Eukaryotic Gene Expression

Prof. P. N. Rangarajan

Department of Biochemistry

Indian Institute of Science, Bangalore

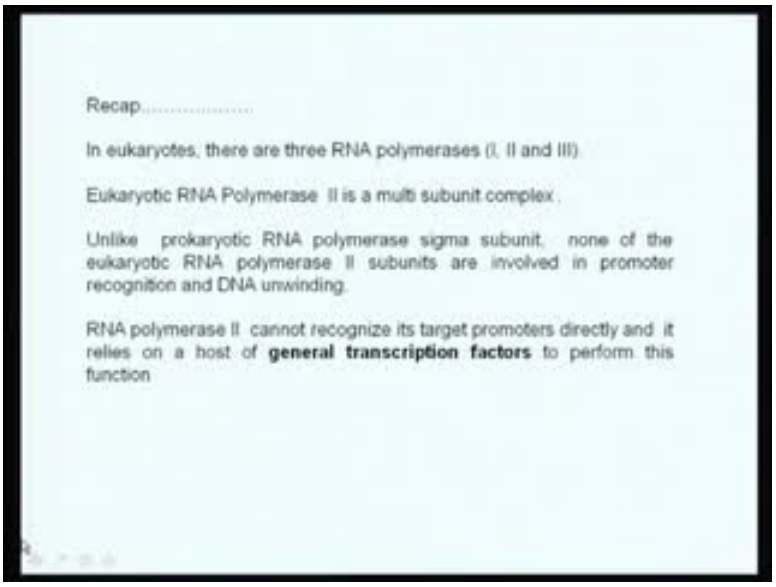
Lecture No. # 02

Gene Regulation in Eukaryotes:

Diversity in Core Promoter Elements

The impression that the sequence, which we have been talking in the previous class, as a TATA box is not just there to invite or polymerase general transcription factor into the promoter, and then just help in transcript creation, but the core promoter elements consist of not only TATA box, but many other sequences. And, these core promoter elements not only help the RNA polymerase to initiate transcription, but there is lot of diversity in the core promoter elements itself. And, this diversity itself can contribute to differential gene regulation. So, the purpose of this talk is going to make you understand that the sequences like TATA box and many other core promoter elements, their job is not only to just initiate transcription, but also to bring in diversity or contribute to differential gene expression you may carry over. That is going to be the crux of today's talk.

(Refer Slide Time: 01:07)

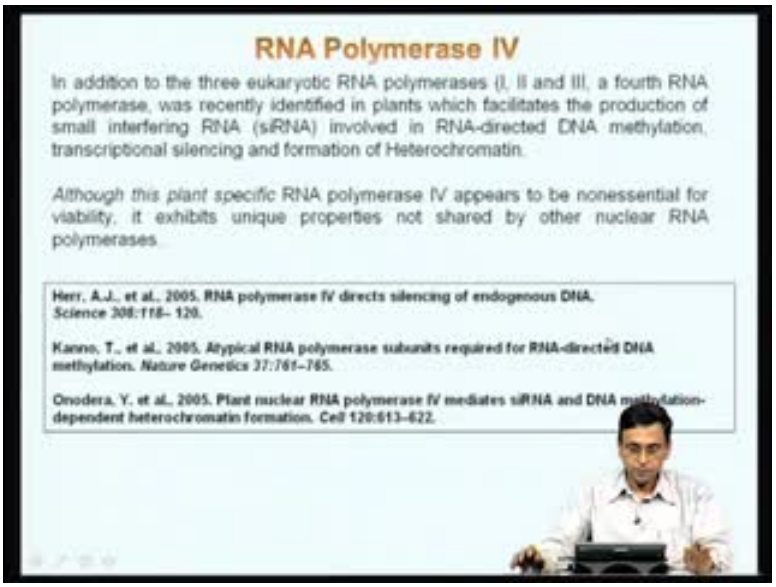


Let us now just recapitulate what we discussed in the previous class. In case you have no time to look at the previous lectures, the summary of what we discussed in the previous lecture is that in eukaryotes, there are about three RNA polymerases, which we call as I, II and III. This is unlike what happens in prokaryotes, where there is only one RNA polymerase, which transcribes tRNA rRNA as well as mRNA; whereas, in the case of eukaryotes, the RNA polymerase has evolved into three different RNA polymerases. RNA polymerase I takes care of repose warning transcription; RNA polymerase II takes care of transcription of protein coding genes; and, RNA polymerase III takes care of tRNA, 5 sRNA and other RNA genes.

We also discussed in the previous class that eukaryotic RNA polymerase, actually is a multi subunit complex. Unlike prokaryotic subunits, there were only four subunits; the eukaryotic RNA polymerase has lot more subunits. And, unlike the prokaryotic RNA polymerase sigma subunits, which actually can recognize the Pribnow box, that is, the minus 10 box and the minus 35 sequence. None of the eukaryotic RNA polymerase 2 subunits are involved in promoter recognition and DNA unwinding. This is another important point we discussed last time; that unlike the sigma subunit of RNA polymerase in bacteria, the eukaryotic RNA polymerase subunits, none of the subunits can actually recognize the promoter. So, the eukaryotic RNA polymerase is kind of a blind. It does not know where the promoter is.

In fact, if you want to do a cell-free transcription, in a test tube, if you want to do transcription, and if you put the DNA template containing promoter in the coding region, and if you also put all the nucleotides, and if you just put purified eukaryotic RNA polymerase, it just goes and randomly binds everywhere, and initiates transcription very randomly, which is very deleterious to the cell, because RNA polymerase every time has to bind in a specific place and initiate transcription from a specific point. Then only, you will get a same kind of mRNA every time the gene is transcribed. So, it is very important to understand that the general transcription factors, that is TFIIA to TFIIH play this key role of bringing RNA polymerase to the exact site in the promoter, so that the RNA polymerase transcription initiates transcription exactly from the same point every time. And, the RNA polymerase II cannot recognize its target promoters directly and it relies on a host of general transcription factors to perform this function. This is what we discussed in detail in the last class.

