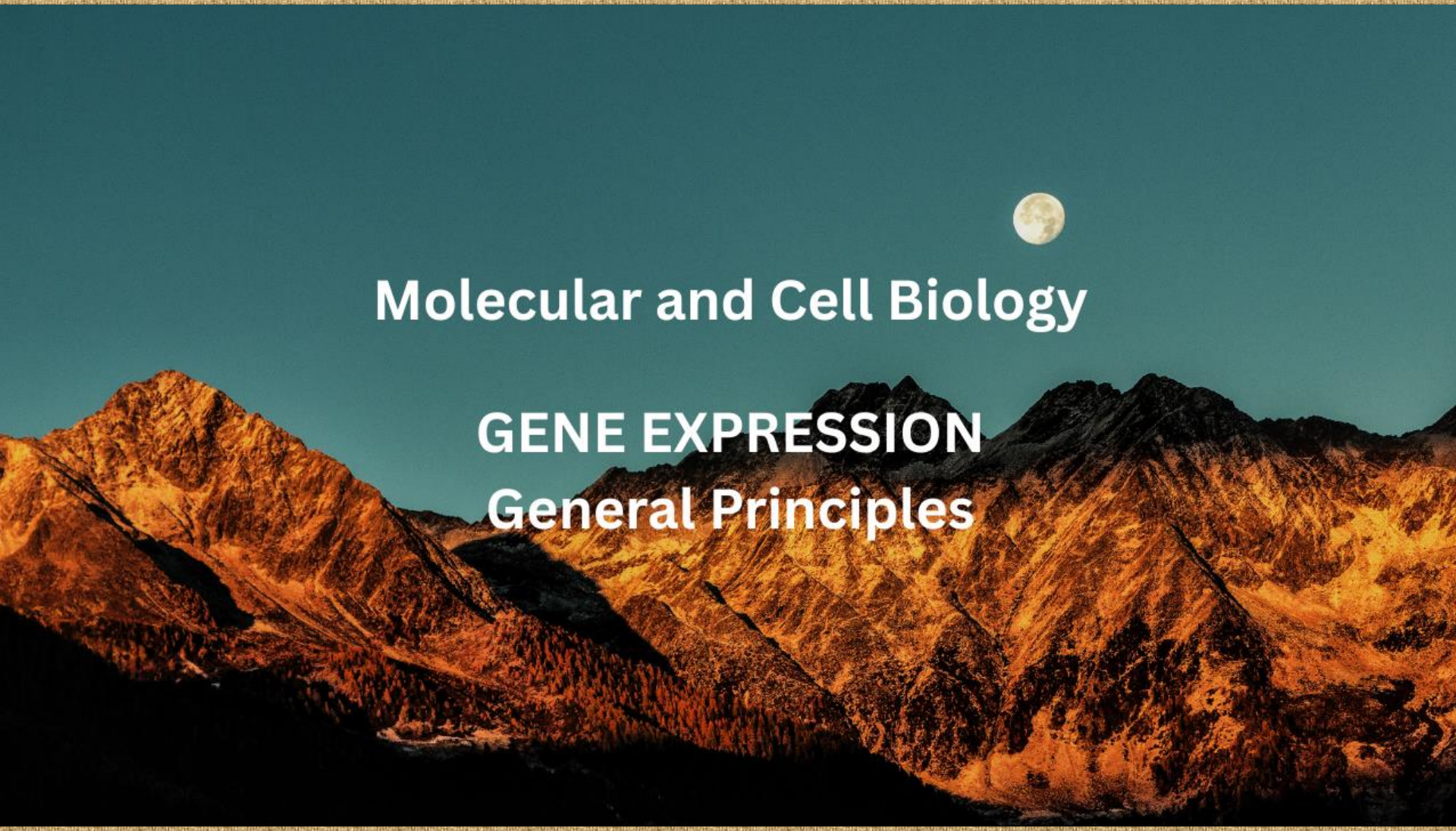
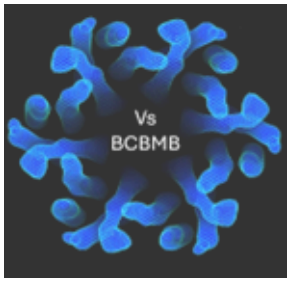

Molecular and Cell Biology
GENE EXPRESSION
General Principles

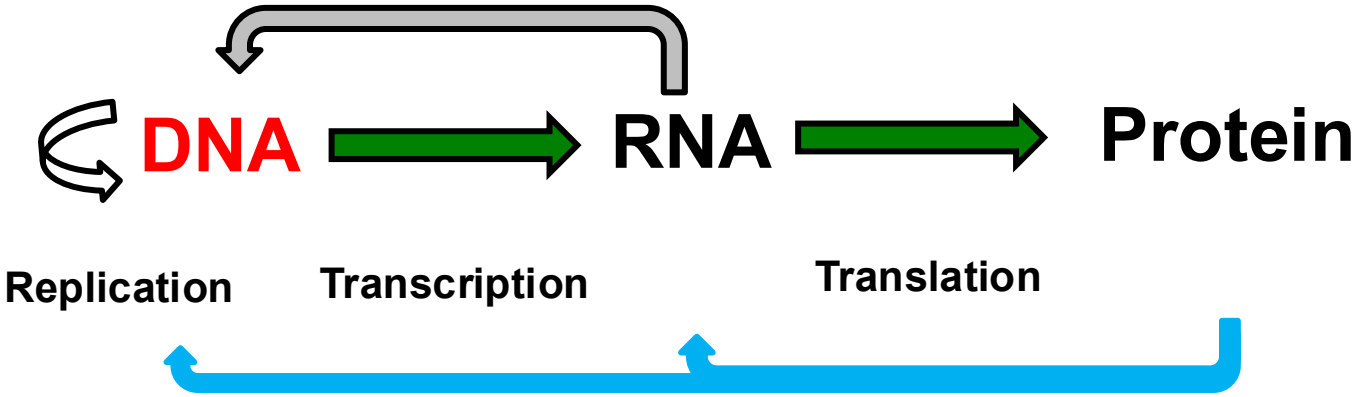


General Principles of Gene Expression
Setting the Stage

this chapter continues the general narrative of building a cell from the inside out and follows the chapters about Genomes and Chromatin.

looking at the "Central Dogma" to remind ourselves about the bigger picture

Reverse Transcription (the anti-dogma)

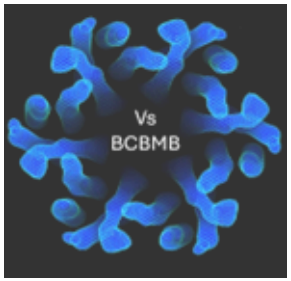


our narrative will unfold by looking at how cells work with DNA to

- retrieve information from DNA through the process of transcription
- how to cross from nucleic acids into protein space
- how to regulate gene expression

before looping back to take a look at

- how to duplicate the instruction manual, and
- how to regulate that duplication and associated cell division



General Principles of Gene Expression
Setting the Stage

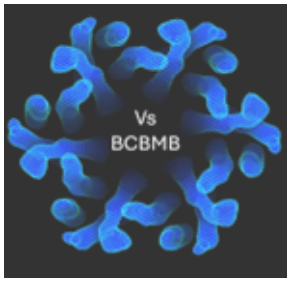
Goals

by the end of this chapter you will have a basic understanding of

- the chemical principles of the first step in gene expression – transcription
- how cells find the genes they need to transcribe
- the basic components of the transcription machinery how they assemble at the right sites
- differences in transcriptional outcomes in prokaryotes and eukaryotes
- types of transcripts produced

aspects of how gene expression is regulated are covered in a separate chapter

If you already have some prior knowledge about transcription, you will benefit from summarizing what you know before continuing - not just to "compare notes", but also because it will actively link the content of this chapter to what you already know



Transcription – The Basic Reaction

Before doing anything else – lets look at the basic reaction that occurs…because it will remind and solidify the importance of molecular asymmetry of nucleic acids.

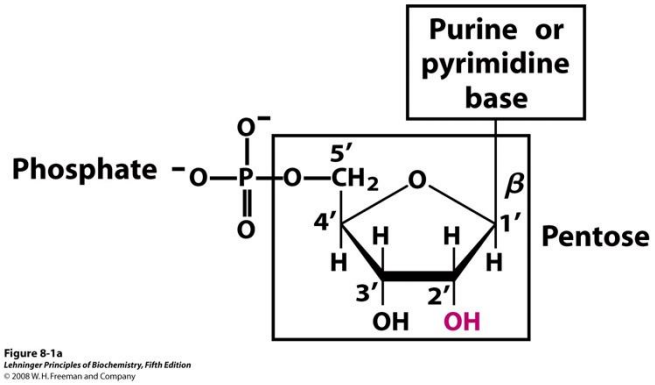


Figure 8-1a
Lehninger Principles of Biochemistry, Fifth Edition
© 2008 W. H. Freeman and Company

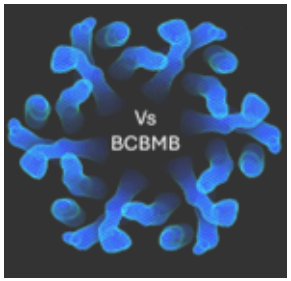
Slightly extending what we introduced in the "Fundamentals - NUCLEIC ACIDS" chapter

- The building blocks of nucleic acids are “nucleotides”.
- Each nucleotide is built from a nucleobase, a sugar (ribose [shown] or 2’-deoxyribose [red –OH = H]), and a phosphate
- Carbons in the sugar are numbered 1’ through 5’ (because plain numerals count in the nucleobase)
- The nucleobase is attached at 1’, the phosphate at 5’ and the –OH at 3’ is present in both ribose and deoxyribose.

While the generalized structure shown in the picture holds for what you will observe in the fully formed polymer, a different form of these molecules is needed for the actual synthesis of the polymer.

Specifically … what do you already know about this?

The substrates for synthesis of nucleic acids are ……..*complete the sentence if you can ….*

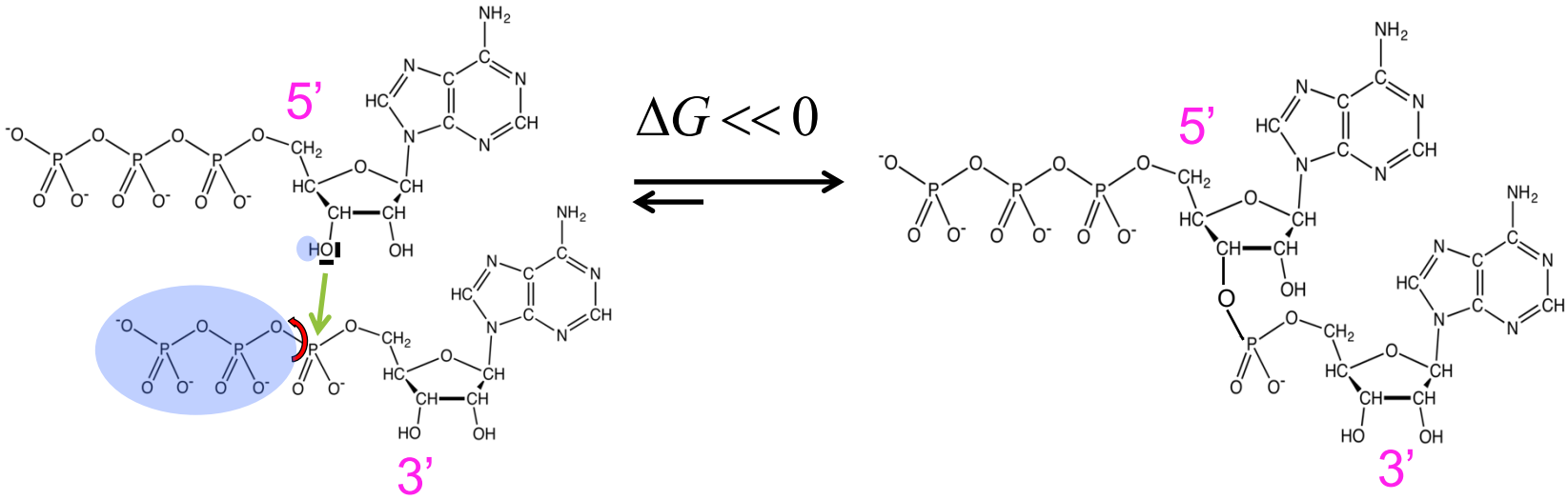


Transcription – The Basic Reaction

The substrates for synthesis of nucleic acids are
NTPs (nucleoside-triphosphates) for RNA
and dNTPs (deoxynucleoside-triphosphates) for DNA

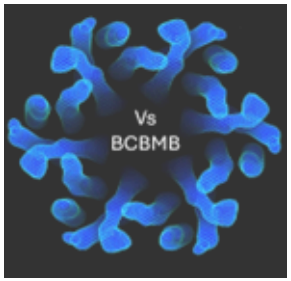
Synthesis of RNA uses “NTPs” (nucleoside-triphosphates) instead of the basic nucleotides as reactants because the elimination of inorganic pyrophosphate (PP_i) during the condensation reaction is very exergonic = energetic favorable (especially if coupled to the further hydrolysis of PP_i to two inorganic phosphates)

For instance: 2 ATP → pppApA-OH + PP_i



This reaction can be repeated by extending the growing chain at the 3’ –OH group
➔ eventually creates an asymmetric polymer that has **directionality 5’ (phosphate end) → 3’ (hydroxyl end)**

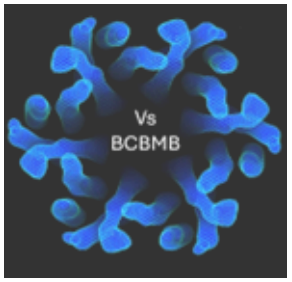
Of note: in RNA the nucleobase Uracil replaces Thymine (forming A::U pairs instead of A::T pairs)
This seemingly "random" fact is not random …. it represent one chemical marker that distinguishes RNA from DNA and has to do with chemical stabilities of the different nucleotides. We will look at this more in the Advanced Biochemistry - NUCLEIC ACIDS chapter where you will learn why using "U" in DNA would have catastrophic consequences



Transcription - Challenges

With the chemistry out of the way….lets next create a big picture view of transcription by verbalizing the challenges that we will encounter.....

....*what do you think they are?*



Transcription - Challenges

With the chemistry out of the way….lets next create a big picture view of transcription by verbalizing the challenges that we will encounter.....

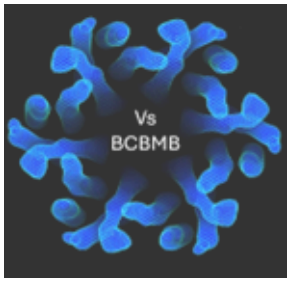
- **How to find what you want to read out:**

Did you "find" them all?

- **How to decide which DNA strand to use**

- **How to actually make the transcript**

- **If necessary, how to process the transcript to prepare it for the next step = translation.**



Transcription - Challenges

With the chemistry out of the way….lets next create a big picture view of transcription by verbalizing the challenges that we will encounter.....

