Biochemistry 3107 - Fall 2003

The Bacterial Promoter

E. coli Promoters

Unlike DNA polymerase, RNA polymerase does not require a primer. It can synthesize RNA ***de novo*****.** However, RNA polymerase does not initiate RNA synthesis randomly in the cell. It will only initiate synthesis at specific places on the DNA template called **promoters**.

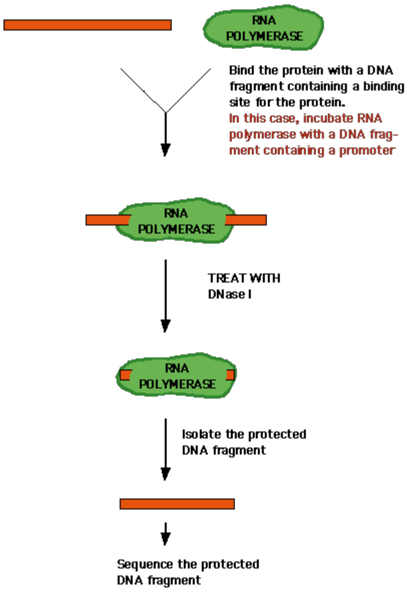
A great deal of research has focused on determining exactly what nucleotide sequences define a promoter. What sequence does RNA polymerase recognise and bind?

We can summarize the general approach to this question as follows:

- Can a specific small region of DNA be identified which binds with RNA polymerase?
- What is the sequence of such an RNA polymerase binding region?
- Are there are any nucleotides that are common to all such RNA polymerase binding regions?
- If so, can it be proven whether such nucleotides have a functional significance? Is there genetic or physical evidence to support their importance?

The first conserved sequence

In 1975 **David Pribnow** examined and compared the sequence of **five** RNA polymerase binding DNA fragments whose sequence had been determined by himself, by Heinz Schaller and by others. These sequences were determined after isolating the RNA polymerase binding fragments by a simple protocol:


RNA POLYMERASE
Bind the protein with a DNA fragment containing a binding site for the protein.
In this case, incubate RNA polymerase with a DNA fragment containing a promoter
RNA POLYMERASE
TREAT WITH DNase I
RNA POLYMERASE
Isolate the protected DNA fragment
Sequence the protected DNA fragment

[S33-2]

Note the following points about this experimental protocol:

- In this experiment, the enzyme **DNase I** is used to digest away all of the dsDNA that is not bound and is therefore not protected by RNA polymerase. This is **NOT** a **footprinting** experiment. That technique uses DNase I in a very different way.

  [S33-3]

- The fragments that are protected by RNA polymerase can be isolated because proteins and protein-DNA complexes bind to nitrocellulose whereas DNA does not. The DNA fragments that are bound as protein-DNA complexes can be released from those complexes for sequencing.

- The **Maxam-Gilbert** technique of DNA sequencing must be used to determine the sequence of these fragments - the **Sanger** dideoxy technique cannot be used.

When Pribnow & Schaller compared the sequences of the protected fragments so that the startpoints of transcription were more or less aligned with one another, the following results were obtained:

sequence	promoter
TGCTTCTGACTATAATAGACAGG**G**TAAAGACCTGATTTTTGA	fd
AAGTAACATGCAGTAAGATACAAATC**G**CTAGGTAACACTATCAG	T7 A2 promoter
GTAAACACGGTACGATGTACCAC**A**TGAAACGACAGTGAGTCA	T7 A3 promoter

ACCTCTGGCGGTGATAATGGTTGC**A**TGTACTAAGGAGGTTG	lambda P_R promoter
GCTTCCGGCTCGTATAATGTGTGG**A**ATTGTGAGCGGATAACAA	*lac*UV5 promoter
TA AT A T	**common bases that occur in at least 4 of the 5 sequences**

[S33-4]

Note the following:

- The startpoint of transcription is shown by the **red** base; the startpoint of transcription for the bottom two promoters is one base to the left of that for the top three.
- The first 4 sequences are bacteriophage promoters; the fifth sequence is from *E. coli*.

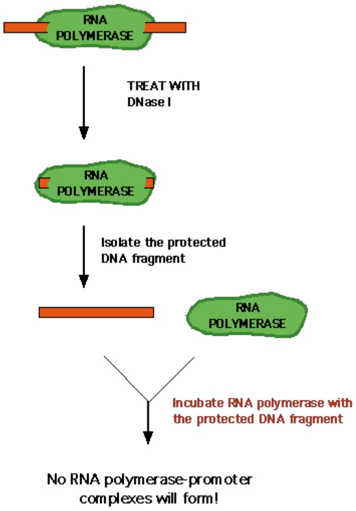
From this analysis, a conserved sequence centred 10 bp upstream of the startpoint of transcription was identified. The sequence was initially called the **Pribnow Box**. However, it is more usual now to refer to it simply as the **-10 region**.