(Refer Slide Time: 03:19)

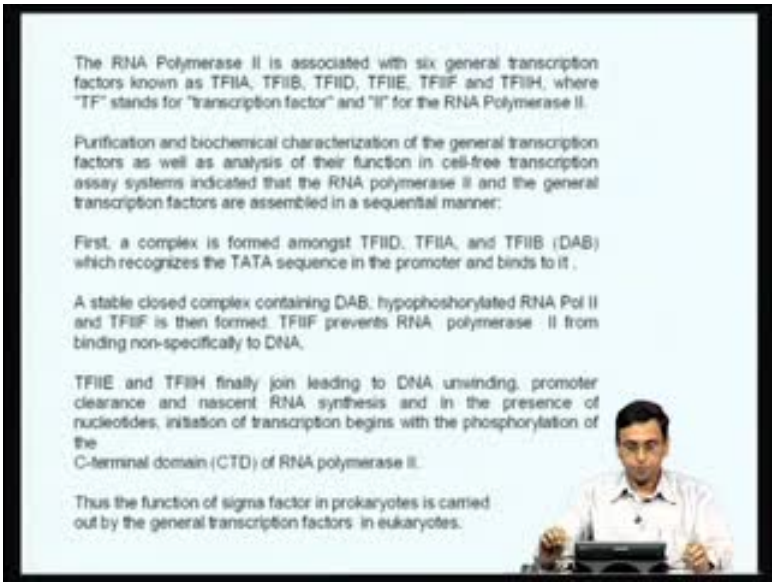


Now, I forgot to mention in the last class that in addition to the three RNA polymerases, there is also a very new RNA polymerase, was discovered in plants and is actually called as RNA polymerase IV. In fact, this was just discovered just in 2005. And, as we can see here, three papers were published more or less simultaneously in science, nature and cell about the discovery of this RNA polymerase, which has now been called as RNA polymerase IV. So, in addition to the three eukaryotic RNA polymerases I, II and III, a fourth RNA polymerase, was recently identified in plants and this RNA polymerase actually facilitates the production of small interfering RNA. Now, this point I do not want to digress into explain into what is siRNA and so on and so forth; we will discuss it much later in this course. Just remember this RNA polymerase IV, which was newly discovered in plants, actually is involved in the transcription of production of siRNA. And, this siRNA is actually involved in RNA directed DNA methylation transcriptional silencing and formation of heterochromatin.

I do not want to explain all these terminologies now; we will discuss all these things when we actually come to the role of chromatin in regulation of gene expression and we talk about histones, and so and so forth. At that time, we will explain in the detail, but just remember at this time a point that a novel RNA polymerase refer to the RNA polymerase IV, has been discovered in plants. So, in addition to the polymerase I, II and III, which most of you have studied in text books, here now, we have a novel RNA polymerase, which is called RNA polymerase IV. It is also important to remember, this

RNA polymerase IV is not essential. That means if you delete this RNA polymerase code in gene, it is not that the plants will die; the plant also continue to survive. So, it is called as a non-essential gene, but it does expect to exhibit many unique properties, which the other nuclear RNA polymerases do not exhibit. Now, we will not go into too many details about the RNA polymerase structure and function, and so on and so forth. Just remember at this point that a new RNA polymerase called RNA polymerase IV has been discovered in plants and it seems to be involved in the production of the siRNA, which is involved in gene silencing. We will discuss this later in the course.

(Refer Slide Time: 05:18)



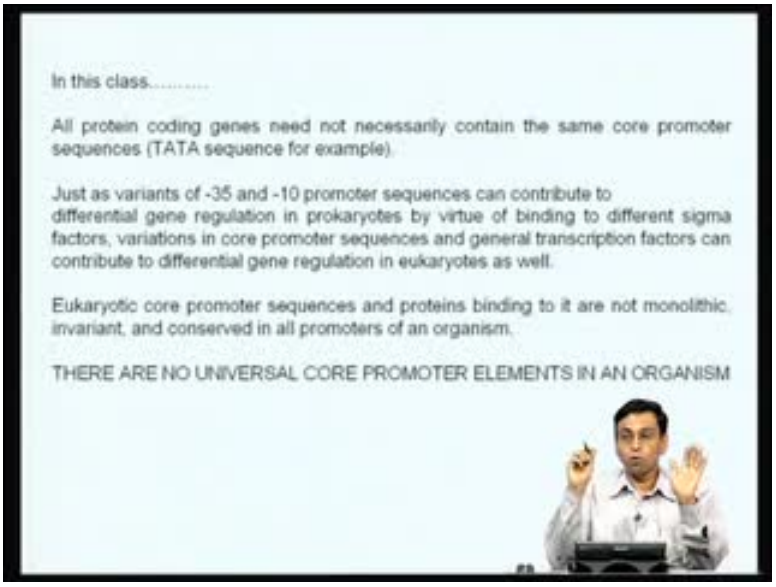
Now, I also discussed, in the last class that the RNA polymerase II is associated with six general transcription factors like I mentioned TFIIA to TFIIH, and TF in all these case terminologies stands for transcription factor, and II stands for transcription factor II, that is, the protein coding genes or the RNA polymerase II.

Now, purification and biochemical characterization of all the general transcription factors as well as analysis of their function is cell-free transcription assay systems indicated that RNA polymerase II and the general transcription factors are assembled in a sequential manner. This also we discussed in detail, wherein I actually told you that first, a complex called a DAB complex involved (( )) TFIID, TFIIA and TFIIB, is first assembled. And, this is the one that actually recognizes the TATA sequence in the promoter region and the DAB complex is first formed. And, once the DAB complex is

formed, the RNA polymerase II in come conjunction with TFIIF is recruited to this previous session complex, and the TFIIF actually prevents RNA polymerase from binding non-specifically to DNA. So, the specificity of RNA polymerase to promote is actually brought out by general transcription factors.

And finally, the TFIIE and TFIIH join the previous session complex resulting in DNA unwinding, promoter clearance, and RNA synthesis begins in the presence of nucleotides. And, the signal for the initiation of transcription is the phosphorylation of the C-terminal domain of RNA polymerase II by TFIIH. So, the crux of what I told you yesterday is that the function of sigma factor in prokaryotes, namely the promoter recognition in DNA unwinding, is actually carried out by general transcription factors in eukaryotes.