- **How to find what you want to read out:**

This is a challenge because in a human genome you have ~6x10^9bp of potential coding regions. The average length of a primary transcript (don't worry about exact nomenclature here…we will get to that later) is 10,000-30,000 bp. Your chance of transcribing exactly the stretch needed when it is needed is – for all practical purposes - "0"

- **How to decide which DNA strand to use**

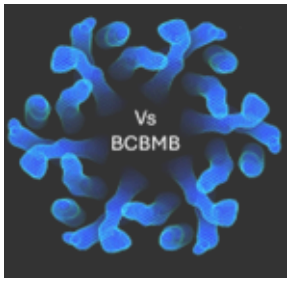
Even if you find the right place …. you still have two choices because DNA is double stranded = either could serve as template for the 5'→3' synthesis of RNA, but only one of these two potential transcripts will make sense

- **How to actually make the transcript**

Even if you have decided which strand to use as template… you still need to actually make the RNA strand because it will not make itself ..how do you get this started?, how do you deal with the idiosyncrasies of the chromatin structure?, how "carefully" will you be makin the transcript?, how fast can you go?, how do you know that you are done? … all relevant questions

- **If necessary, how to process the transcript to prep it for the next step = translation.**

This brings us to characteristic differences in prokaryotes and eukaryotes …
If you worked through the "Genomes" chapter, you may remember that eukaryotic genes have non-coding regions inserted into their genes (slide 37) … these "introns" need to be removed to produce a transcript that can be translated into a protein = how does the process of intron removal work?



Transcription - Challenges

Starting to address these points, we want to first think more carefully about the first big challenge….

- **How to find what you want to read out**

The justification we gave (big genome, lots of potential target sites) is quite intuitive and not particularly difficult to appreciate.

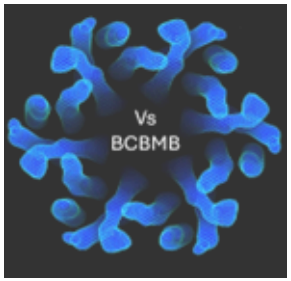
Letting this enormous mismatch between the size of eukaryotic genomes and the size of the transcripts sink in, you may come to realize that just this mismatch by itself already gives you some "pointers/hints" about mechanism.

That is because even without knowing anything about the mechanisms you may be able to intuitively infer that this challenge has at least **two layers of complexity**:

- a purely **spatial challenge** (“do you find it, or does it find you?”) and a
- **chemical/molecular challenge** (how do I recognize where to start?)

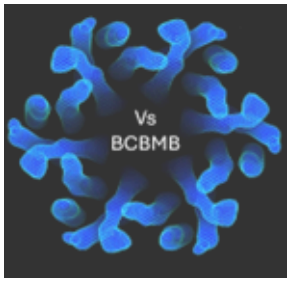
The **first point simply asks what is easier/more feasible** to do: pre-assemble the necessary transcription machinery in certain locations and "drag" to these locations the genes that need transcribing … or leave genes in their normal location and let all components needed for transcription go on a random "hunting mission" to find what needs to be transcribed and when.

Posed this way **… you may start to see that biology, at times, is quite logical …. Why?**



Transcription - Challenges

because …you already know the answer by taking a look at your own life realities



Transcription - Challenges

take "Office Hours" as example: …. do you and your peers go to the professor's office (or any agreed upon location), or does your instructor individually tries to find each of you at a time when you are available?

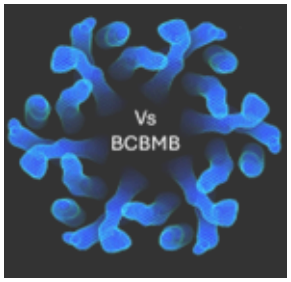
or "lunch time in the cafeteria" ….do you and your peers go there to select your lunch from various food stations…..or is there delivery service where staff from the different stations (greens, sandwiches, entrée options, sides, ….) individually bring your individual choices to your dorm room on demand?

In most places YOU will go to get what you want…. = "you find it" (not the other way round).

Same is true in eukaryotes: genes that need to be expressed are brought to sites where transcription happens (more about this later).

Interestingly though: the small size of prokaryotic cells + their smaller genome make the task of finding what you look for so much easier that in prokaryotes, the transcription machinery is the "hunter", finding what needs transcribing.

… casting all this into more formal language and putting some visuals to it …..



Transcription – How to Find What You Want

Finding the information a cell needs to transcribe is a formidable challenge. The spatial aspects of it "**do you find it (= the gene), or does it find you**"?

Answer: depends on cell type – *prokaryotes …mostly "you find it",* but in eukaryotes that approach would be hopeless → mostly follows a "it will find you" approach. **Why would it be hopeless to try find transcription start sites in eukaryotes?**

4.6×10^6 bp total

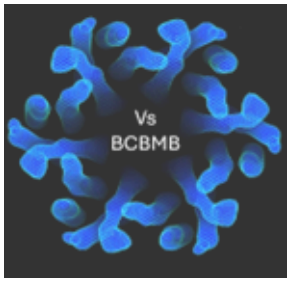
Answer: size of genome

- bacterial genomes are small and lean, occupying a total volume that is far smaller than a eukaryotic nucleus → [concentrations] of all required components are large enough to keep a random walk process feasible

- eukaryotic genome: too large to find small targets by random searches (**remember:** only ~1.5% code for transcripts) → need to be more organized → "transcription factories" (=functionally defined compartments within nucleus that are dedicated to carry out transcription)

~47-250$\times 10^6$ bp **per chromosome**
(picture shows metaphase chromosomes)

The idea of "**transcription factories**" makes intuitive sense
….but also creates real issues – **why?**



Transcription – How to Find What You Want

The idea of "**transcription factories**" in eukaryotes makes intuitive sense ….but also creates real issues – **why?**

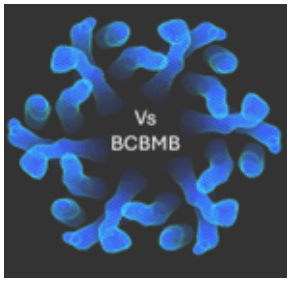
Answer: extreme compaction/folding of genome during cell division (= metaphase chromosomes) prohibits access to "readable" information!

➔ after cell division, compaction of chromosomes **must** be partially reversed to make the information physically accessible for read out.

This raises two questions:

- will chromosomes be evenly spread out across entire nucleus in this "unfolding"? (=random dispersion)
- are there defined and/or static places for establishing the factories?

~47-250x10^6 bp **per chromosome**
(picture shows metaphase chromosomes)



Transcription – How to Find What You Want

Answer 1:
random dispersion? NO

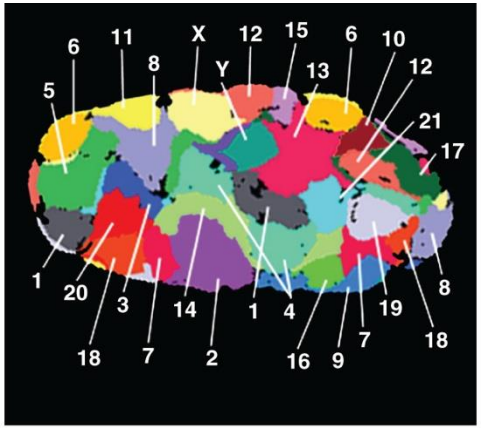
"random dispersion" of the chromosomes would be a disaster
for two reasons:

- **physical entanglement** would create impossible challenges to refold chromosomes for separating them after duplication
- **functionally related genes (either similar functions [A, A', A'', A'",...] or different functions within a particular process [A,B,C, ..]) tend to cluster on chromosomes**
  = keeping them spatially close helps with copying the right things at the right times.

Conclusion:

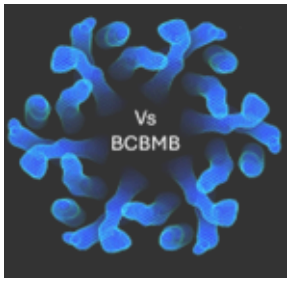
- unfolding metaphase chromosomes after cell division yields **euchromatin chromosome territories** that are folded "loosely" enough to allow information to be accessed for readout.

➔Need to build **many transcription factories to serve the different territories.**



"painting" each chromosome with a different color allows analysis of chromosome territories during interphase. Note how these territories are an example for the concept of compartmentalization

Amazing trivia: the positions and relative spatial relations of chromosome territories within the interphase nucleus are **non-random and largely conserved** across different cell types and, in many cases, across different species



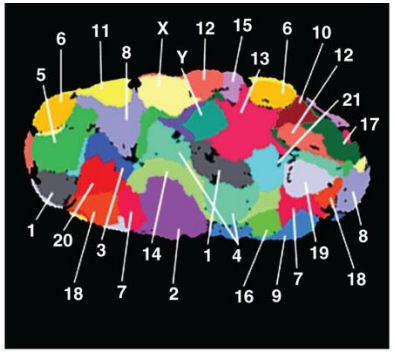
Transcription – How to Find What You Want

Answer 2:
defined/static places for formation of factories? still in progress

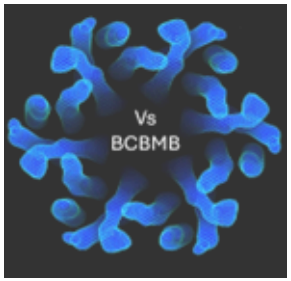
while stable over significant periods of time, factories are heterogeneous and dynamic = some are specialized to make only a few defined transcripts, others are more generic and switch between substrates; factories can disassemble and re-emerge in different positions.

current mechanistic studies favor a model in which transcription factories **are MLOs (membrane less organelles).** These structures are biological condensates and arise through "liquid-liquid phase separations" that are triggered if certain types of biological macromolecules (for instance: a combination of nucleic acids and certain proteins) mix at certain ratios.
These specialized structures can be tuned to be very selective for what can come inside and – if necessary – be dispersed very quickly by chemically modifying one or more of the components required for the phase separation to persist (which is one reason that this model is very attractive)

Another attractive feature of MLO model: it is easy to appreciate how many transcription factories can be established all over the interphase nucleus because these structures exploit local variances in chemical composition rather than requiring dedicated molecular architectures that are harder to establish/maintain



Regardless of precise mechanism: "transcription factories" are yet another example for compartmentalization!



Transcription - Challenges

Looping back to the two levels of mechanistic complexity:

- *a purely **spatial challenge** ("do you find it, or does it find you?") and a*
- ***chemical/molecular challenge** (how do I recognize where to start?)*

You now have an intuitive answer for the first
big space, lots of moving parts/components/unknown location of target → target needs to find the place where transcription happens because only that maintains a small "search radius" for all the components of the transcription machinery (preventing them from "getting lost in space")

The second part of the challenge is easier to understand ….

… you still need to figure out where to start, even when you "feed" the right stretch of DNA to the transcription machinery, ….

….lets take a look how that is done…..

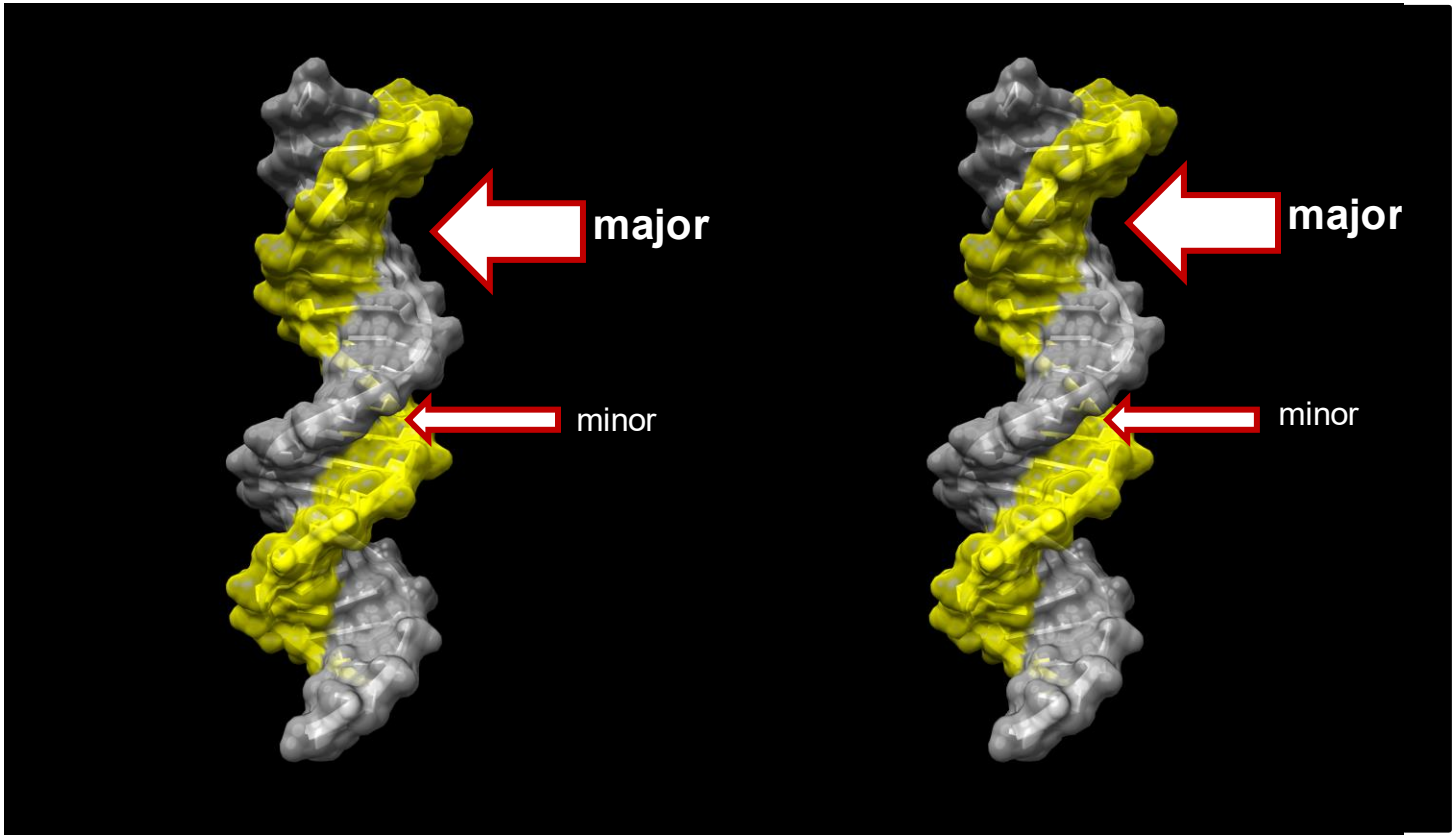
Transcription – Seeing DNA Helps to Understand

To warm up – lets look at DNA structure to refresh/reinforce what you likely know already …
the antiparallel orientation of the two strands causes the formation of two distinct regions – called "grooves" - along the surface of the double strand.

The narrower groove is called "minor groove", the wider one "major groove".

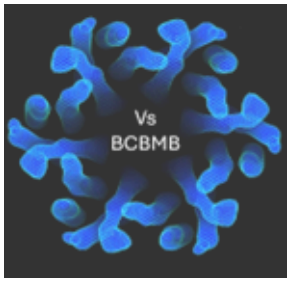
Key distinguishing feature: in the minor groove, the phosphate backbone ridges are closer to each other than in the major groove, creating a high negative charge density + "shielding" the base pairs at the center of the helix from access

These are cross-eyed stereo images – hold ~30cm away from you, cross your eyes and adjust your eyes + head tilt until you achieve 3D view



with those images on your mind - lets do this

How do you find where to start transcription?



Transcription – How to Find What You Want

Answer: since compartmentalization seems to be the answer to "everything" (territories, transcription factories)…is it here as well?