Do not refer this sequence as a TATA box - that name applies only to eukaryotic promoters.

Note also that molecular biologists use a numbering system which has no zero! The first nucleotide of the RNA transcript is numbered **+1**; the nucleotide immediately upstream from that is numbered **-1**.

A second conserved sequence

After Pribnow had determined his conserved sequence, it soon became apparent that this was not sufficient to define an *E. coli* promoter. If the DNase I protected fragments were now re-incubated with RNA polymerase, it was found that they were no longer able to bind to RNA polymerase:


RNA POLYMERASE
TREAT WITH DNase I
RNA POLYMERASE
Isolate the protected DNA fragment
RNA POLYMERASE
Incubate RNA polymerase with the protected DNA fragment
No RNA polymerase-promoter complexes will form!

DNase I digestion must therefore have removed some of the nucleotides that are essential for RNA polymerase to recognise and bind to its promoter.

This experiment demonstrates that the enzyme recognises the upstream sequence before it fully recognises the -10 region.

Later experiments showed that the enzyme actually covers a region of DNA extending from **-55** to **+20.** RNA polymerase binding to a promoter is therefore a complex process involving at least some degree of conformational change in the DNA or the enzyme or both and the sensitivity to DNase I digestion reflects this change.

As a result of further sequence analysis, a second conserved sequence was discovered centred 35 bp upstream of the startpoint of transcription. This conserved sequence is known simply as the **-35 region**.

What is a consensus sequence?

Nucleotide sequences which share a common function, such as binding to the same protein, are often compared to see if they contain common nucleotides at fixed positions. The result usually is that they do -- and the resulting sequence is often called a **consensus sequence**.

It is important to realize that a consensus sequence is a **statistical** creature. It does not necessarily tell one what features are being recognised by any given protein though it is probably true to say that it does give one a pretty good idea. The true significance of bases belonging to any consensus sequence must be confirmed by genetic analysis, by physical analysis or by biochemical analysis.

We discussed as an example, how a consensus student ID number can be determined by comparing the student IDs of all the students enrolled in this course.

There are 58 students enrolled in the course this year. As you know, all MUN student IDs are 9 digit numbers, the first four of which indicate your year of first enrollment in the University. The following table is an analysis of the occurrence of the numbers 0..9 as each of the 7 possible digits:

No. of times each of the digits 0..9 occurs

9			8	8		4	6	3	**10**
8						4	**9**	6	7
7						7	5	4	7
6						6	5	7	5
5					1	6	4	4	**5**
4					1	3	7	**9**	4
3					8	**9**	2	**10**	6
2	**50**			1	16	**9**	8	4	4
1				**20**	14	3	4	7	7
0	8	**58**	**50**	**29**	**18**	7	8	4	**3**

The most common digit at each position is bold-faced in the table.

50 of this year's class have a **200** as the first three digits in your ID; and eight have **009** as the first three digits in their ID.

A majority of you (29) have **0** as the fourth digit; but 20 of you have **1** as the fourth digit.

The fifth digit is pretty evenly split between **0**, **1** and **2** with a majority (18) having **0**.

For the other positions, we need to establish a rule. Since each number should occur 5.8 times (58/10) on average, we might take twice this value - 12 - as the number of occurences that we will consider to be significant. Those values are in red in the above table.

These rules allows us to define a consensus for only five of the nine positions.

Now we can write the consensus ID for the students enrolled in this course as:

$$2_{50}0_{58}0_{50}0_{29}0_{18}({}^{2}/{}_{3})_{9}8_{9}3_{10}9_{10}$$

or, more simply:

$$20000(2/3)839$$

No one has this ID, of course, which should show you the limitations of a consensus ID. It does illustrate the fact that most of you have IDs starting with 20. It might also lead us to ask why so many of you have a 0 at the fourth position or a 1 at the fifth position. Similarly, a consensus sequence can help focus our thoughts and experiments on those nucleotides that might be especially important.

However, by writing down just one number as the consensus ID, we are also ignoring the fact that none of you have 4, 6, 7, 8 or 9 as the fifth digit in your ID or why so few of you have 2 at the eighth, position. Why should this be?

Similarly, consensus sequences do not tell the whole story. In particular they do not tell you which bases are never found at certain positions. Those reasons might be just as important for protein-DNA recognition as the reasons why certain bases are found.

Tom Schneider has written a lot about the limitations of consensus sequences and he has devised an alternative representation, called a **sequence logo** which show pictorially just how important every base in a site really might be.