(Refer Slide Time: 06:51)

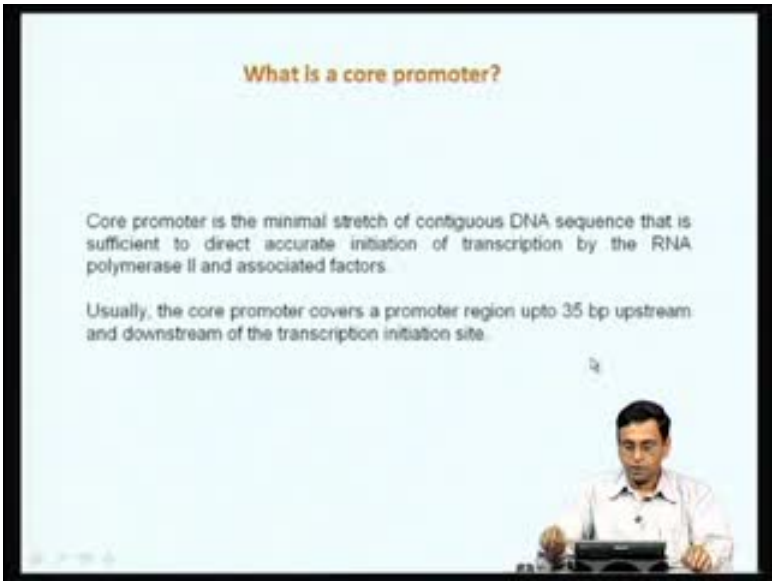


Now, in this class, what I would like to emphasize is that so far, we thought the core promoter means only the TATA box. Now, I do not want to give you the impression that it is only the TATA box that is actually involved in the initiation of transcription and this is the only element that constitutes the core promoter element. Now, the crux of today's class is to tell you that all protein coding genes need not necessarily contain the same core promoter sequences. There is lot of variation in this core promoter sequences and TATA box is actually present only in the subset of protein coding genes in eukaryotes. There are many other core promoter sequences, which are involved in transcription

initiation and this diversity in core promoter sequence itself can contribute to differential gene regulation. This is what is going to be crux of today's talk.

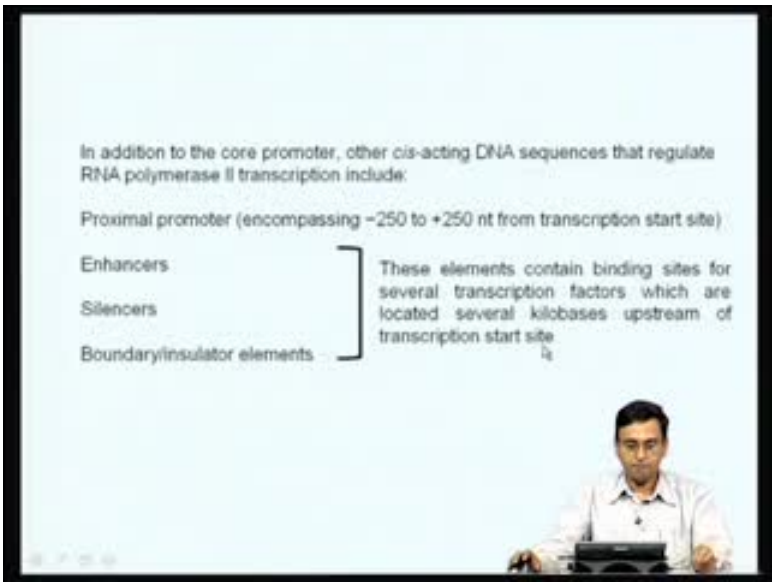
Just as variants of minus 35 and minus 10 promoter sequences can contribute to differential gene regulation prokaryotes by virtue of binding to different sigma factors, variations in core promoter sequences and general transcription factors binding to it can also contribute to differential gene regulation in eukaryotes. So, this is what is going to be the crux of today's talk. So, eukaryotic core promoter sequences and proteins binding to it are not monolithic. That means there is not just one type of core promoter sequence and the proteins binding to it, there is lot of variation in that, and in fact, they are highly invariant and they need not be conserved in all promoters of an organism. So, there is not just one single TATA box, one single TFIID, and so on and so forth; there is lot of variations in this. And, in this class and then in the next class, I am going to discuss about the diversity in this core promoter elements as well as diversity in the general transcription factors, and how this diversity itself can contribute to differential gene regulation. So, the bottom line I want to say is that there are no universal core promoter elements in an organism, which means it is not that TATA box is the only core promoter element, which is present in all the protein coding genes, there is lot of variations; there is nothing like a universal core promoter element, which is present in all the eukaryotic promoters. That is what the take home message is.

(Refer Slide Time: 08:49)



Let us now try to understand what exactly we mean by a core promoter. Now, core promoter can be defined as the minimal stretch of a contiguous DNA sequence that is sufficient to direct accurate initiation of transcription by RNA polymerase II and its associated factors. This is the simplest definition one can think about a core promoter. And, usually, the core promoter covers a promoter region about 35 base pairs upstream or 35 base pairs downstream of the transcription start site. Now, we all know the transcription start site is recognized as plus 1. We are going to discuss today about the promoter elements, which is about 35 base pairs upstream or 35 base pairs downstream of this transcription start site, and this is what constitutes a core promoter sequence. Now, let us try to understand what is happening in around the core promoter sequence, what are the variations in this core promoter sequences.