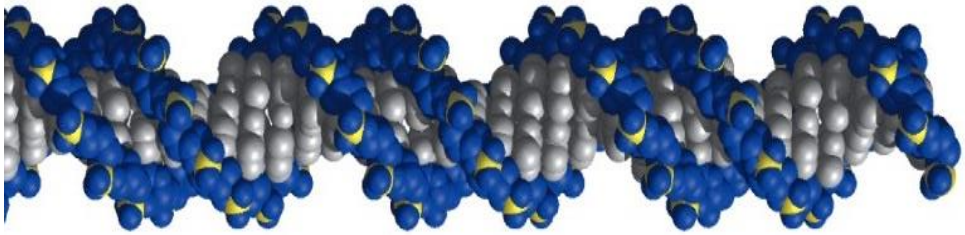
Answer:
yes BUT
not in the sense of a distinct macroscopic/physical structure
(like an MLO, or chromosome territory)

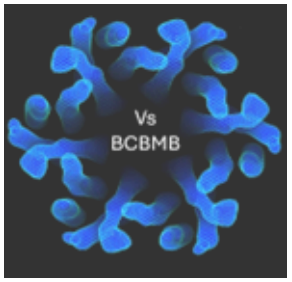
→ the **second component** of how to solve the search problem comes from **chemical compartmentalization** = you create a **signal** at the DNA sequence level that says "**start here ….**"

➔ the structure that marks transcription start sites is called **promoter**

Instead of just showing you a cartoon of a promotor (like most textbooks or instructors will do) – lets think about it based on the DNA structure you just looked at because that will allow you to **understand** most of why promoters look they way they do

Lets start with a simple question: to serve as signal, promoters rely on sequence → **will this sequence-dependent signal be found in the "minor or major groove"?** *….your turn!!…..*





Transcription – How to Find What You Want

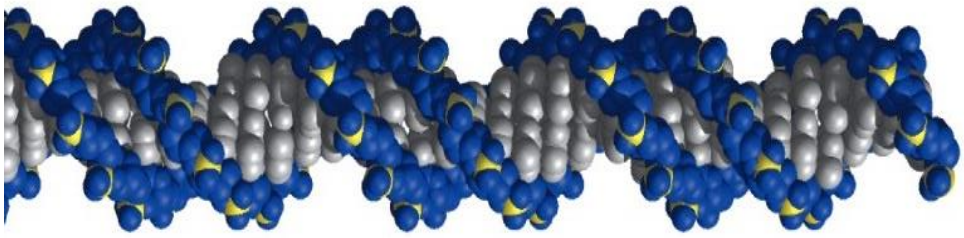
Answer: since compartmentalization seems to be the answer to "everything" (territories, transcription factories)…is it here as well?

Answer:
yes BUT
not in the sense of a distinct macroscopic/physical structure (like an MLO, or chromosome territory)
→ the **second component** of how to solve the search problem comes from **chemical compartmentalization** = you create a **signal** at the DNA sequence level that says "**start here ….**"
➔ the structure that marks transcription start sites is called **promoter**

Instead of just showing you a cartoon of a promotor (like most textbooks or instructors will do) – lets think about it based on the DNA structure you just looked at because that will allow you to **understand** most of why promoters look they way they do

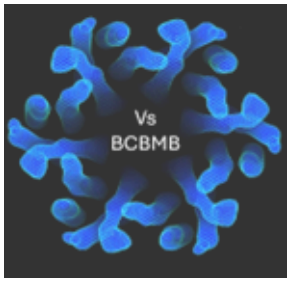
Let start with a simple question: to serve as signal, promoters rely on sequence → **will this sequence-dependent signal be found in the "minor or major groove"?**

Answer: major groove – because in it the bases are exposed enough to allow you to "look" at sequence and to evaluate whether it says "start here….".



scrutinizing DNA structure → roughly how many bp are "visible" in any one turn of the major groove?

Answer: look for yourself….



Transcription – How to Find What You Want

Answer: since compartmentalization seems to be the answer to "everything" (territories, transcription factories)…is it here as well?

Answer:
yes BUT
not in the sense of a distinct macroscopic/physical structure (like an MLO, or chromosome territory)

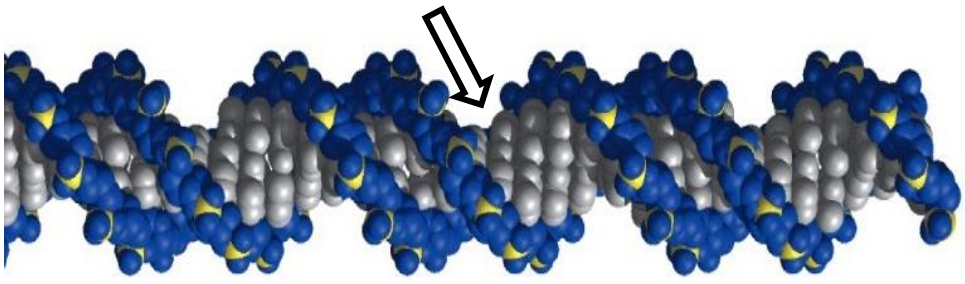
→ the **second component** of how to solve the search problem comes from **chemical compartmentalization** = you create a **signal** at the DNA sequence level that says "**start here ….**"

➔ the structure that marks transcription start sites is called **promoter**

Instead of just showing you a cartoon of a promotor (like most textbooks or instructors will do) – lets think about it based on what we know already because that will allow you to **understand** most of why promoters look they way they do

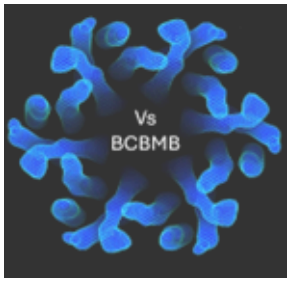
Let start with a simple question: to serve as signal, promoters rely on sequence → **will this sequence-dependent signal be found in the "minor or major groove"?**

Answer: major groove – because it allows to "look" at sequence and to evaluate whether it says "start here….".



scrutinizing DNA structure → roughly how many bp are "visible" in any one turn of the major groove?

Answer: 4 for sure, ~6 if you extend 1bp to the left/right of the center of the groove.

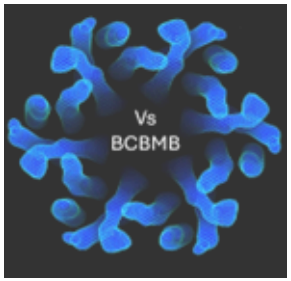


Transcription – How to Find What You Want

Being able to see ~6bp in the major groove, what is the probability for any 6bp sequence to occur?
(why would that be relevant? Answer: if it occurs too often, then it's not a good signal)

Answer: P(hexanucleotide) = $(1/4)^6$ =1/4096 → expect a given hexanucleotide to occur every ~4.1kb – is that good or bad?

Answer: *… try …. (hint: has something to do with what you learned in the Genomes chapter?)*



Transcription – How to Find What You Want

Being able to see ~6bp in the major groove, what is the probability for any 6bp sequence to occur?
(why would that be relevant? Answer: if it occurs too often, then it's not a good signal)

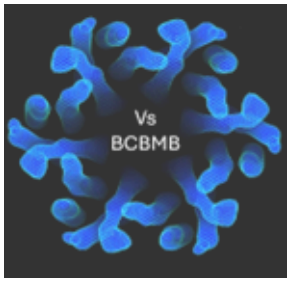
Answer: P(hexanucleotide) = $(1/4)^6$ =1/4096 → expect a given hexanucleotide to occur every ~4.1kb – is that good or bad?

Answer: depends (yet again) on what organism you look at

E coli: average size of gene is ~1kb ➔ a ~4.1kb average spacing for any given signal seems not completely unreasonable; you'd expect ~1122 copies of the hexanucleotide by chance, which is less than the 4255 you need to encode the entire E coli proteome.

- **BUT** "random" means "random" = copies would not actually **be** regularly spaced (= need to consider distribution which is very broad ranging from ten's of bp to tens of thousands of bp)
- **Also**: mutations could **not** be tolerated (5bp sequence occurs every ~1kb; 4bp occur every 256 bp)

→ **summary**: not entirely impossible, but **not likely to be viable**



Transcription – How to Find What You Want

Being able to see ~6bp in the major groove, what is the probability for any 6bp sequence to occur?
(why would that be relevant? Answer: if it occurs too often, then it's not a good signal)

Answer: P(hexanucleotide) = $(1/4)^6$ =1/4096 → expect a given hexanucleotide to occur every ~4.1kb – is that good or bad?

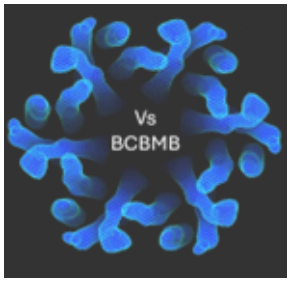
Answer: depends (yet again) on what organism you look at

E coli: average size of gene is ~1kb ➔ a ~4.1kb average spacing for any given signal seems not completely unreasonable; you'd expect ~1122 copies of the hexanucleotide by chance, which is less than the 4255 you need to encode the entire E coli proteome.

- **BUT** "random" means "random" = copies would not actually **be** regularly spaced (= need to consider distribution which is very broad ranging from ten's of bp to tens of thousands of bp)
- **Also**: mutations could not be tolerated (5bp sequence occurs every ~1kb, 4bp occur every 256 bp) → **summary**: not entirely impossible, but **not likely to be viable**

H sapiens: average size of protein still requires ~1kb of direct code, **BUT** accounting for introns, the actual average gene size is ~28,000 bp ➔ **game over….. any given hexanucleotide occurs too often to serve as start signal. ➔ How to solve this problem?**

Answer: …*can you think of anything? … be bold…..*



Transcription – How to Find What You Want

Being able to see ~6bp in the major groove, what is the probability of any 6bp sequence to occur?
(why would that be relevant? Answer: if it occurs too often, then it's not a good signal)

Answer: P(hexanucleotide) = $(1/4)^6$ =1/4096 → expect a given hexanucleotide to occur every ~4.1kb – is that good or bad?

Answer: depends (yet again) on what organism you look at

E coli: average size of gene is ~1kb ➔ a ~4.1kb average spacing for any given signal seems not completely unreasonable; you'd expect ~1122 copies of the hexanucleotide by chance, which is less than the 4255 you need to encode the entire E coli proteome.

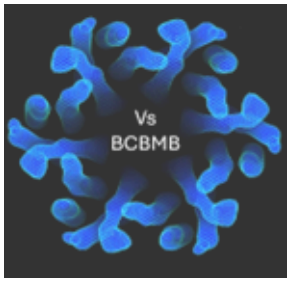
- **BUT** "random" means "random" = copies would not actually **be** regularly spaced (= need to consider distribution which is very broad ranging from ten's of bp to tens of thousands of bp)
- **Also consider**: mutations could not be tolerated (5bp sequence occurs every ~1kb, 4bp occur every 256 bp) → **summary**: not entirely impossible, but **not likely to be viable**

H sapiens: average size of protein still requires ~1kb of direct code, **BUT** accounting for introns, the actual average gene size is ~28 **k**bp ➔ **game over….. any given hexanucleotide occurs too often to serve as start signal. ➔ How to solve this problem?**

Answer: use combination of two or more short sequences
(eg A and B, both hexanucleotides) ➔

$$P(A \cap B) = 5.9x10^{-8}$$

At $5.9x10^{-8}$ (= occurence every $16.8x10^6$ bp) the propability of finding two defined, short recognition sequences in proximity to each other is <1 occurence in the entire prokaryotic genome, and even in a human genome you'd expect less than 1 occurence if you further enforce that the two elements must have a certain distance from each other



Transcription – How to Find What You Want

H sapiens: average size of protein still requires ~1kb of direct code, **BUT** accounting for introns, the actual average gene size is ~28 **k**bp ➔ **game over….. any given hexanucleotide occurs too often to serve as start signal. ➔ How to solve this problem?**

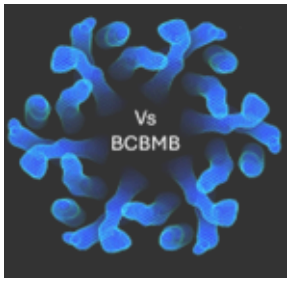
Answer: use combination of two or more short sequences (eg A and B, both hexanucleotides)

$$P(A \cap B) = 5.9x10^{-8}$$

At $5.9x10^{-8}$ (= occurence every $16.8x10^{6}$ bp) the propability of finding two defined, short recognition sequences in proximity to each other is <1 occurence in the entire prokaryotic genome, and even in a human genome you'd expect less than 1 occurence if you further enforce that the two elements must have a certain distance from each other

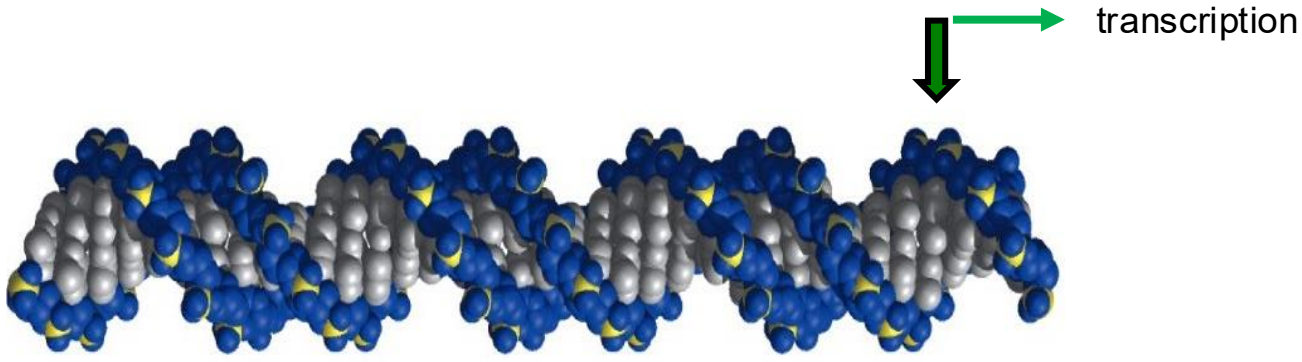
➔ **promotors are "coincidence detectors" (> 1 condition met at once).**

coincidence detection is EVERYWHERE in biology
(eg...protein folding, molecular recognition, enzyme catalysis, molecular signaling, ...
keep watching out for it as we go....!)

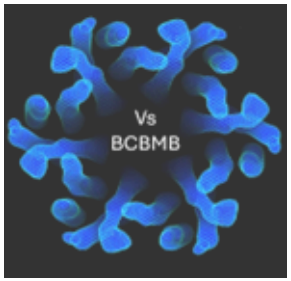


Transcription – How to Find What You Want

→ Now just need to figure out how to place these elements. Assume you want to **start** transcribing at the position of the **green arrow** and proceed to the right ... **how far to the left (called "upstream" of the transcription start site) would you place the first and second elements?**

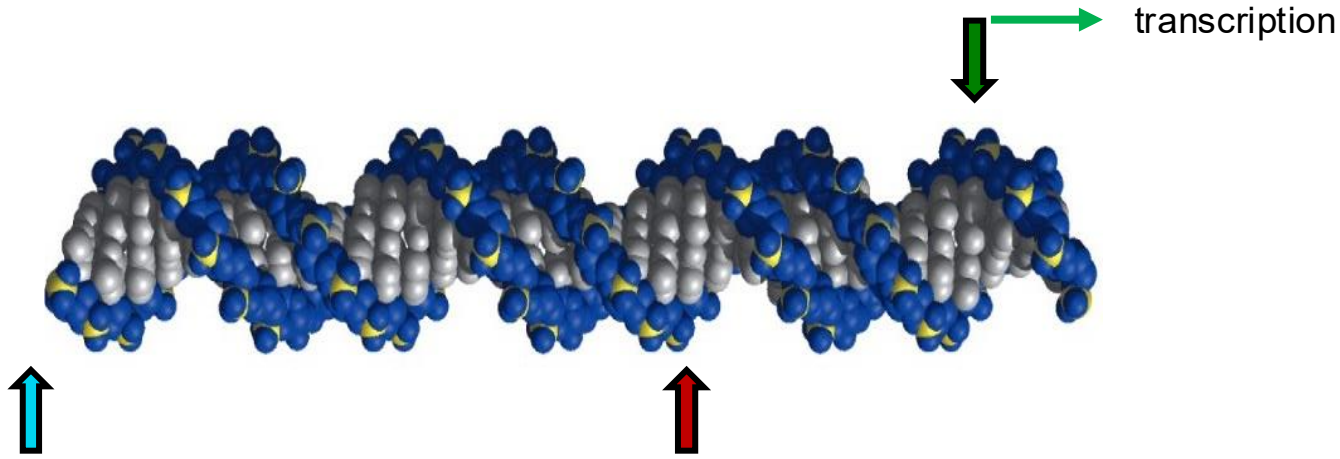


….try to visualize it …
…what does your intuition tell you??...