Finally, for the record, the consensus class ID for previous years was:

1996	9426586
1997	952(3/6)428
1998	951947(0/8)
1999	9606305
2000	98164(0/2)4
2001	991(8/0)5(8/6)9
2002	99100(8/4/3)(5/0)

The *E. coli* promoter consensus sequence

Analysis of many *E. coli* promoters has revealed that there are 3 conserved elements in the *E. coli* promoter:

-35 sequence

Centred 35 base pairs upstream of the start-point of transcription, this sequence element has the consensus sequence **TTGACA**.

-10 sequence

Centred 10 base pairs upstream of the start-point of transcription, this sequence element has the consensus sequence **TATAAT**.

spacer

The distance between the above two conserved elements is also important and is conserved at **17±1** base pairs.

TTGACA ---- 17±1 ---- TATAAT

Just as we discussed above in considering the case of student IDs, we can ask three broad questions about the *E. coli* consensus promoter elements:

- Are all positions in the consensus sequences conserved equally well?
- Is there any practical meaning to the notion of a consensus sequence promoter?
- Is there any evidence that the particular bases in the consensus sequences are especially important?

Compilation analyses of promoter sequences

Over the years, a number of authors have compiled lists of *E. coli* promoters and analysed their sequences. From these the original consensus sequence elements and sequences have been confirmed. However, as more and more promoters were sequenced, it has also become clear that not all bases are as well conserved.

Harley & Reynolds' Compilation

In 1987, **Calvin Harley** and **Robert Reynolds** of McMaster University published an analysis of 263 **phage**, **plasmid** and **bacterial** promoters:

> "In the final compilation, all bases in the -35 (TTGACA) and -10 (TATAAT) hexamers were highly conserved, 92% of promoters had inter-region spacing or 17±1 bp, and 75% of the uniquely defined start points initiated 7±1 bases downstream of the -10 region."

The degree of conservation (%) of each base in their compilation of the consensus hexamers was:

$$\mathbf{T_{78}T_{82}G_{68}A_{58}C_{52}A_{54} \text{ -- } 16_{21}17_{52}18_{19} \text{ -- } T_{82}A_{89}T_{52}A_{59}A_{49}T_{89}}$$

Lisser and Margalit's compilation

One criticism that has always been levelled at promoter compilations such as Harley and Reynolds' is that it is biased. The bias arises in a couple of ways:

- it is easier to study strong promoters rather than weak promoters.
- bacteriophage promoters are inherently strong promoters and therefore may not be representative of a typical *E. coli* promoter.

For this reason, in 1993 **Shlomit Lisser** and **Hanah Margalit** looked exclusively at *E. coli* promoters. Their results showed some changes in the degree of conservation of some bases, but overall, the same results were obtained:

$$\mathbf{T_{69}T_{79}G_{61}A_{56}C_{54}A_{54} \text{ -- } 16_{17}17_{43}18_{17} \text{ -- } T_{77}A_{76}T_{60}A_{61}A_{56}T_{82}}$$

Compare results from the two compilations. Notice that while the degree of conservation of some of the most highly conserved nucleotides has decreased, that of some of the less highly conserved nucleotides has increased.

[25-5]

Finally, note that different forms of RNA polymerase will recognise different promoters. The different sigma factors confer different sequence specificity. Strictly speaking, therefore, a consensus sequence only applies to a particular RNA polymerase with its associated sigma subunit.

[S33-6]

The meaning of a promoter consensus sequence

Why is so much variability allowed in the sequence of an *E. coli* promoter?

The answer lies in the fact that, while *E. coli* RNA polymerase is designed to transcribe mRNA, not all mRNA molecules need be synthesized in the same amount. The easiest way to control the level of mRNA synthesis is to vary promoter sequences so that RNA polymerase will recognise some very well (those from which lots of mRNA is required) and some at all well (those from which little mRNA is required).

Thus there are **strong** promoters and **weak** promoters:

A strong promoter

The ***recA*** promoter is a strong promoter:

TTGATA -- 16 -- TATAAT

TTGACA -- 17 -- TATAAT

It differs from the consensus *E. coli* promoter in only one nucleotide and by one base pair in the spacer length.

A weak promoter

The ***araBAD*** promoter is a weak promoter:

CTGACG -- 18 -- TACTGT

TTGACA -- 17 -- TATAAT

This promoter differs from the consensus by 5 nucleotides and by one bp in the spacer length.

From this kind of observation, it has become clear that a consensus promoter sequence is likely to be a strong promoter. In fact, a promoter with the consensus sequence has not so far (to my knowledge) been found in *E. coli*. Such a promoter is likely to be too strong!

[Lod10-11ab]

Genetic evidence for the importance of the consensus sequence elements

A number of mutations have been characterized over the years in different promoters. For the most part, these mutations occur within the conserved hexamer elements or alter the spacer length between the elements.