(Refer Slide Time: 09:36)

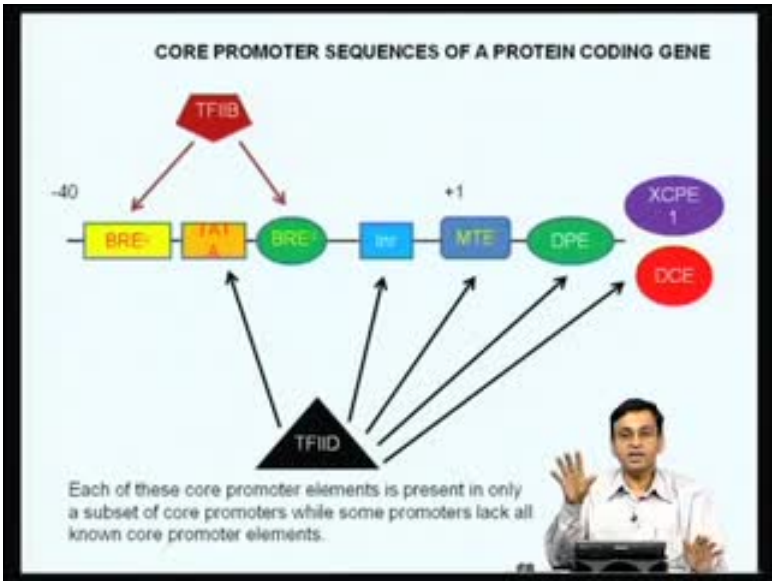


In today's class, we are going to discuss only about core promoter sequences, does not mean there are no other such acting elements. So, roughly, the protein coding gene of a eukaryote… There are number of other regulator elements outside also, contribute to differential gene regulation. For example, there is something called a proximal promoter, which encompasses about minus 250 to plus 250 nucleotides from the transcription start site. Today, we are going to talk about sequences only about plus 30 to minus 30 of the in and around the transcription start side. But, if you go further beyond either upstream or downstream, these are actually referred to as the proximal promoter sequences, and

there are protein factors, which bind to these sequences and they also contribute to differential gene regulation. We will talk about this at later stages.

There are also what are called as enhancers, silencers, boundary or insulator elements. And, all these elements usually are present several kilobases upstream for the transcription start site. And, they actually serve as binding sites for several transcription factors. So, it is the combined action of all these promoter elements and all these protein factors that ultimately contribute to differential regulation in eukaryotes. But, today, we are going to focus only about the core promoter sequences, which are present about 30 to 40 base pairs upstream and downstream of the transcription session site.

(Refer Slide Time: 10:53)



Now, as I mentioned earlier, in the first class, I kept on talking only about TATA box. What I am going to tell today is that TATA box is not the only core promoter sequence, which is present in the eukaryotic promoter; there are several variants and some of the well characterized core promoter sequences are shown in this particular cartoon. For example, there is something called as a BRE; it is actually called as TFIIB response element; whereas, (( )) there are two variants called BRE U, BRU D; U refers to upstream and D refers to downstream. We will discuss what exactly all these things are a little bit later. There is of course, what is called as an initiator motive, which is again a very important core promoter element. There is again something called an MTE, DPE, XCPE, DCE, and so on and so forth. What I want to emphasize from this cartoon is that

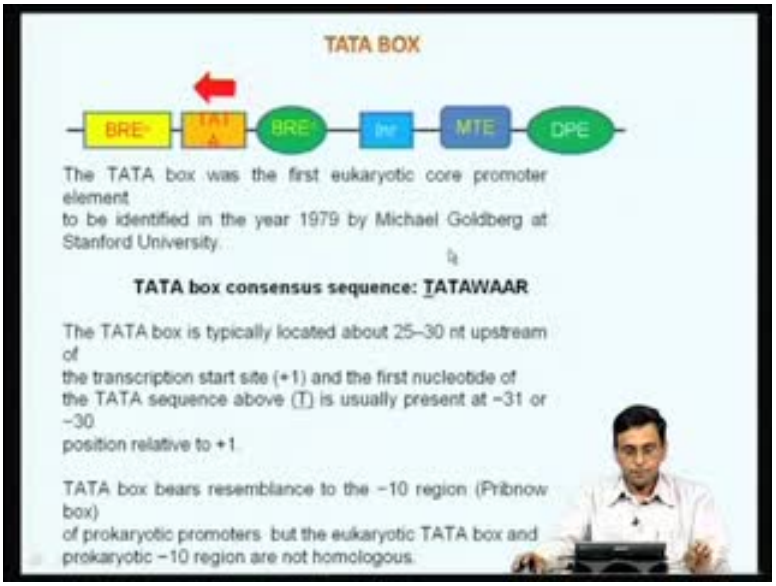
TATA box is not the only core promoter sequence of a protein coding gene; there are many other core promoter elements, which have been discovered and these sequences are also presented a number of eukaryote promoters. And, TATA box need not be present all the time in all the protein coding genes.

Now, what is also been demonstrated is that TATA box is not the only sequence to which this TFIID actually binds. As you know, TFIID is nothing but the TATA binding protein and a bunch of protein called TAFs; know TBP associated factors. And, I told you in the last class, it is the TFIID, which is actually involved in the promoter recognition and the TBP component of TFIID actually recognizes TATA box and the DAB complex is therefore, formed in the TATA box. But, the TFIID not only recognizes the TATA box, it also recognizes several other core promoter elements like initiator, MTE, DPE, DCE, and so on and so forth. So, already you can see, there is diversity here. So, remember TATA box is not the only core promoter element recognized by the TFIID general transcription factor; TFIID can also recognize a number of other core promoter sequences, such as initiator, MTE, DPE, and DCE, and so on and so forth. Similarly, in the last class, I told you, the only component of the general transcription machinery, which can bind DNA is the TFIID; that is, TBP and associated factors. But, now, later studies have actually shown even the TFIIB component of the general transcription factors can recognize DNA. And in fact, the sequences which are recognized by TFIIB is actually called as TFIIB recognition element or BRE. So, you can see already we are deviating from some of the basic finding that we discussed in the last class; that as we start analyzing more and more genes in eukaryotes and more and more organisms, you can see there are many exceptions to this general rule that we discussed in the last class. And, as we discussed some of these examples, it becomes very clear; diversity in these core promoter sequences itself can contribute to differential gene regulation in eukaryotes. So, each of these core promoter elements is present in only a subset of core promoters while some promoters lack all known core promoter elements. This is very important for you to know.

All these core promoter elements I have discussed here may be present in one or the other gene. For example, some promoters may contain TATA box; there are also what are known as TATA-less promoters. They do not contain TATA box. So, instead of TATA box, they actually contain what is called as initiator. So, the function of TATA

box in these promoters is actually taken over by the initiator sequence. Like similarly, some contain MTE, DPE, DCE, and so on and so forth, what research have been shown is that there are many promoters, which do not contain any of these core promoter elements suggesting that there may be still unknown core promoter elements, which are yet to be discovered. So, we are still not completed the story; this story is ongoing. As and when we analyze more and more genes, there will be more novel core promoter elements will be discovered. But, so far, at least these many number of core promoter elements have been discovered.