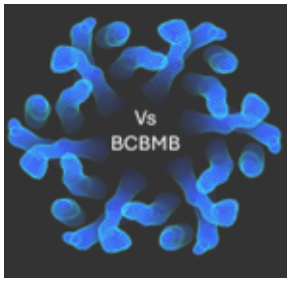
Transcription – How to Find What You Want

→ Now just need to figure out how to place these elements. Assume you want to **start** transcribing at the position of the **green arrow** and proceed to the right ... **how far to the left (called "upstream" of the transcription start site) would you place the first and second elements?**



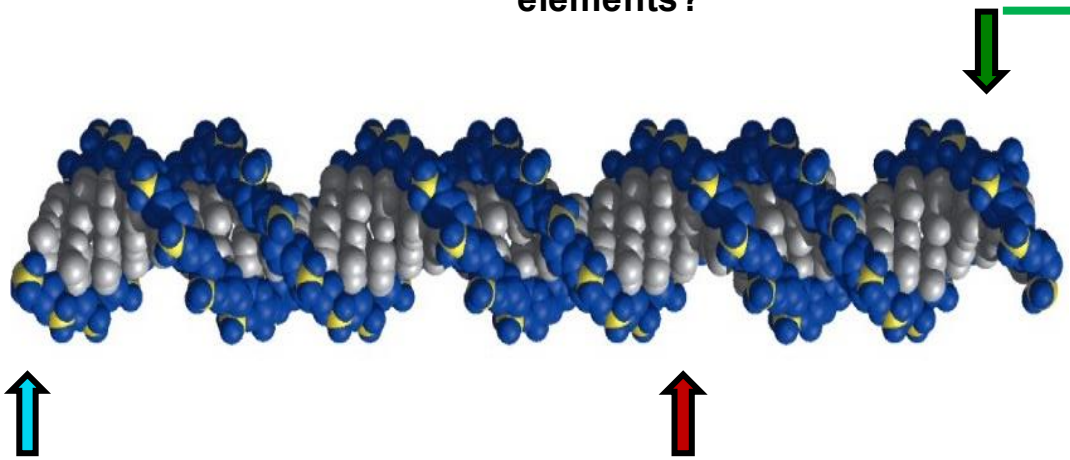
Answer: **~3 tuns (~30-35 bp)** **1 turn (~10bp)** **why?**

Answer: ***...try to justify this design.... (even if you didn't get it quite right yourself)***



Transcription – How to Find What You Want

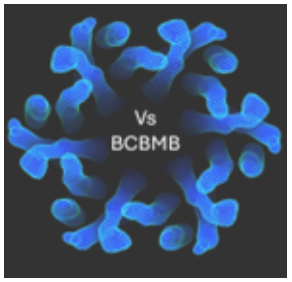
→ Now just need to figure out how to place these elements. Assume you want to **start** transcribing at the position of the **green arrow** and proceed to the right … **how far to the left (called "upstream" of the transcription start site) would you place the first and second elements?**



Answer: ~3 tuns (~30-35 bp) **1 turn (~10bp)** **why?**

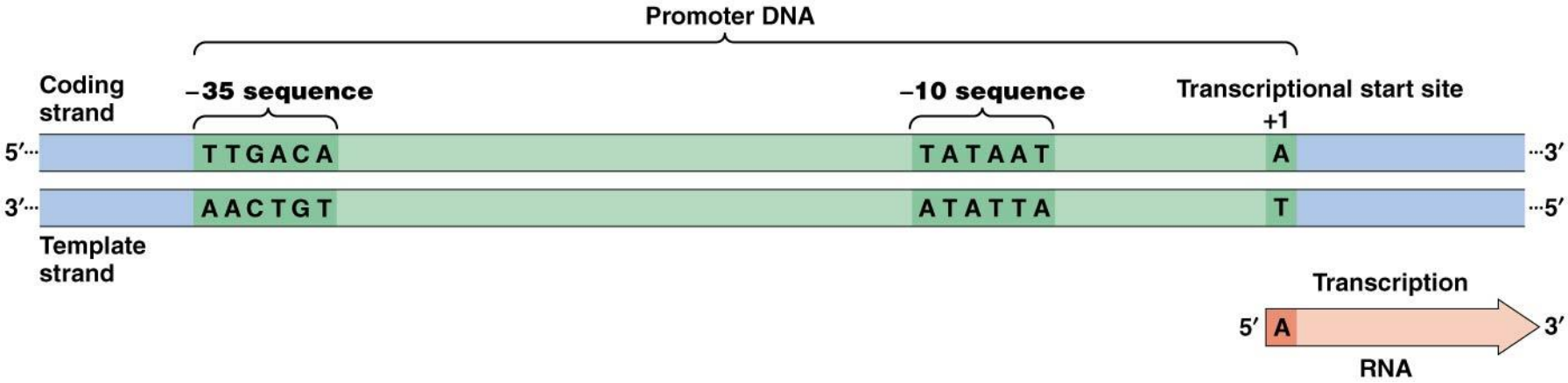
Answer:

- **need to locally unwind the DNA, which opens up a "bubble" to expose the base pairs that serve as "template"** (...opening bubble means strand separation = costs energy = the shorter the segment the better = the first part of the promoter needs to be very close to the start site; more importantly though...opening the bubble causes DNA to get **more tightly** wound on either side of the bubble ... want to be as close as possible to minimize topological stress…(if working through the Chromatin chapter you played with the rope strings as I recommended, you may remember that holding the string at both ends and trying to pry open a bubble by separating the two strands gets very difficult very quickly)
- **Since the transcriptional machinery is made from proteins, you need some space to allow those to engage** (if you go too far away though, it becomes harder to coordinate the DNA:protein interactions)**;** having some space also **increases tolerance for mutations** (because small distortions and mismatches can be compensated by polymer flexibility) **and creates opportunity for regulating the efficiency** (increasing/decreasing efficiency of engagement with transcription machinery by playing with the exact spacing)



Transcription – Canonical Prokaryotic Promoters

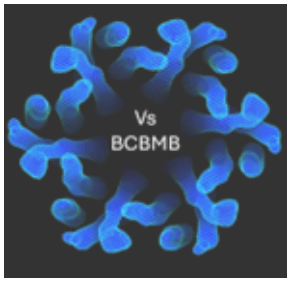
Through the chain of arguments – you basically **deduced the design principles of** what is called a **"core promoter".** What is shown here are the "consensus sequences" = most prokaryotic promoters conform with this, but small variations and certain mutations are permissible



- Acknowledging that synthesis proceeds from 5' → 3', the **strand that contains the TATAAT sequence of a gene's promoter is called coding** strand (because the transcript will have the same sequence except that T is replaced by U); the opposing **complementary strand is called template strand**
- the **transcription start site** is numbered +1 and **typically is an "A"** (not to be confused with the "A" in the ATG translation start codon)
- **sequences 5'** of the transcription start site are called "**upstream**", and assigned negative numbers
- **sequences 3'** of the startside are called "**downstream**", and are assigned positive numbers

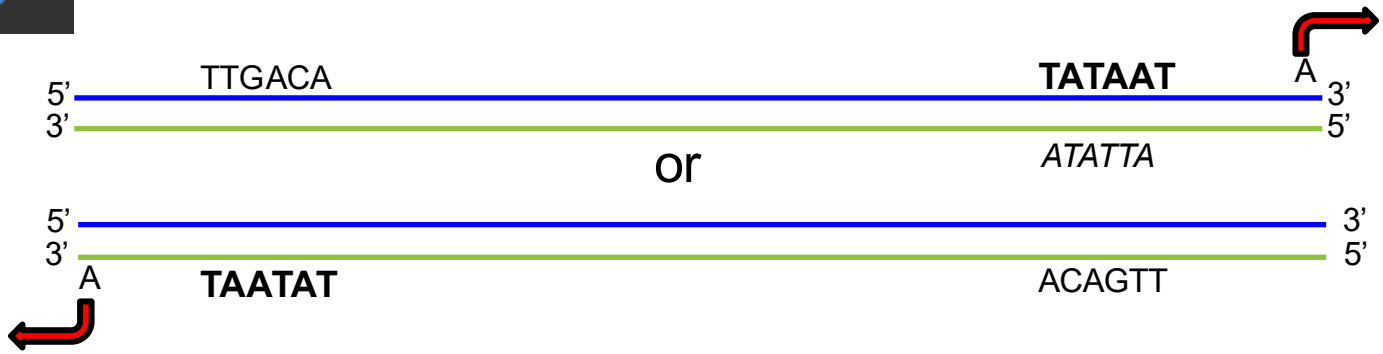
The design of eukaryotic promoters is more complicated; they have additional elements to accommodate a more complex transcription machinery (since this is an introduction to it all, we will not cover it in detail)

Thinking through **design principles of promoters** – you may realize that we **also answered** how the machinery knows **which strand to use – why?**

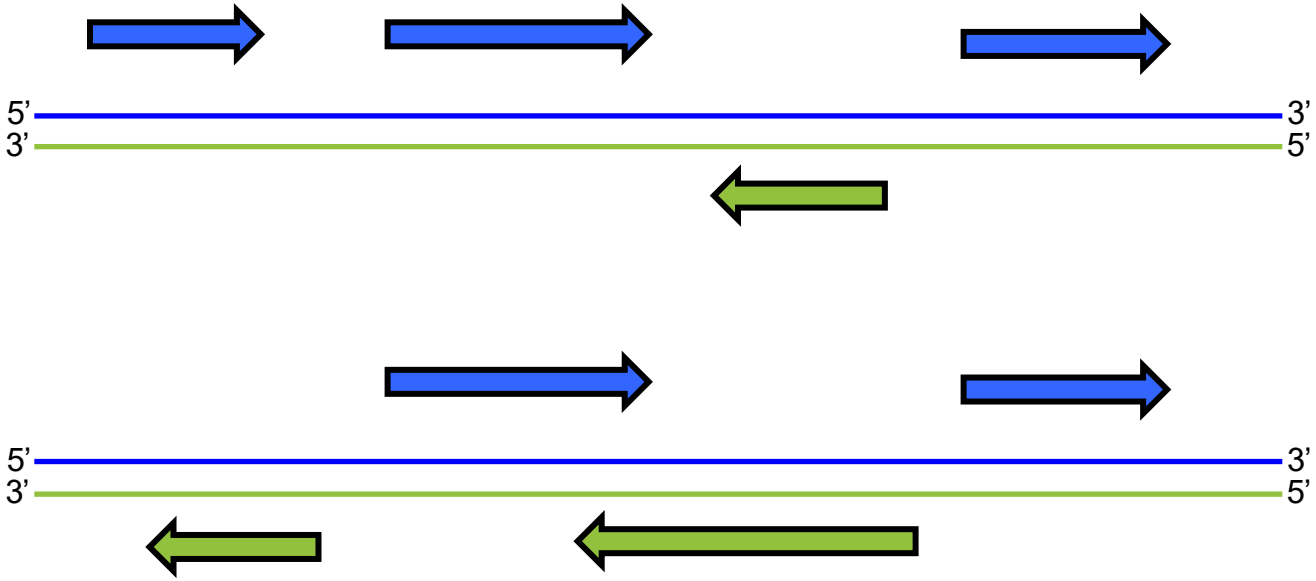


Transcription – Coding vs Template Strand

Answer: asymmetry – the ability to distinguish between "TATAAT" and "ATATTA" (for instance) allows to establish the relative placement of the -35 and -10 elements with respect to each other. This inherent directionality of the promoter **positions and orients** the machinery that will synthesize the RNA.



Take Note: the direction of transcription depends **only** on how the promotor elements are distributed on the two strands. Either strand can equally well serve as coding or template strand. There even are instances where genes overlap (either on the same or opposite strand).



each arrow indicates a transcription start site

Transcription – The Cartoon

Now that we understand – in principle – what a promoter is, lets scrutinize a summary picture of the general sequence of events:

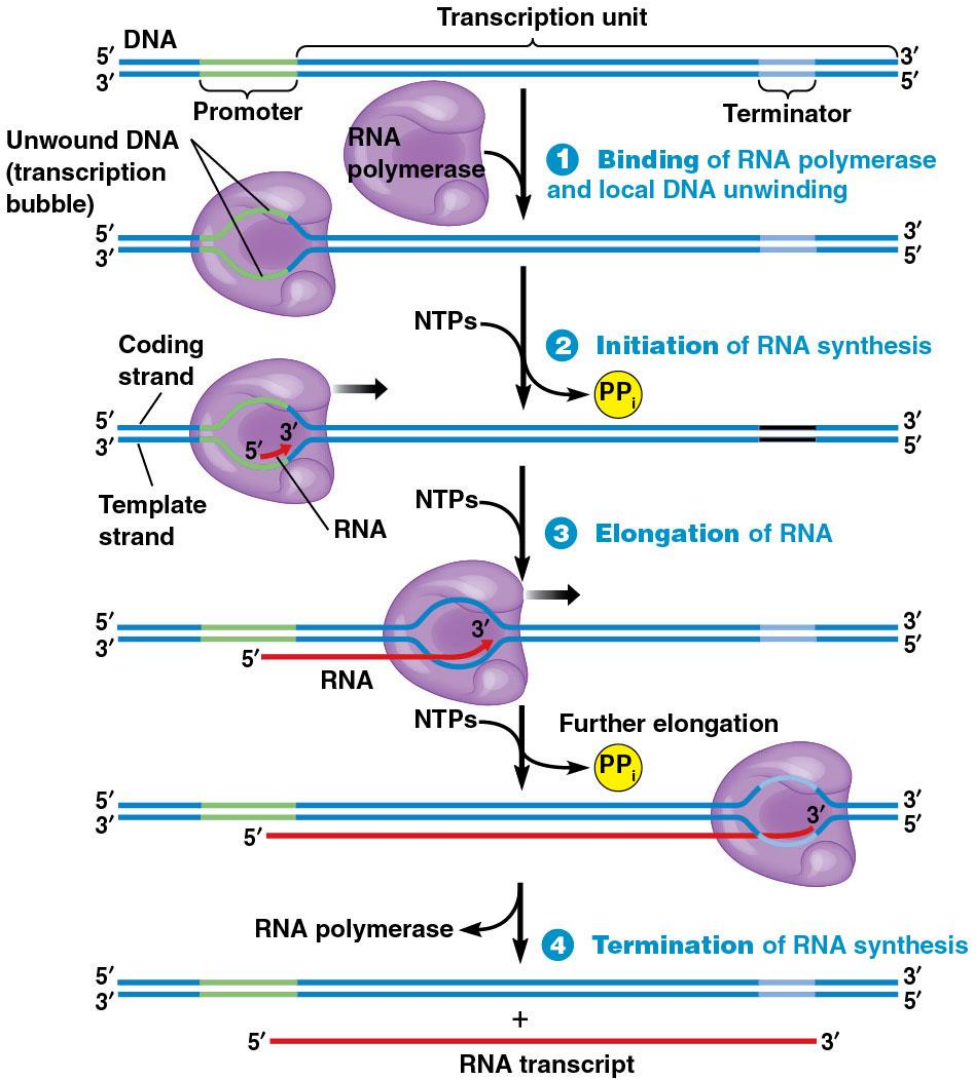
1/ Engagement with the transcription machinery
→Opens transcription bubble by partial unwinding of DNA

2/ Initiation of synthesis (primer generation)

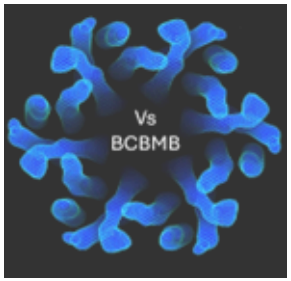
3/ Promoter "escape" and initial elongation

4/ Elongation

5/ Termination and release of machinery + transcript



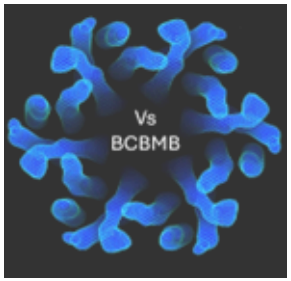
NOTE: cartoons are helpful but can be misleading .... in this case ...cartoon suggests that RNApol is moving along a stationary transcript → true in prokaryotes, but in eukaryotic transcription factory, RNApol is immobilized ➔ the DNA is reeled in.