- If a mutation (nucleotide change) makes the promoter a poorer match to the consensus sequence then the mutation is nearly always a **down** mutation - i.e. the promoter becomes weaker.
- If a mutation makes the promoter a better match to the consensus sequence then it is nearly always an **up** mutation - i.e. the promoter becomes stronger.

[Lod10-11ab][Lod10-11c]

For example, the ***lacUV5*** promoter is a better promoter than the ***lac*** wild-type promoter:

lac UV5	TTTACA -- 18 -- TATAAT
lac wild-type	TTTACA -- 18 -- TATGTT
	TTGACA -- 17 -- TATAAT

The wild-type promoter is of moderate strength - it differs from the consensus in 3 nucleotides and has a nonideal spacer length; the ***lacUV5*** promoter is much stronger - it differs from the consensus in only 1 nucleotide though it still has a nonideal spacer length.

Similarly, while the wild-type bacteriophage lambda **P_{RM}** promoter is a very poor promoter, the **P_{RM}**Up-1 mutant is quite a bit stronger:

P_{RM} wild-type	**TAGATA -- 17 -- TAGATT**
P_{RM}Up-1	**TAGACA -- 17 -- TAGATT**
	TTGACA -- 17 -- TATAAT

Although the wild-type **P_{RM}** promoter differs from the consensus in only 4 nucleotides and has a ideal spacer length, the T-> A change at the second position of the -35 region is particularly deleterious.

Physical evidence for the importance of the consensus sequence elements

Physical evidence for the importance of the conserved sequence elements of the promoter comes from studying the interaction of RNA polymerase with the promoter.

There are a number of physical techniques for studying the interaction between a protein and its DNA binding site. For this course, the technique that we learned about is **DNase footprinting**

[S33-3][Lod10-6]

This technique depends on:

- Relatively nonspecific digestion of DNA - i.e. every site is equally susceptible to digestion.
- Only one strand (of the two) is labelled at a time.
- The digestion is **LIMITED** so that each DNA fragment in a reaction tube is attacked just once by the enzyme.
- The results of DNase I digestion in the presence of protein are compared with those in the absence of protein.

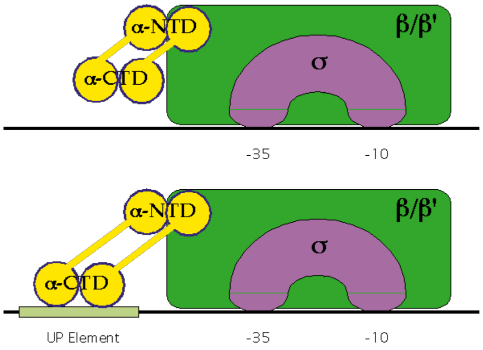
The DNase I footprint of RNA polymerase on an *E. coli* promoter extends from approximately **-50** to **+20**. Thus the conserved hexamer sequences are part of the footprint.

[Lod10-8]

Are there other important promoter sequence elements?

Yes! Although, the -35 and -10 regions define the core promoter in *E. coli*, some promoters contain an additional element upstream of ther -35 region, called the UP element. This element is located between -57 and -47 bp upstream of the startpoint of transcription. and is A/T rich. (5'-AAAATTATTTT-3').

The element is found in strong promoters such as ribosomal RNA (*rrn*) promoters. The carboxy-terminal domain of the α-subunit of RNA polymerase (α-CTD) binds to this element and enhances RNA polymerase binding by a factor of 10. In "normal" promoters, which do not contain the UP element, the α-CTD does not bind to DNA.



[The rrnB promoter]

RESOURCE MATERIAL

VOET, VOET & PRATT

1. Chapter 25, Transcription and RNA Processing, pages 816 - 817

STRYER

1. Chapter 5, Flow of Genetic Information, pages 101-102
2. Chapter 33, RNA Synthesis and Splicing, pages 842-844

LEHNINGER

1. Chapter 24, RNA Metabolism, pages 860 - 863
2. Chapter 27, Regulation of Gene Expression, pages 943 - 944

TAMARIN

1. Chapter 10, pages 238 - 243

WEB SITES

- **Tom Schneider** devised the sequence logo approach for representing conservation in sequence elements. He has a page of more sequence logos among which you can view sequence logos for Yeast TATA Boxes, E. coli Ribosome binding sites and Human Splice Junctions. If you are interested in this area read Dr. Schneider's Poster on Information Theory and Individual Information. *Please note that I do not need you to know much about sequence logos for Biochemistry 3107. I do want you to know about consensus sequences and to be aware of their limitations.*
- Prokaryotic Promoters from MCB 411 and 411H Molecular Biology at the University of Arizona.