(Refer Slide Time: 14:48)



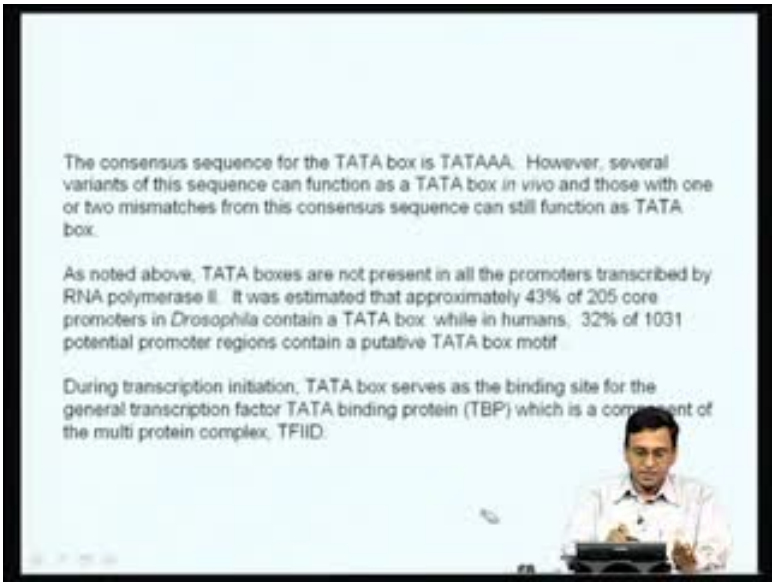
Now, let us do one-by-one and then see what is known about each one of these core promoter elements. Now, the first and the foremost core promoter element, which was identified is the TATA box. So, TATA box is the first eukaryotic core promoter element, which was identified in 1979. Actually, it is a Ph.D. thesis of a person called Michael Goldberg at Stanford University. He actually identified that eukaryotic protein coding genes actually contained a sequence call TATA box, which is actually recognized by the basal transcription machinery and is essential for transcription initiation.

Now, by analyzing a number of eukaryotic promoters, a consensus sequence was drawn for the TATA box and this consensus sequence is TATAWAAR. What is W? W is actually an IUPSC code for nucleotides and I will tell you what exactly W is in my next couple of slides. The TATA box is typically located about 25 to 30 nucleotides upstream

of the transcription start site. As I told you, core promoter elements usually are present about 30 to 40 base pairs upstream or downstream of the transcription start site. And, in many of the eukaryotic promoters, this TATA sequence is actually present about 25 to 30 nucleotide upstream of the transcription start site, which is usually designated as plus 1. So, the first nucleotide of the TATA sequence, that is, what I have underlined here, the T is usually present at minus 31 or minus 30 of the promoter region. For example, if you just draw a eukaryotic promoter…; for example, if this is a transcription start site, which we call as plus 1, the TATA sequence is usually present about minus 30 to minus 31, and this is where the TATA box is present. So, this is what we need to understand.

In the TATA box, remember, we also in the first class discussed about something called as a Pribnow box or minus 10 region and minus 35 region. And, it is these sequences, which are recognized by the sigma factor of the prokaryotic RNA polymerase. So, the TATA box is kind of resembles the minus 10 region of the Pribnow box, but people have clearly shown that it does not mean that the TATA box actually a homolog of the Pribnow box. It is just a coincidence that the TATA box more or less functionally at least resembles the Pribnow box of the eukaryotes.

(Refer Slide Time: 17:21)



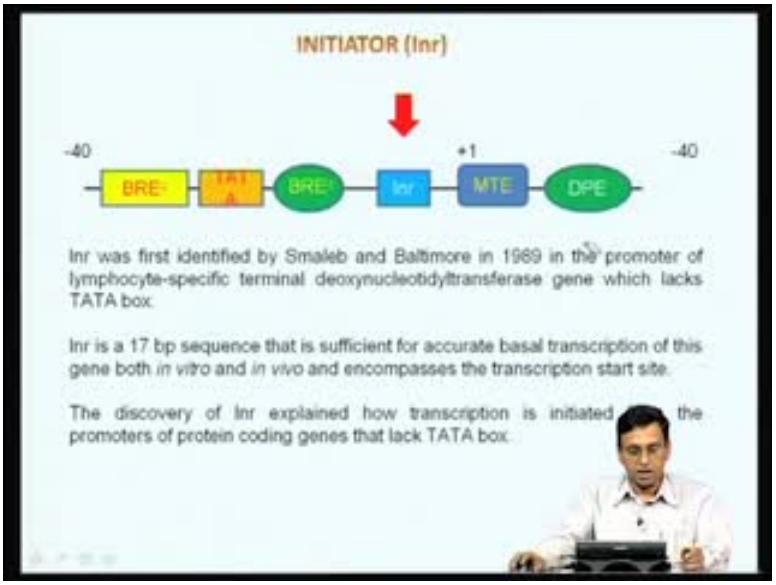
Now, the consensus sequence for the TATA box is as I said TATAAA. But, several variants of this sequence can function as the TATA box in vivo and those with one or two mismatches from the consensus sequence can still function as the TATA box.

Although this is a consensus sequence, there are many variations. As I said, as you start studying more and more of the transcription of the protein coding genes, you soon realize that you start seeing variations of the general role.

The TATA boxes are not present in all the promoters transcribed by RNA polymerase II. Now, most of you who have studied some basic text books in eukaryote gene expression, you probably think, the moment you ask about how transcription is initiated in the eukaryotes, you will say TATA box. TATA box is the place where the preinitiation complex assemble and that is where the transcription initiation takes place. But, remember, TATA box is not the only core promoter element and there are many promoters, which actually lack TATA box. In fact, it was estimated that approximately 43 percent of about 200 odd core promoter sequences analyzed in Drosophila contain a TATA box. So, another 60 percent do not contain. 32 percent of about 1000 potential promoter regions contain a putative TATA box motif. So, there are many promoter sequences, which do not contain for a functional TATA box.