Transcription – RNA Polymerase and Initiation

With the summary of events in mind – lets next turn attention to the machinery that is involved. Specifically, we want to ask – **is transcription accomplished by a single protein?** (as the cartoon suggests)

....*what are your thoughts?....*



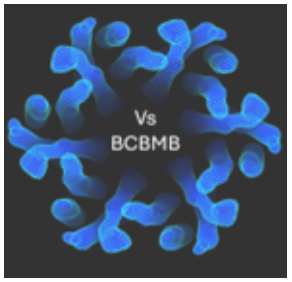
Transcription – RNA Polymerase and Initiation

With the summary of events in mind – lets next turn attention to the machinery that is involved. Specifically, we want to ask – **is transcription accomplished by a single protein?**

Answer: no.
This should not come as surprise, given that initiating and carrying out transcription involves several functions: (a) recognize promoter element**s**, (b) unwind DNA, (c) synthesize RNA, (d) possibly participate in termination in some specific way.

→ ...why can't that all be done by one multidomain protein?
Answer: well, in principle it could ....but remember that it all **started with a system that was "still learning how to do this". At that stage, exploiting quaternary structure was more likely** than the emergence of one enormous multidomain protein that could do it all.

Why?



Transcription – RNA Polymerase and Initiation

With the summary of events in mind – lets next turn attention to the machinery that is involved. Specifically, we want to ask – **is transcription accomplished by a single protein?**

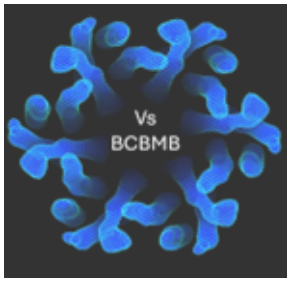
Answer: no.
This should not come as surprise, given that initiating and carrying out transcription involves several functions: (a) recognize promoter element**s**, (b) unwind DNA, (c) synthesize RNA, (d) possibly participate in termination in some specific way.

→ ...why can't that all be done by one multidomain protein?
Answer: well, in principle it could ....but remember that it all **started with a system that was "still learning how to do this". At that stage, exploiting quaternary structure was more likely** than the emergence of one enormous multidomain protein that could do it all.

Why?
To go it all in one step you would need: efficient protein synthesis, have to navigate folding of a very long polypeptide + all the different functions need to be fully operational right away + communicate properly with each other = unlikely;
more likely to go "piecemeal" ...have something than can regonize the elements, something that knows how to "melt DNA", something that has a crude catalytic ability to synthesize.... (multiple pieces also allow for regulation more easily than if it all were on a single polypeptide)

With this in mind ... we want to look at the "simplest" solution

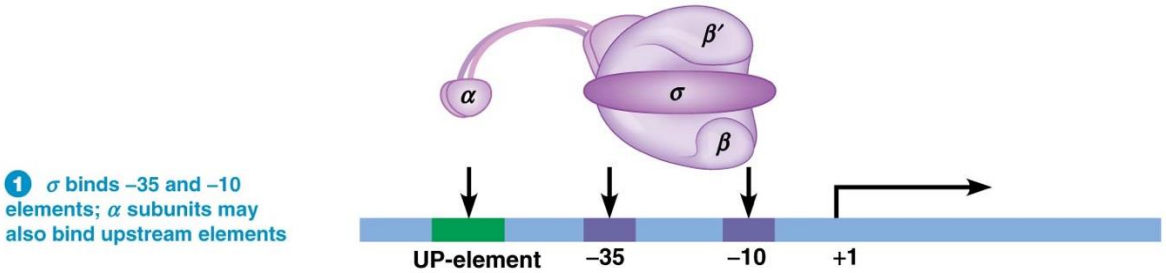
Transcription in Prokaryotes



Transcription – RNA Polymerase and Initiation

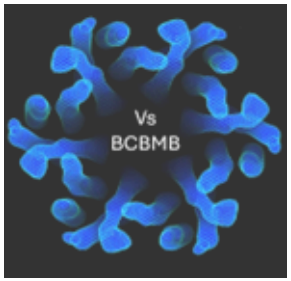
With this in mind ... we want to look at the "simplest" solution

Transcription in Prokaryotes



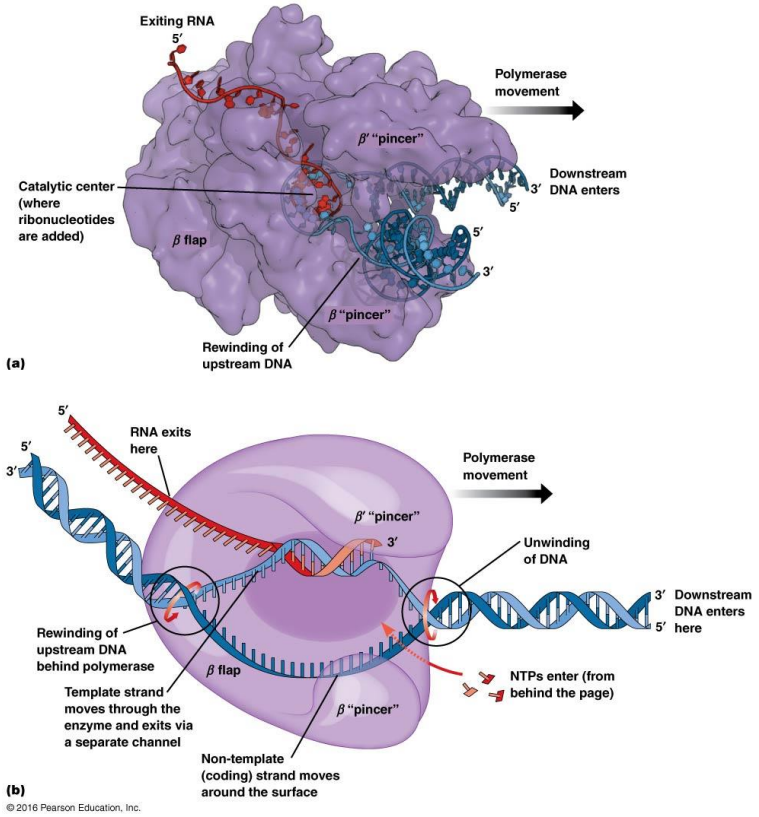
The cartoon illustrates the general topology of the **prokarytic transcription initiation complex**: σ-factor, and RNA polymerase, which consists of 4 polypeptide subunits: $\alpha_2\beta\beta'$

The σ-factor is responsible for recognition of the -10/-35 sequences and is released after the complex breaks free from the promoter. Similarly, another protein factor, called “rho-factor (ρ-factor)” is needed for termination of transcription at some loci (many genes can terminate transcription independent of that though).



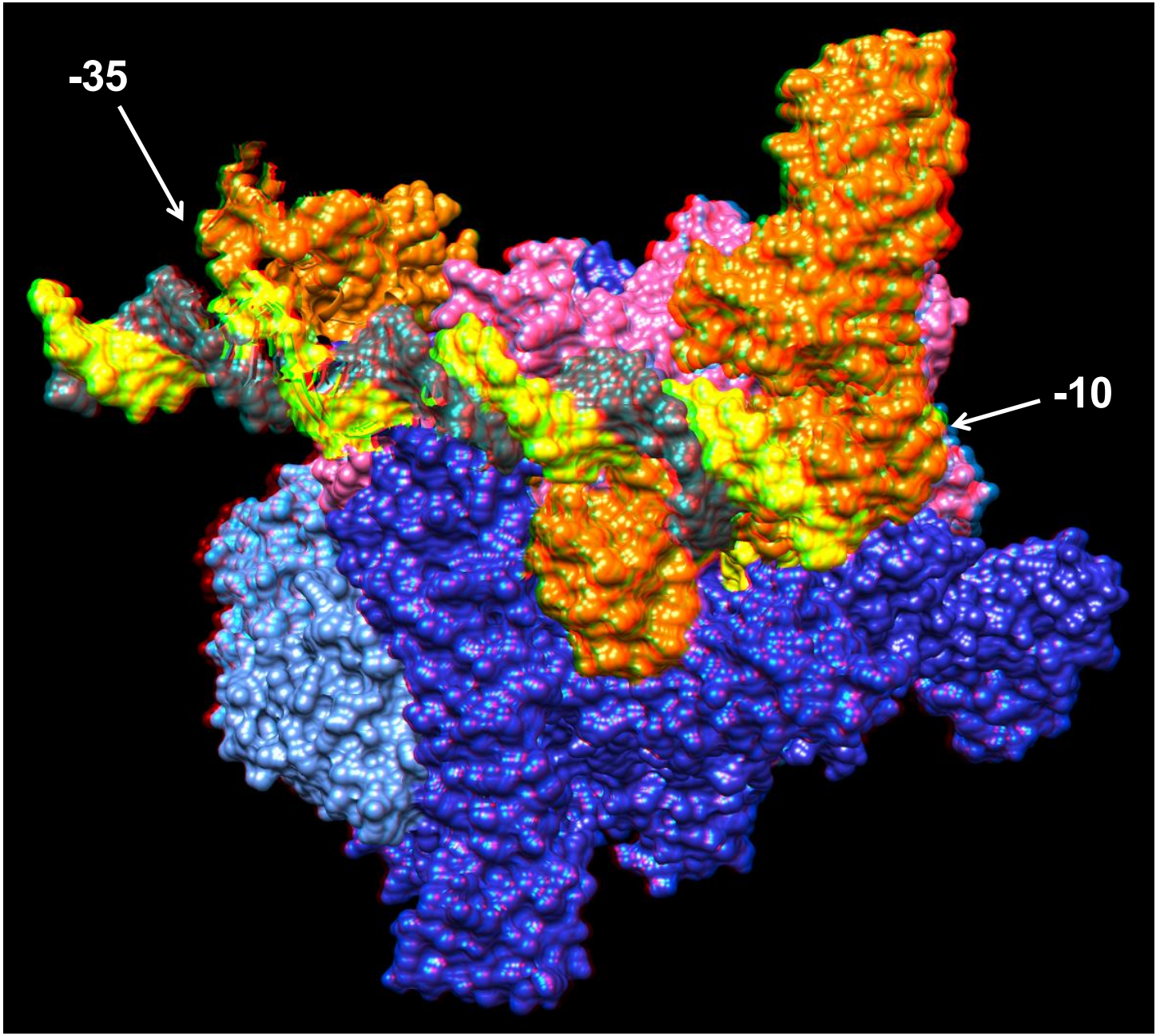
Transcription – From Cartoon to Reality….

To give you just a glimpse of how complex this machinerie is in reality.....lets do a side-by-side comparison of the RNA-polymerase only (left figure) and in complex with the σ-factor (PDB: 6k4y)

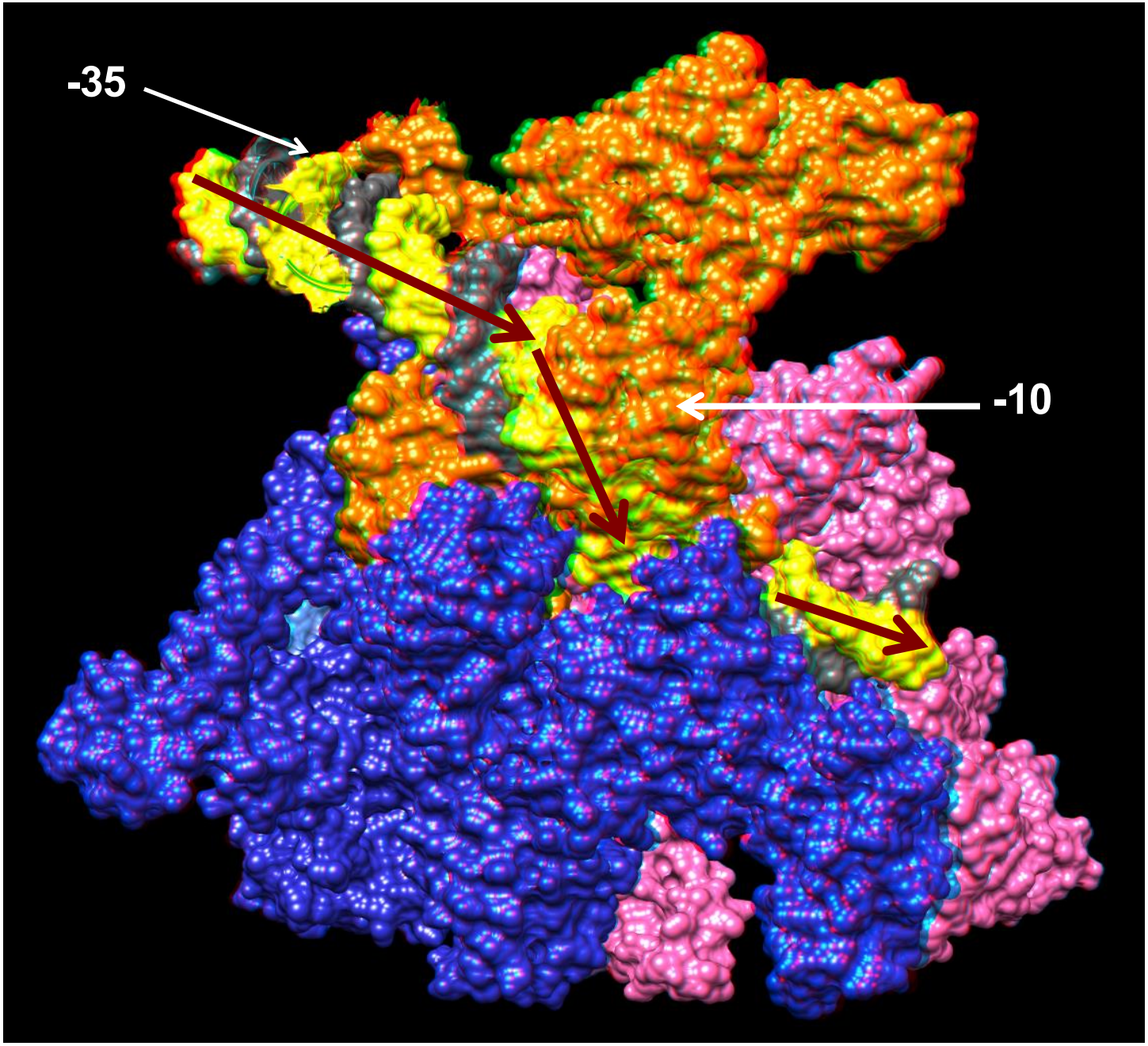


The **stereo ribbon diagram** to the right is for viewing with "red-cyan glasses" (cross-eyed versions are shown after that) Colors have been attributed as follows: yellow and grey for the two DNA strands; orange for the σ-factor, light blue for the α_2 subunits of RNA polymerase, and pink/dark blue for the ββ′ subunits. Looking at the complexity of it...you may appreciate better why this has not evolved as **one** polypeptide .... – to give you an even better sense for the "complementarity" of the structures – two space filling models are shown on the next two slides

