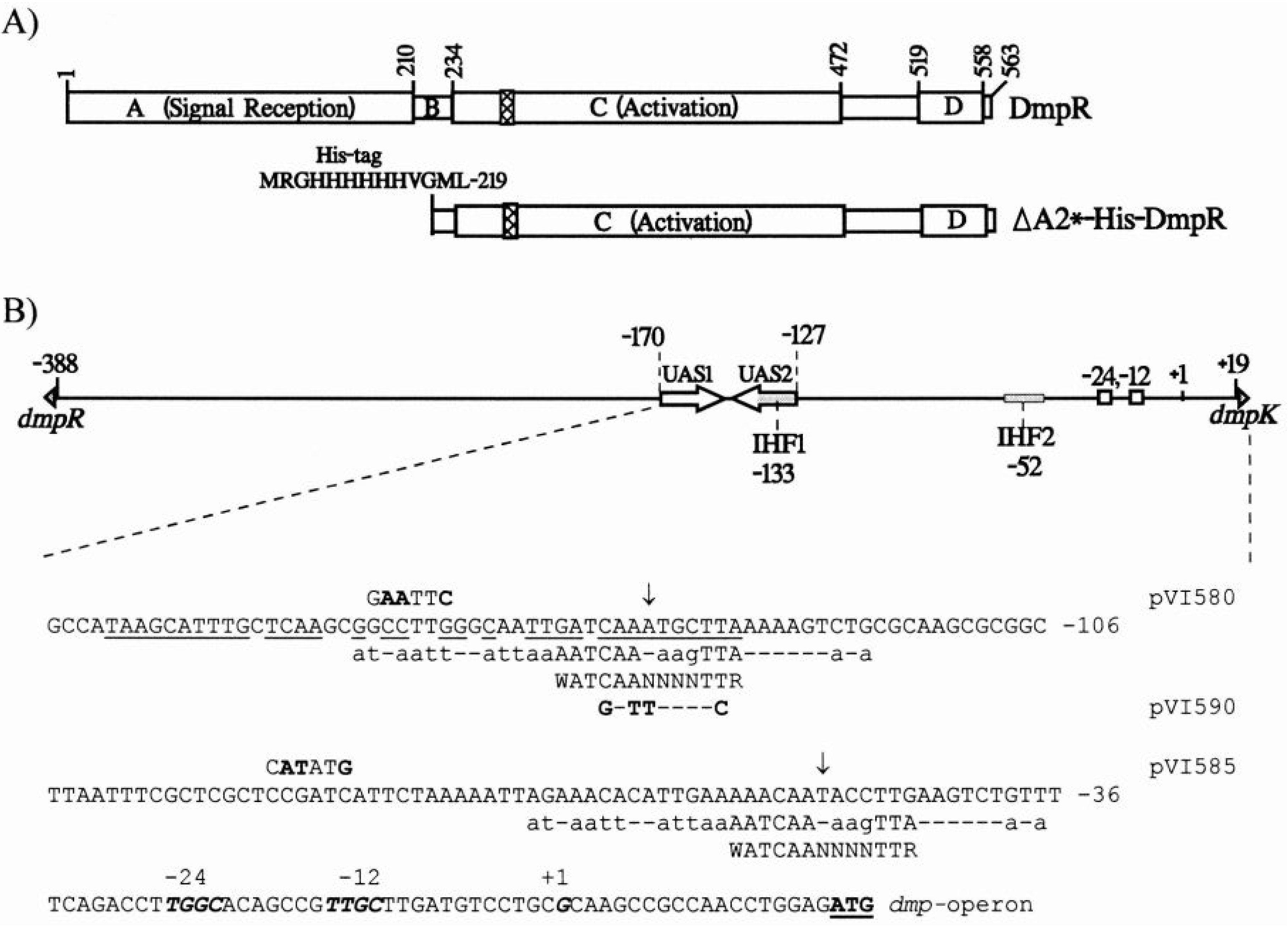
c. The promoter elements are shown in Part B as the -24 and -12 boxes. The transcription start site is +1. (This is all "normal" labeling and numbering. The only thing unusual here is that this is a σ^{54} promoter, so the promoter elements are -12 and -24, rather than the more common -10 and -35. The Fig is so clear that you probably should be able to figure that out, even if you did not remember the σ^{54} promoter.)

d. *dmpK*. It says so.

e. UAS1-UAS2 region. The arrows show the inverted repeat.

f. IHF1 differs from the consensus only at the first position. IHF2, only at the 3rd position.

> It is probably best to draw no particular conclusion from this, other than that both sites are plausible IHF sites. For one thing, the issue of the IHF consensus is murky. In fact, the line above the W… line is a more recent (and more complex!) description of the IHF binding consensus.
>
> This question is based on Sze et al (2001).


A)
1
210
234
472
519
558
563
A (Signal Reception)
B
C (Activation)
D
DmpR
His-tag
MRGHHHHHHVGML-219
C (Activation)
D
ΔA2*-His-DmpR
B)
-170
-127
-388
UAS1
UAS2
-24,-12
+1
+19
dmpR
dmpK
IHF1
-133
IHF2
-52
GAATTC
pVI580
GCCATAAGCATTTGCTCAAGCGGCCTTGGGCAATTGATCAAATGCTTAAAAAGTCTGCGCAAGCGCGGC -106
at-aatt--attaaAATCAA-aagTTA------a-a
WATCAANNNNTTR
G-TT----C
pVI590
CATATG
pVI585
TTAATTTCGCTCGCTCCGATCATTCTAAAAATTAGAAACACATTGAAAAACAATACCTTGAAGTCTGTTT -36
at-aatt--attaaAATCAA-aagTTA------a-a
WATCAANNNNTTR
-24 -12 +1
TCAGACCTTGGCACAGCCGTTGCTTGATGTCCTGCGCAAGCCGCCAACCTGGAGATG dmp-operon