Now, what is the function of TATA box? During transcription initiation, TATA box actually serves as the binding site for the general transcription factor TATA binding protein, which is a component of the multi subunit complex, TFIID, which we have discussed in detail in the last class.

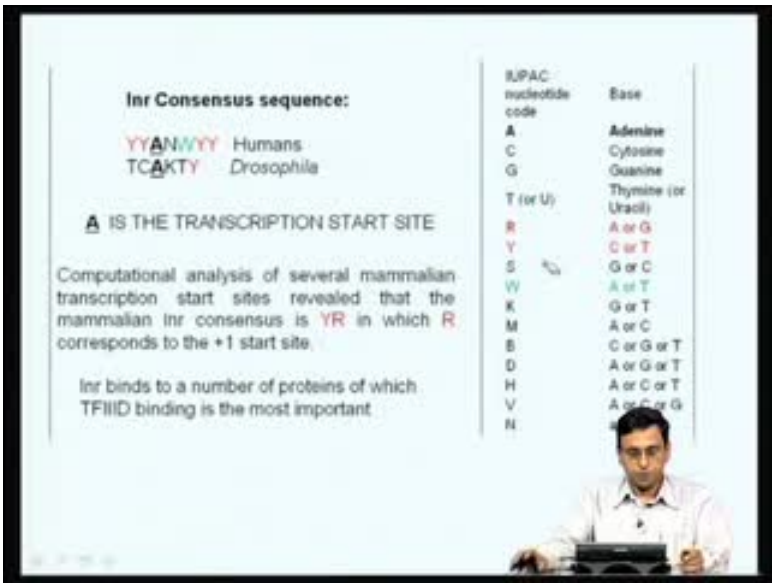
(Refer Slide Time: 18:54)



Now, the next important core promoter element, which was discovered after TATA box is what is called as an initiator motif. Again, this motif called as initiator was first identified by Smaleb and Baltimore in 1989 in the promoter of lymphocyte-specific terminal deoxynucleotidetransferase gene, which actually does not contain TATA box. So, they asked the question, if this promoter does not contain a TATA box, how does the RNA polymerase II recognize this promoter, how does general transcription factor go and bind and is soon identified? There is actually something called an initiator motif, in and around which the preinitiation complex is actually formed. Therefore, in TATA-less promoters, the initiator motif actually serves the function of a TATA box for promoter recognition. So, initiator is nothing but a 17 base pairs sequence that is sufficient for accurate basal transcription of both in vitro and in vivo and it actually encompasses the transcription start site. Now, unless the TATA box, which is usually present are on minus 30 or minus 31 from the transcription start site, the initiator motif seems to be present in and around the transcription start site. This is the basic difference between the initiator and the TATA motif.

Now, the discovery of the initiator actually explained how transcription can be initiated from promoters of protein coding genes, which actually lack a TATA box. So, once people started realizing that TATA box need not be present in all the promoters, people started examining how this TATA-less promoters are transcribed. And, as a consequence of this research, at least one core promoter motif called initiator was identified in 1989 by Baltimore’s group.

(Refer Slide Time: 20:25)

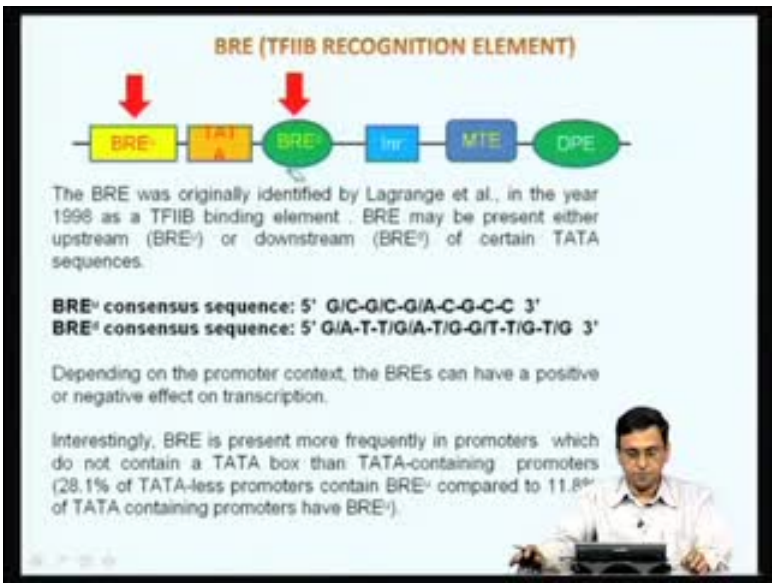


Now, if you look at this, again by analyzing a number of TATA-less promoters in a wide variety of eukaryotes like flies, humans, and so on and so forth, people have identified what is called as a consensus sequence for initiator motif. In humans, the consensus sequence is YYANWYY. Now, I would like to mention here, probably, when we talk about nucleotides, you are only aware of A T G and C in the case of DNA, and U, urasil in the case of RNA. But, IUPAC, that is, International Union of Pure and Applied Chemistry are actually come up with number of nucleotide codes. For example, suppose I want to say that this particular sequence contains either A or G; that means it contains purines, then I use a letter R to designate that R means it can be either A or G. Similarly, Y means it is pyrimidines, that is, it can be either C or T. Similarly, W means it can be either A or T. So, the IUPSC has come up with a number of nucleotide codes, single letter codes to designate a combination of nucleotides. Therefore, when I say W, W actually means this region may contain either A or T. Similarly, Y means it can be either C or T, and so on and so forth. So, I have given the actually IUPSC code here for various nucleotides. So, in this consensus sequence, this underlined A actually serves as the transcription start site. So, the initiator motif actually encompasses the transcription start site in many of the TATA-less promoters.

Computation analysis of several mammalian transcription start sites revealed that the mammalian initiator consensus is YR in which R corresponds to the plus 1 of the transcription start site. So, again, R means it is a purine; it can be either A or G. Now,

initiator binds to a number of proteins of which TFIID binding is the most important. So, remember, the TFIID component of the general transcription factor can not only bind a TATA box, but it can also bind to an initiator motif in TATA-less promoters.