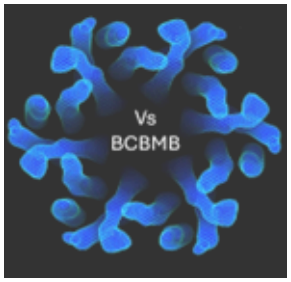
Transcription – Complementarity in Action….(red-cyan stereo image)



Transcription – Complementarity in Action….(red-cyan stereo image)

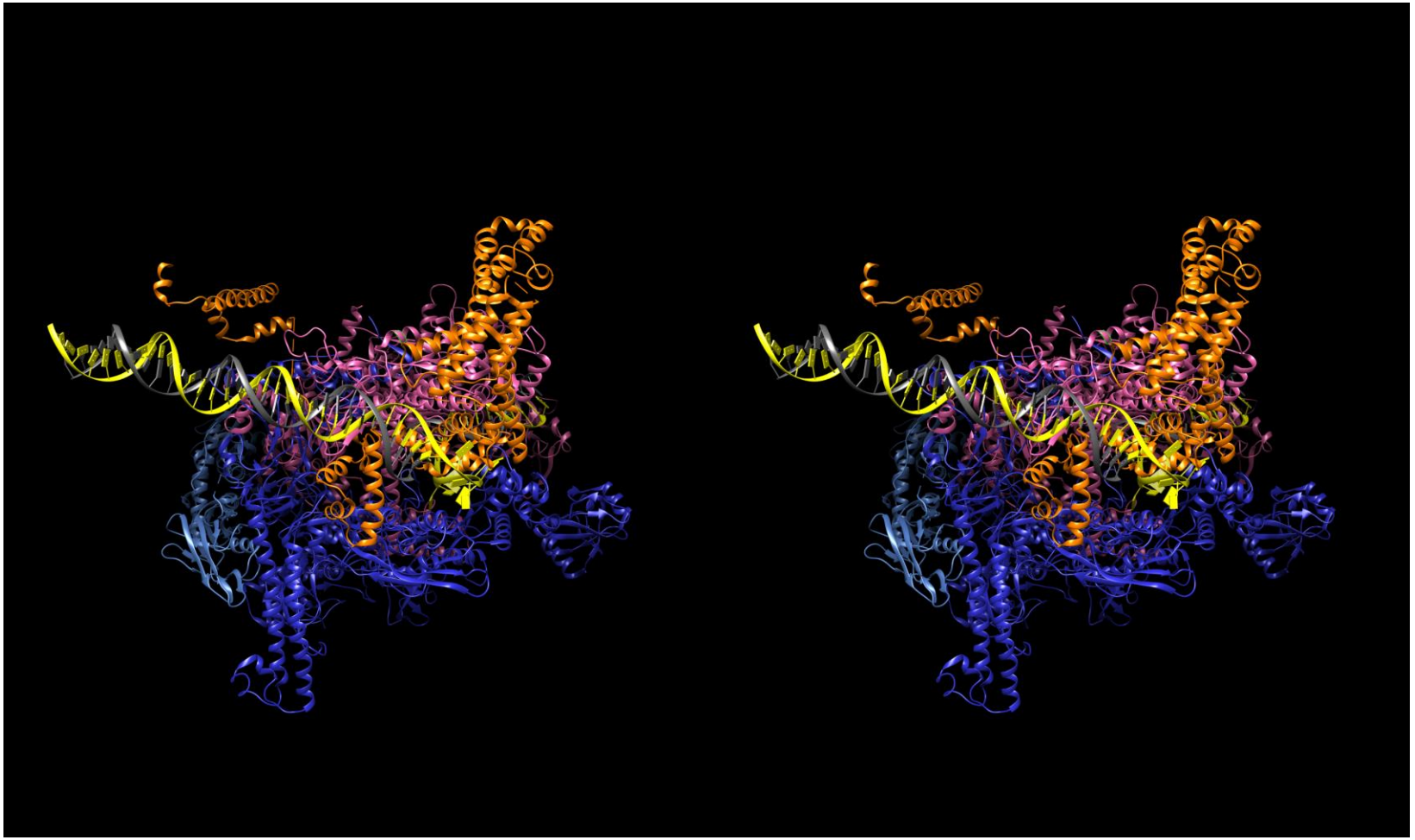


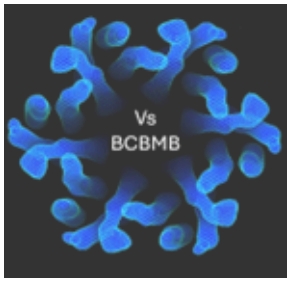
Note the bent path (black arrows) that DNA takes through the catalytically active site of RNA polymerase



Transcription – From Cartoon to Reality….

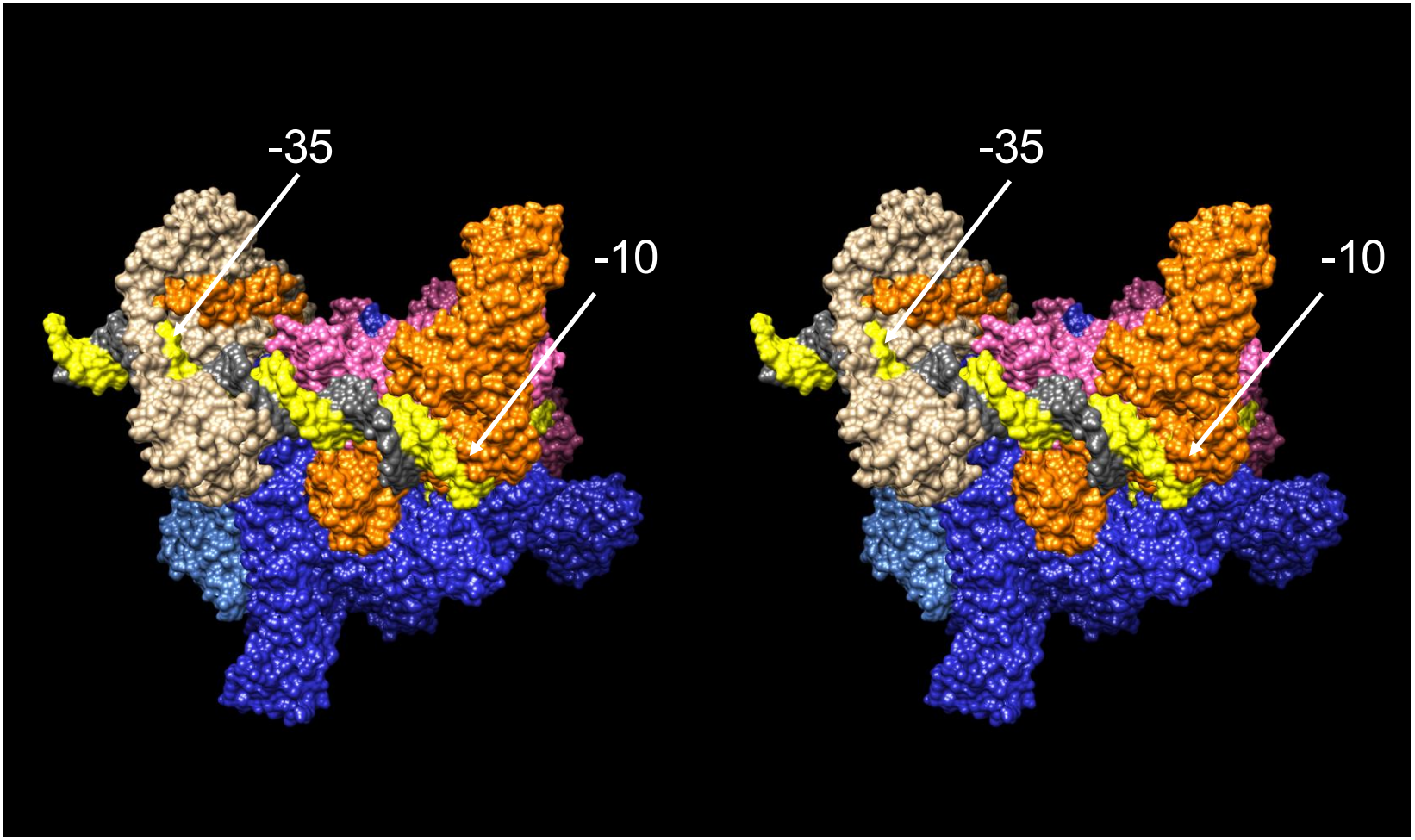
here are the cross-eyed versions

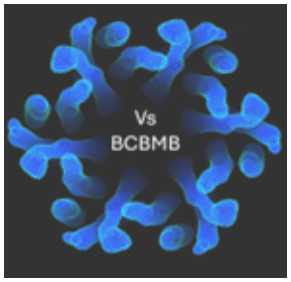




Transcription – From Cartoon to Reality….

here are the cross-eyed versions

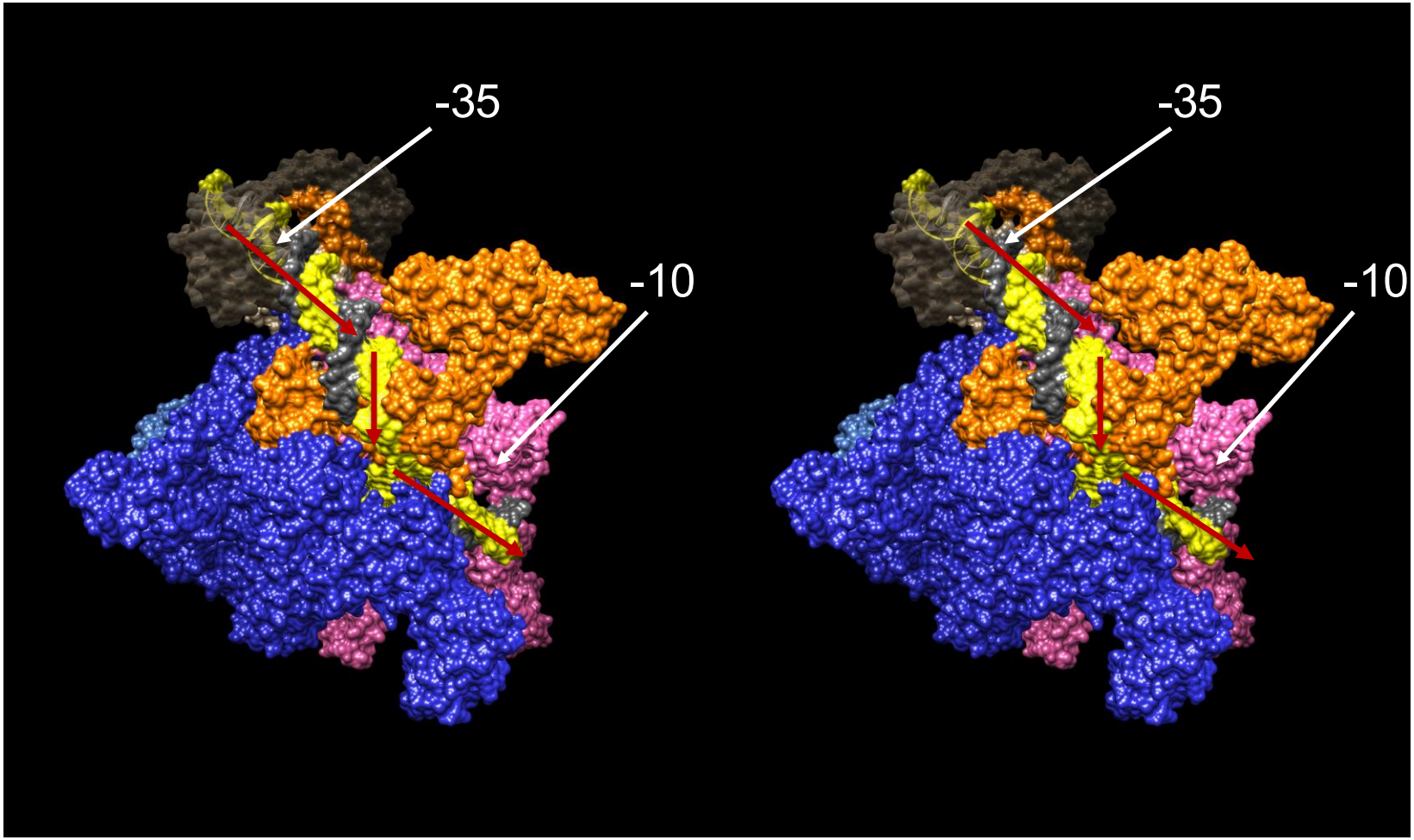


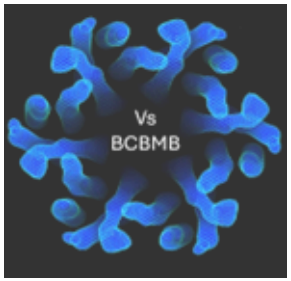


Transcription – From Cartoon to Reality….

here are the cross-eyed versions

Take note of the "criss-cross"/bent path the DNA takes when it is engaged to the initiation complex



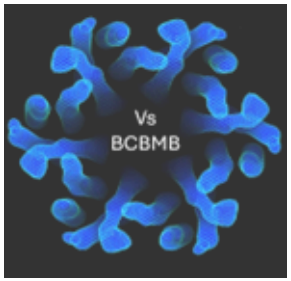


Transcription – When Complicated is Not Complex Enough

If the RNA polymerase of prokaryotes looks intimidating, then eukaryotic transcription will be scary.

Why?

…what are your thoughts?....



Transcription – When Complicated is Not Complex Enough

If the RNA polymerase of prokaryotes looks intimidating, then eukaryotic transcription will be scary.

Why?

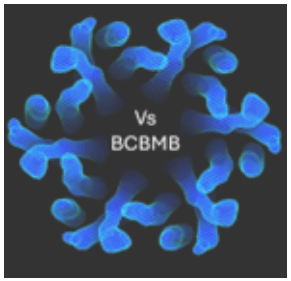
…what are your thoughts?....

Answer

far larger number of proteins is involved in initiation. More specifically - **eukaryotic RNA polymerases require other protein complexes, called "Transcription Factors (TFs)", to bind to DNA and to initiate transcription.**

The need for additional components arises from factors like:

- **differences in DNA structural organization** ("naked" in prokaryotes vs nucleosome based chromatin in eukaryotes; chemical modification [acetylation] of histones allows temporary displacement or removal)
- **promoter design** (number of elements in promotor) and **spatial characteristics of the promoter**
- need to interact with other **regulatory proteins**
- **number and properties of different RNA polymerases** (prokaryotes have only 1, eukaryotes have 3 that transcribe different targets (RNA Pol I: ribosomal RNA; Pol II: messenger RNAs, regulatory RNAs; Pol III: transfer RNAs)



Transcription – When Complicated is Not Complex Enough

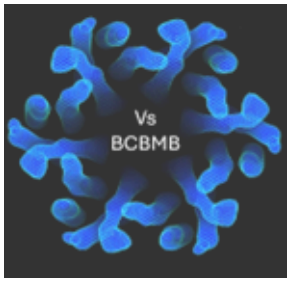
Adding to the need for more protein components
the individual components themselves are more complex too.…

For instance: **RNA-polymerases** – **each** of the three eukaryotic polymerases is larger and structurally more complex than the prokaryotic enzyme. Case in point: RNA-polymerase II is made from 12 subunits (compare: 4 in prokaryotes).

the most important conclusion here:
→ many of the components do no longer interact with DNA (or the emerging RNA) directly, but mediate
protein:protein interactions that reflect the shift from mere transcription to the hugely more complex regulatory aspects of this process in eukaryotes compared to prokaryotes

What is Likely to be the Biggest Consequence of This Increase in Complexity For the Logistics of the Process?

…try to come up with an idea….



Transcription – When Complicated is Not Complex Enough

Adding to the need for more protein components
some components themselves are more complex too….

For instance: **RNA-polymerases** – **each** of the three eukaryotic polymerases is larger and structurally more complex than the prokaryotic enzyme. Case in point: RNA-polymerase II is made from 12 subunits (compare: 4 in prokaryotes).

the most important conclusion here:
→ many of the components do no longer interact with DNA (or the emerging RNA) directly, but mediate
pure protein:protein interactions that reflect the shift from mere transcription to the hugely more complex regulatory aspects of this process in eukaryotes compared to prokaryotes

What is Likely to be the Biggest Consequence of This Increase in Complexity for the Logistics of the Process?

The larger size of the enzymes and requirements for accessory proteins like transcription factors result in a
strictly sequential process
that first assembles a so-called "preinitiation complex" on the promoter (~50-100 proteins).

Once assembled, transcription is enabled/"switched on" through the binding of a last piece: TFII-H. This factor has two functions: it can unwind DNA (= a helicase) and it can activate RNA polymerase II through chemical modifcations that allow the enzyme to escape from the preinitiation complex.
A summary of the entire process is shown on the next slide.

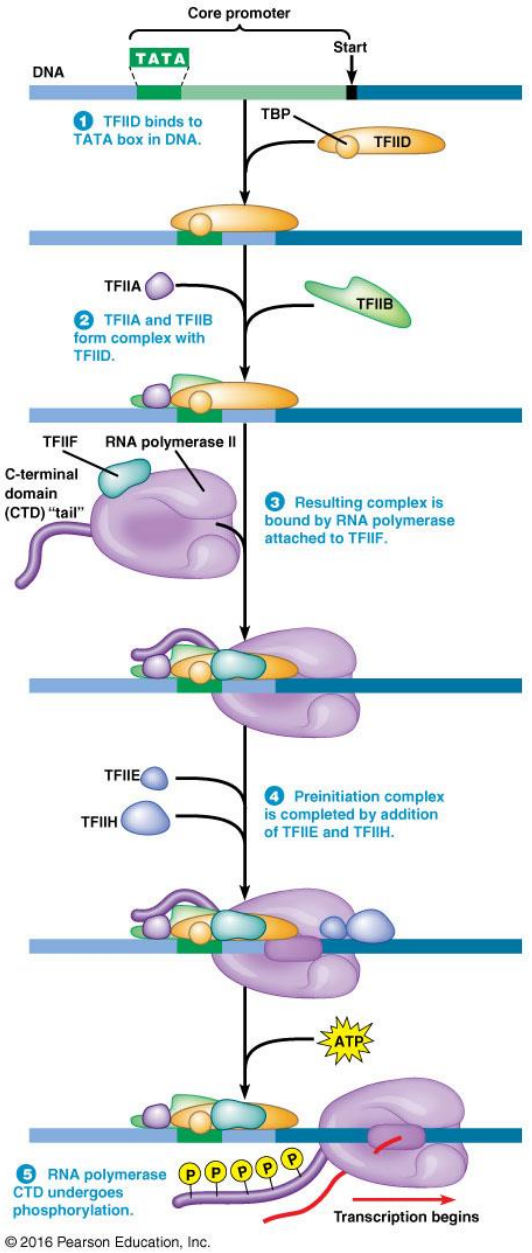
Transcription – When Complicated is Not Complex Enough

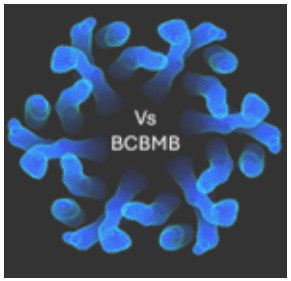
Looking at this ….**don't** panic ....nobody in their right mind would/should expect you to memorize this scheme. The sole point here is to give you a basic appreciation how complicated initation of transcription is in eukaryotes.

Why would I want you to acutely take note of this?

Answer

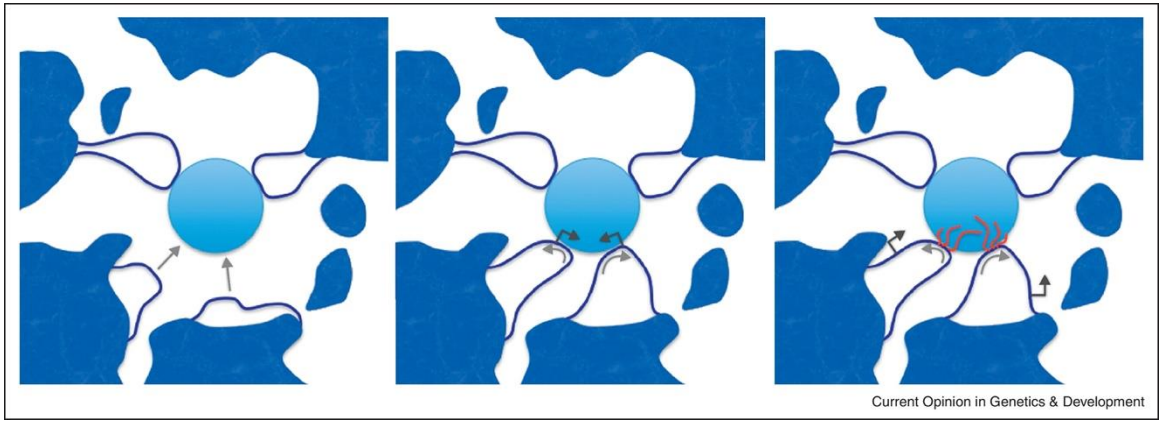
because the detail makes it easier to appreciate why eukaryotes build "transcription factories" = specialized compartments within the nucleus that carry out most of the transcriptional activities. A simplified version of what a transcription factory is is shown on the next slide





Transcription – When Complicated is Not Complex Enough

Answer:
A simplified version of what a transcription factory (blue circle in center) is is shown below:

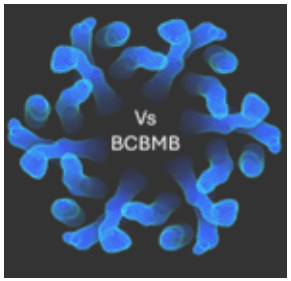


Edelman LB, and Fraser P (2012) Curr Opin Gen&Dev 22(2):110-114

tying things back into the big questions we had at the onset…

note how in this model, euchromatin DNA loops are "reeled in" (=actively transported to the factory = "it (the gene) finds you (the factory)") **and slides "past" an immobilized RNA-polymerase complex** (located somewhere on the surface of the blue circle)**.**

Having several complexes immobilized in close vicinity triggers multiple initiation events = multiple transcripts



Some Trivia Before Moving On

Having dealt with the basics of transcription …. a few more simple questions remain

once it gets going – how fast is it going?

Prokaryotes … 40-100 nt/s (less than a minute for a typical transcript)

Eukaryotes…much slower @ 20-40 nt/s

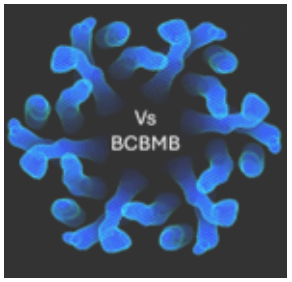
The reason will become obvious in a moment…(third question)

how accurate is the process?

...take a guess by completing the two gaps!

RNA polymerase makes ____ error(s) in any _____ nucleotides it adds to the transcript.

(ideally...how many errors would you want? Would that really be necessary? If not, why not? And in this case, what would be a good compromise between speed-accuracy)



Some Trivia Before Moving On

Having dealt with the basics of transcription …. a few more simple questions remain

once it gets going – how fast is it going?

Prokaryotes … 40-100 nt/s (less than a minute for a typical transcript)
Eukaryotes…much slower @ 20-40 nt/s
The reason will become obvious in a moment…(third question)

how accurate is the process?

1 error for every 10,000-100,000 nucleotides added (how close was your guess?)

If this sounds impressive to you … it isn't. In fact, this accuracy is LOUSY compared to DNA (better than 1 in a billion) because RNA-polymerases lack what is called "proofreading activity" = RNA polymerases do not check if they added the right nucleotide based on the template … most of the time they do because wrong nucleotides will not stick around long enough to be incorporated, but when they do stay long enough, RNA polymerase just adds them without checking…

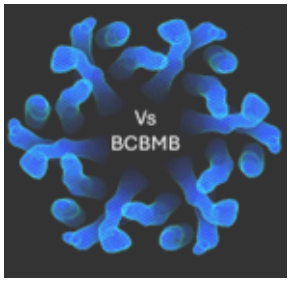
this seems odd …but works for a number of reasons: (1) RNA molecules are relatively short lived; (2) just as for proteins, RNA function is unlikely to be badly impacted by a single mistake (covered in more detail in the Advanced Biochemistry chapters), and (3) even if the error destroys that RNA molecule's function, it only affects a single molecule (out of hundreds or thousands that are made).
Conclusion: "quick and dirty" works just fine in this case.

how does the eukaryotic machinery deals with the nucleosomes?

nucleosomes are "roadblocks" and need to be displaced to allow RNA polymerase to pass
this displacement involves chemical modifications, sliding and/or physical removal of nucleosomes, and reconstitution of chromatin structure behind the RNA polymerase complex → explains the slower rate if transcription in eukaryotes

and lastly … are all transcripts functionally the same?

No … different types. Most common ribosomal RNA (rRNA), transfer RNA (tRNA), and messenger RNA (mRNA)
and then there is a **myriad of other non-coding RNA's** (small nuclear, small nucleolar, small interfering, small non-coding, long non-coding, micro, piwi, cajal body specific, extracellular RNA's) with regulatory, RNA-processing, or unknown functions



Transcription – The Menace of Introns and Other Things

In all organisms: **rRNA and tRNA transcripts need further processing** to give mature products. For more details on that see Advanced Biochemistry Chapters

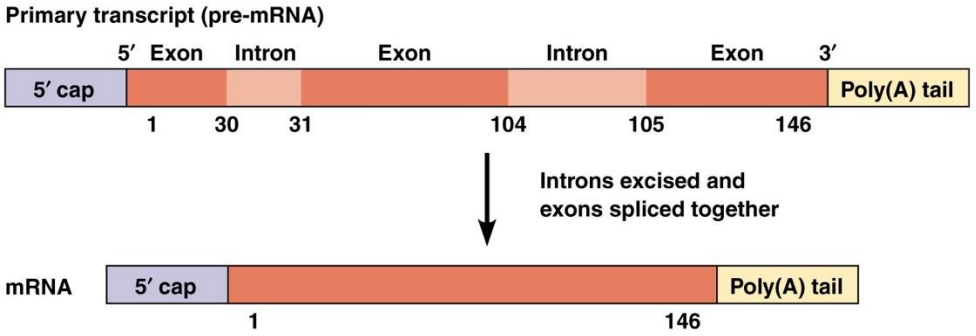
In stark contrast:
mRNAs are ready to go in prokaryotes (in fact translation happens cotranscriptionally)

BUT

raw mRNAs in eukaryotes (**called primary transcript, or pre-RNAs) require extensive additional processing steps to**

- **remove non-coding regions –** called introns. Introns interrupt the coding regions that are called exons
- **chemically modify the 5' end of the mRNA** (= capping) → increases stability of mRNA
- **chemically modify the 3' end of the mRNA** by adding a "polyA-tail" (polyadenylation) → increases mRNA stability, aids export from nucleus, helps ribosomes engage messages.

The only process that we want to look at in this basic chapter is removal of the introns by a process that is called **"mRNA-splicing".** Shown below is the summary for the process: input → output



Position numbers
(1,30, 31, …)
refer to amino acid number in sequence of protein
not bp in the transcript

Pre-mRNA Splicing

If you felt that the sequential assembly of a eukaryotic transcription preinitiation complex looked somewhat complicated, then pre-mRNA splicing is even more complex…..here is the cartoon:

That does not look too bad … in fact...may look easier than formation of the preinitiation complex because the logistics seem intuitive:

- bind 5' border + 3' border of intron
- bring together
- cut intron out & join exons
- release and deposit an "exon junction complex" to singal that this has been processed

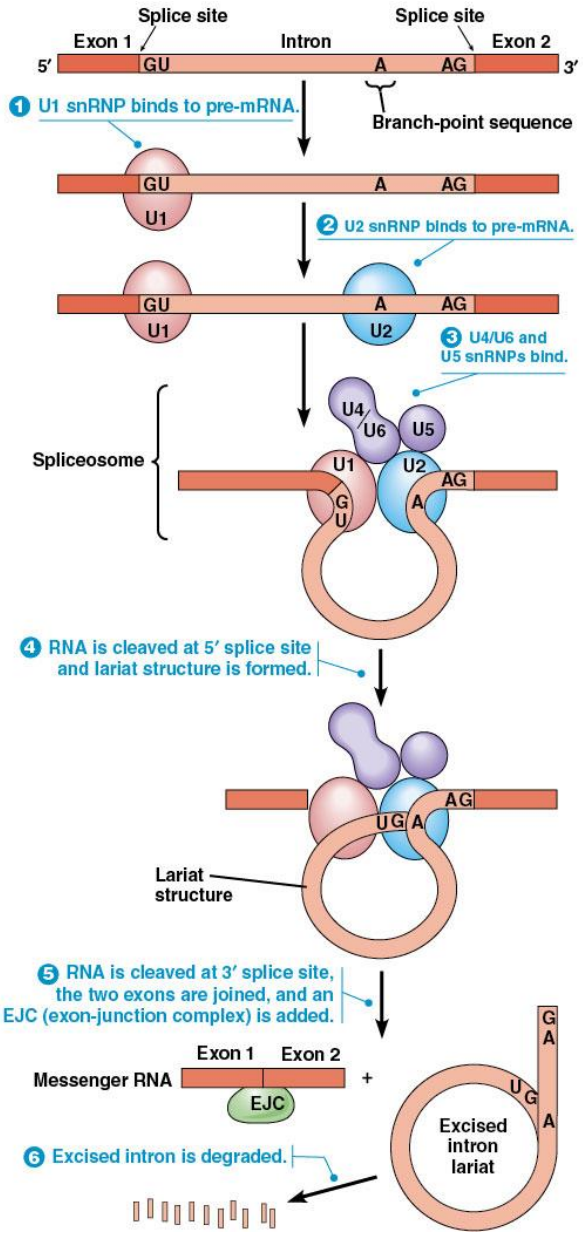
BUT

This simplicity is gone the instant you take a closer look at the "recipes" for making U1, U2, ….. (next slide … to avoid you "shutting down")

and consider that beyond those players additional proteins are needed ....