(Refer Slide Time: 22:22)



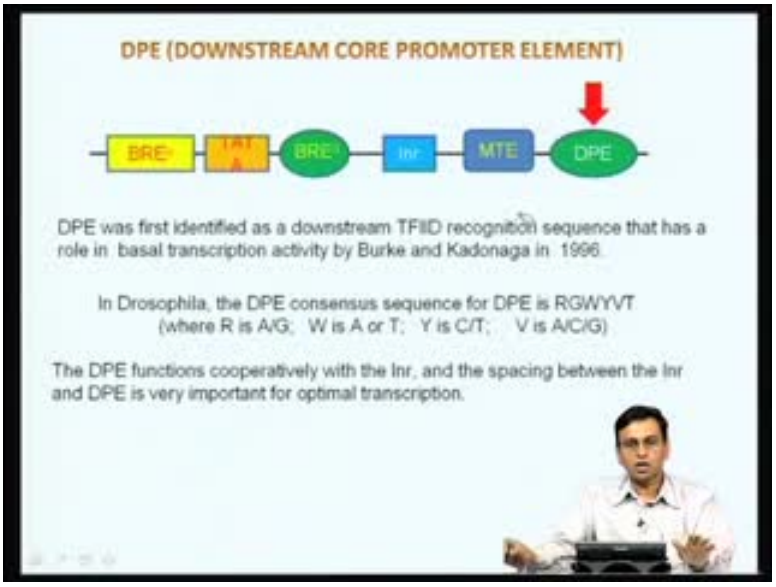
Now, the second important cis-acting element or the core promoter element, which was discovered in addition to TATA box initiator is what is called as the BRE, which actually stands for TFIIB recognition element, BRE (TFIIB recognition element). The BRE was originally identified by Lagrange et al., in the year 1998. And, this was actually found or identified as a TFIIB binding element; that is, the sequence to which the TFIIB transcription factor goes and binds. And, the BRE may be present either upstream; when it is present upstream, it is called BRE U or it may be present downstream of the TATA box when it is called as BRE D. So, there are two variants; when the BRE is present upstream of the TATA box, it is called BRE U; when it is present downstream of TATA box, it is called as BRE D.

Now, depending on the promoter contest, the BREs can have either positive or negative effect on the transcription. So, remember, this is what I kept talking about. So, do not assume that when I say core promoter sequences, their job is just to recruit RNA polymerase to the promoter. Variations in the core promoter sequence itself can lead to differential gene regulation. That is the crux of the thing. So, depending upon what kind of other promoter sequences are there, these BRE sequences can either activate

transcription or they can actually repress transcription depending upon what kind of promoter you are talking about.

The BRE is actually present more frequently in promoters, which do not contain a TATA box than those promoters, which contain a TATA box. For example, 28.1 percent of TATA-less promoters contain BRE U compared to 11.8 percent of TATA containing promoters, which have BRE U. So, usually, those promoters which do not contain a TATA box seem to contain other core promoter sequences like initiator, BRE, and so on and so forth. So, you can already see the diversity in the core promoter elements. Some eukaryotic promoters contain the typical TATA box; whereas, in some cases, the TATA box function is replaced by initiator; in certain other promoters, neither you have initiator nor you have the TATA box, but you may actually have something called as a BRE.

(Refer Slide Time: 24:33)

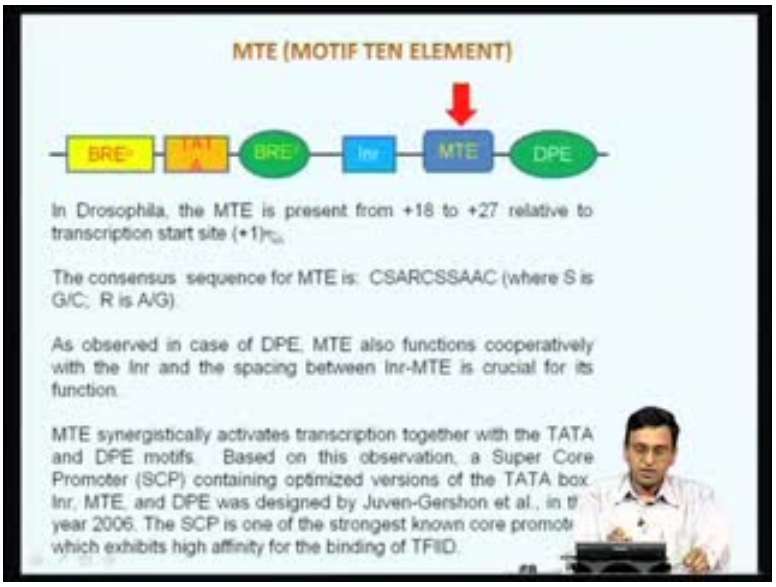


Now, the other core promoter element, which was discovered, is what is called as a DPE, which actually stands for downstream core promoter element, because this is actually present downstream of the transcription start site. Now, initiator is actually present on the transcription start site; plus 1 is somewhere here. So, DPE is actually present downstream of the transcription start site. Therefore, it is called as downstream core promoter element or DPE.

Now, the DPE was again first identified as a downstream TFIID recognition sequence that has a role in basal transcription activity by Burke and Kadonaga in 1996. Now, again, you see this is an important observation. Do not assume that the TFIID or the TBP always need to bind to a region upstream of the transcription start site; it can bind right near the transcription start site; that is, in the case of the initiator, the TFII chaotic can also bind to core promoter elements, which are kind of situated little bit downstream of the transcription start site. So, you can already see there is lot of variations.

Now, in the Drosophila, the DPE consensus sequence is RGWYVT and again, I have explained here what actually each one of these letters means as per the IUPSC code. Now, the DPE functions cooperatively with the initiator, and the spacing between initiator and DPE is very important for optimal transcription. So, you can see the exact position of the DPE, downstream of the transcription start site, that is, this distance between initiator and DPE is highly conserved, and this distance is very important for the functional role of downstream core promoter element in many of the eukaryotic promoters.