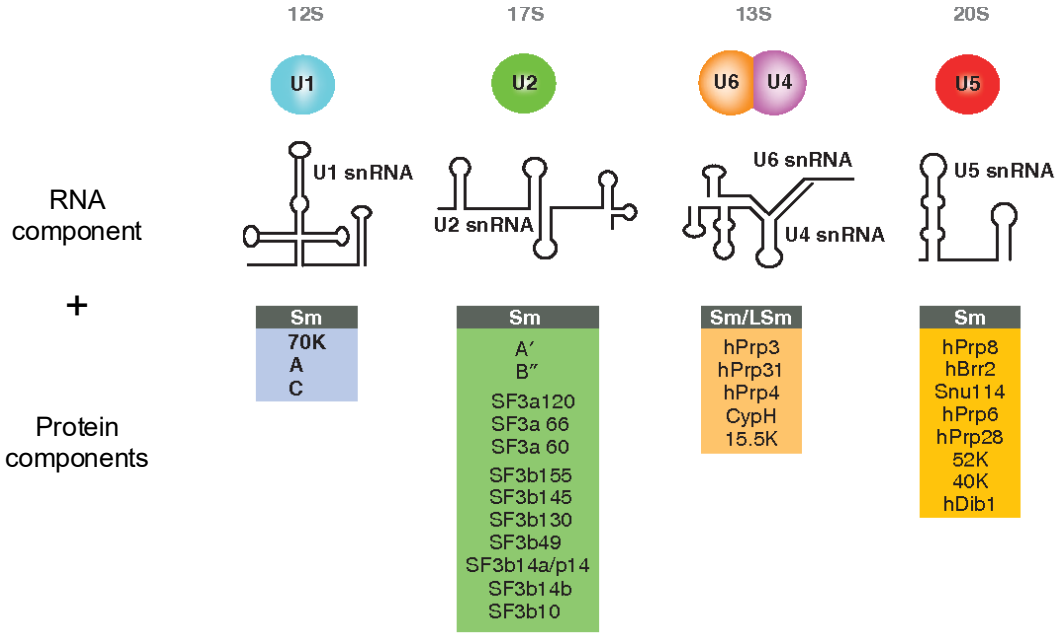
for a **total of ~175-200 proteins that associate with the "spliceosome" at some point during the process**

...this is INSANE!



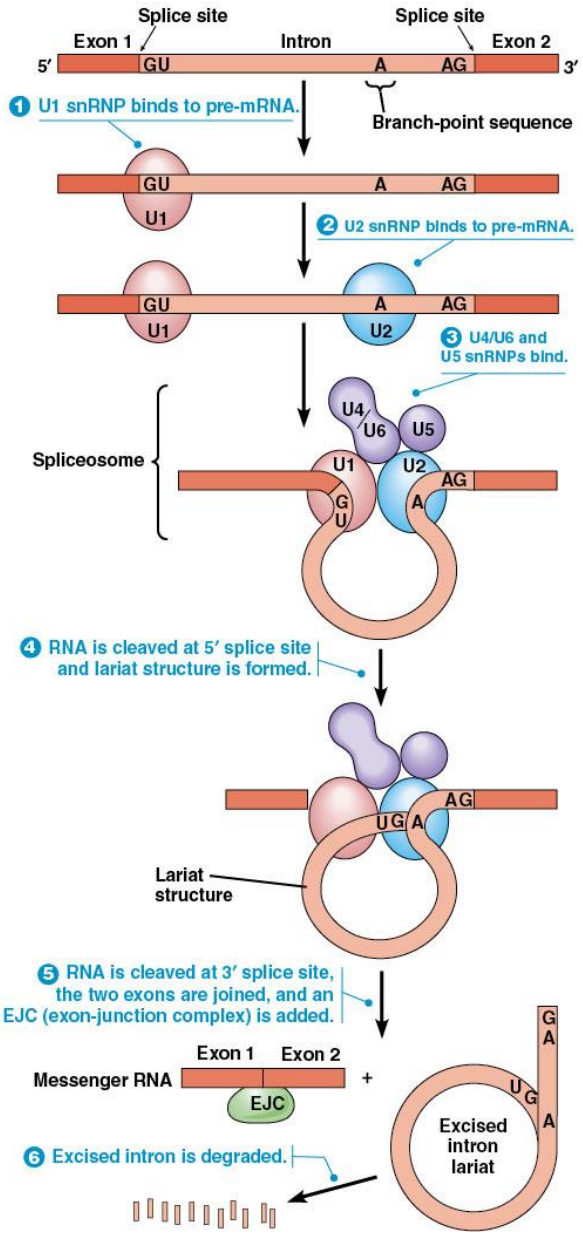
Pre-mRNA Splicing

Recipes for making U1, U2, U5, U4/U6



Will, CL and Lührmann, R (2011) Cold Spring Harb Perspect Biol. Jul 1;3(7).

Take note how each of these so called
small nuclear ribonucleoproteins (snRNPs, or "snurps")
contains a short snRNA that serves as scaffold to assemble the proteins, helps aligning the mRNA substrate and actively participates in the splicing reaction



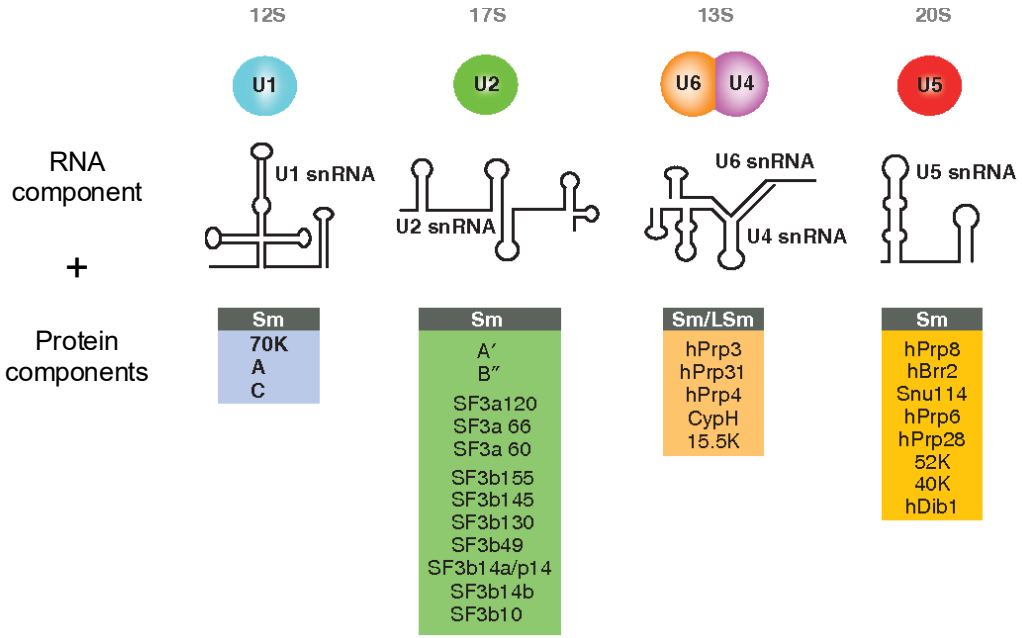
Pre-mRNA Splicing

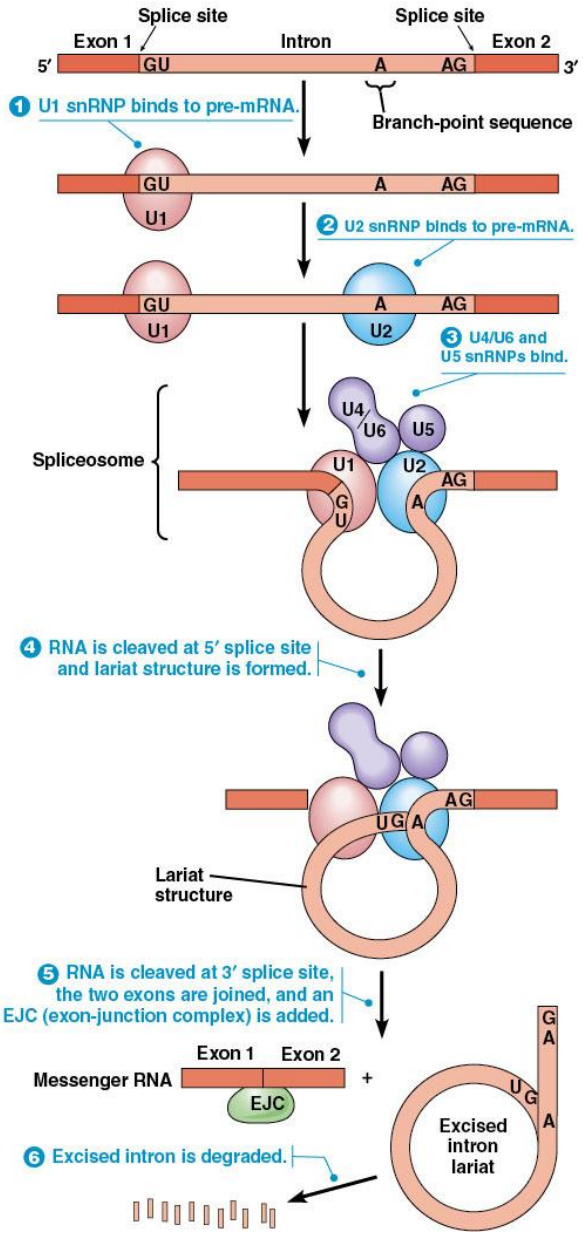
....and one more time: a **total of ~175-200 proteins that associate with the spliceosome at some point during the process...**

There is no expectation that you retain any of the details ...this ALL just goes to illustrate just how altogehter unlikely "life" is ....

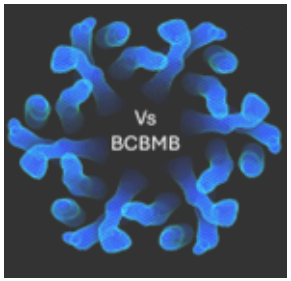
How did ALL of this come about?
How does ALL of this work reproducibly? ...
How can ALL of this be controlled? .....

Biology is AMAZING … finding answers is solving an impossibly complicated puzzle where you don't even know all the pieces yet!
Who wouldn't want to study this? (well maybe … you … in this case, apologies for all of this)





Will, CL and Lührmann, R (2011) Cold Spring Harb Perspect Biol. Jul 1;3(7).



Pre-mRNA Splicing – Why?

why would Nature come up with and sustain such an insanely complicated process?

(a primer on that was already given on slide 37 of "GENOMES" chapter …. here is a little more detail on the aspects that are not related to making coding regions less susceptible to random mutations).

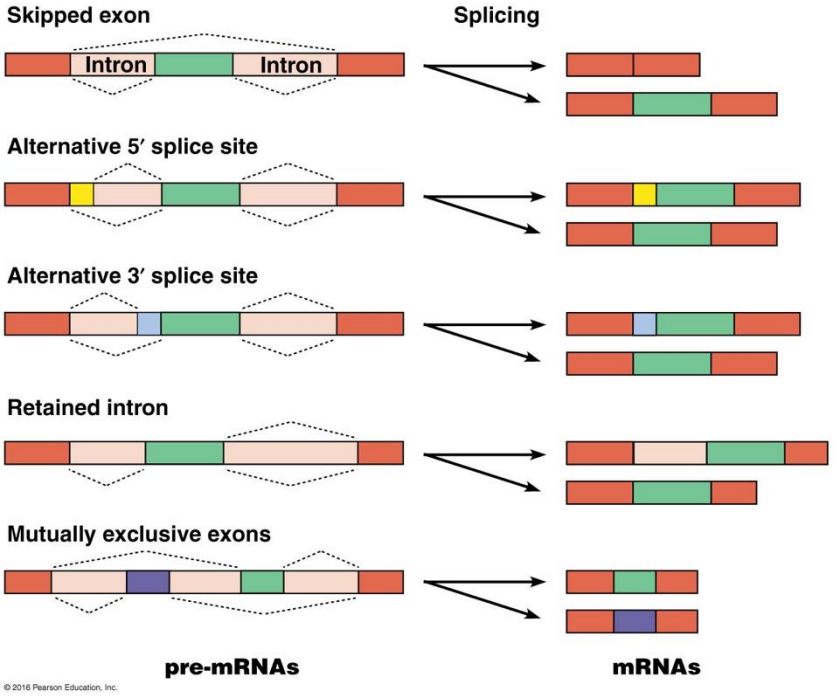
Answer: in contrast to the structural insanity of the molecular mechanism for intron removal …introns **are really** useful for a number of reasons:

- At genomic level, introns serve regulatory roles in gene expression (more detail in Regulation of Gene Expression Chapter)
- After excision, further processing of some introns leads to small non-coding RNAs that regulate gene expression at both the transcriptional as well as translational level.
- **Best understood and wildly important**: the existence of introns allows for an enormous expansion of the molecular diversity through

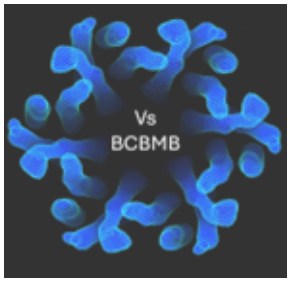
"alternate splicing"
(= generation of different mRNAs from the same pre-mRNA)

and

"exon shuffling"
(= creating of new genes by duplication and/or physical movement of exons to a different location in the genome)



→ alternative splicing allows the generation of >100,000 different polypeptides from a limited repertoire of just ~20,000 protein coding primary genes.
If you add to this chemical modifications that occur after protein synthesis, you have created a truly unlimited repertoire of protein building blocks to execute the genetic instruction manual.



Summary Transcription General

- Transcription is the **first step in transforming DNA code into molecular/chemical diversity**

- Transcription involves the **synthesis of RNA molecules in a DNA-dependent manner**

- both DNA strands can serve as templates for RNA synthesis, but for any given gene only strand will be copied (the coding strand)
- Transcription **is initiated at specialized DNA regions called promoters** which recruit necessary proteins + properly orient the apparatus to transcribe the gene in question.

- **Promoters are built from at least two short (6bp) sequence blocks that are appropriately spaced** upstream of the start site **and** from each other (eg prokaryotes at “-10” and “-35”).

- The design of promoters requires >1 recognition sequence to avoid erroneous initiation = specificity is achieved by **coincidence detection (= AT LEAST two elements that are at the RIGHT distance/spacing)**. Coincidence detection also confers robustness (= point mutations and slight changes in spacings become permissible)

- Promoters are recognized by the transcriptional apparatus that in prokaryotes consists of a sigma factor and RNA-polymerase (which is build from 4 protein subunits). Transcription initiation in eukaryotes is MUCH more complex, involving a sequential process that assembles a pre-initiation complex (general transcription factor complexes + one of three different polymerases, each of which area also made from multiple subunits)

- The complexity of eukaryotic transcription initiation explains the existence of “transcription factories”. **Unlike** in prokaryotes, eukaryotic genes to be transcribed are looped and transported **to** a nearby transcription factory.

- In prokaryotes: transcription yields a fully functional, ready-to-go mRNA; in eukaryotes the presence of introns in protein coding sequences requires extensive post-transcriptional processing to generate the appropriate messenger RNA. The splicing apparatus is mindbogglingly complex involving up to 200 different proteins. This complexity is off-set by the gain in functional expansion = 20,000 genes can yield hundreds of thousands different proteins.