(Refer Slide Time: 26:07)



The other important core promoter element, which has been discovered, is known as the MTE or motif ten element. Again, this was first discovered in Drosophila and this MTE is actually present from plus 18 to plus 27 relative to the transcription start site. So, remember, again, the two elements that we are discussing so for, that is, DPE and MTE,

are actually present downstream to the transcription start site; whereas, all other elements discussed so far are either present in and around the transcription start site or upstream of the transcription start site. Two of the core promoter sequences, DPE and MTE, are actually present downstream to the transcription start site. And, in the case of MTE, which is the motif ten element, first identified in Drosophila is somewhere between plus 18 to plus 27 relative to the transcription start site. So, the consensus sequence for MTE is again shown here, where S means it can be either G or C; and, R is actually purine; it can be either A or G as per the IUPSC nucleotide codes.

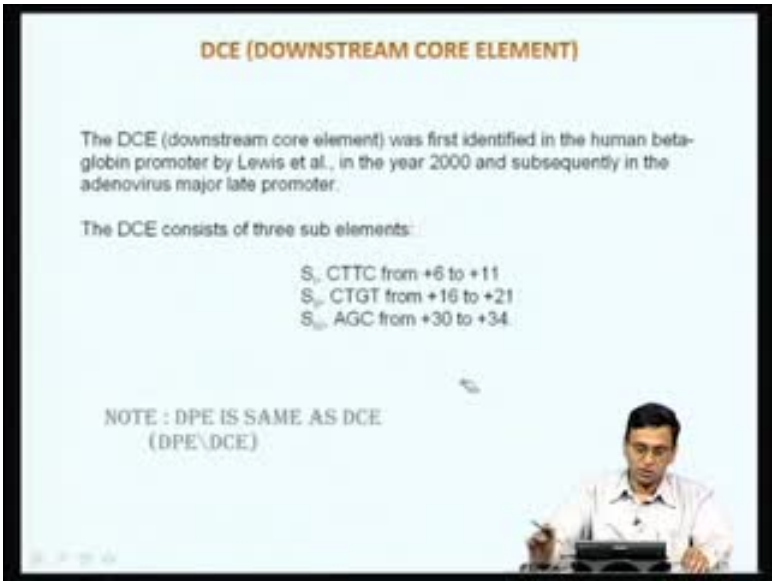
As observed in the case of DPE, the MTE also functions cooperatively with initiator and again the spacing between initiator and MTE is crucial for its function. So, remember, both in the case of the digital core promoter elements, that means, digital to the transcription start site, the spacing between initiator and MTE, that is, this region in the case of MTE (Refer Slide Time: 27:27), and similarly, in the case of DPE, this spacing between initiator and DPE, the spacing is very important for their function. That means this (( )) we had positioned exactly at a specific distance from the transcription start site. This is very important.

Many a times, we keep asking what is the significance of understanding all these jargons; why should we try to understand all these core promoter elements; and so on so forth. In fact, the lecture title is eukaryotic gene expression basics and benefits. As we discuss some of these basic aspects, you may actually find many of these things boring. But, as and when I am going to give you examples of how some of these understandings, some of these core promoter elements, have actually led to some of the application. For example, I will give one example here; the MTE, which is the motif ten element, synergistically activates transcription together with the TATA and DPE motifs. So, when you have TATA box alone or TATA plus DPE motifs, you get a certain level of transcription. But, if you meet MTE in addition to these two elements, you get much higher levels of the transcription suggesting that the MTE synergistically activates transcription in conjunction with TATA and DPE motifs. Based on this basic observation, something called as a super core promoter (SCP) was actually constructed.

Now, this is kind of a synthetic promoter, which contains the best TATA box sequence, that is, the most optimal TATA box sequence to which TFIID binds very efficiently. The best initiator motif based on analysis of various number of Drosophila promoters and the

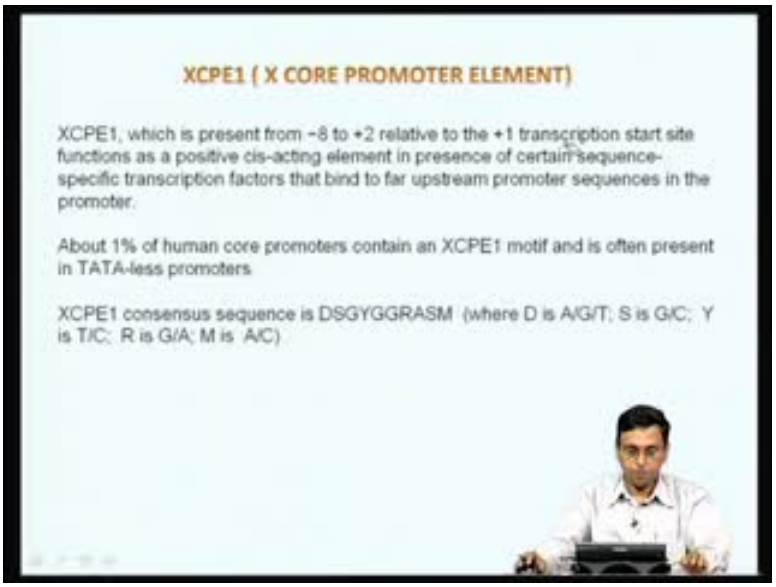
best MTE and DPE, was designated by this people in 2006. And therefore, they are able to kind of construct a new eukaryotic promoter, which is one of the strongest known core promoters, which exhibits very high affinity for binding to TFIID. So, by actually understanding how these core promoter elements actually function and how these core promoter elements interact with each other, it is actually possible to discover or construct novel promoter sequences to which the TFIID can bind very efficiently. And, such promoters can actually be used for recombinant protein production in using many expressions. So, you can see already some kind of applications came out of many of these identifications of these core promoter elements. So, there have been benefits arising out of some of these basic researches.

(Refer Slide Time: 29:39)



Now, the other important core promoter element, which was identified in some of the promoters, is what is called as DCE, named as the downstream core element. The downstream core element was first identified in the human beta-globin promoter by Lewis et al., in the year 2000 and subsequently in the adenovirus major late promoter. And, this DCE actually consists of three other sub elements called S I, S II and S III. The sequence is given here and they are actually present at different regions from the transcription start site. So, what I told you so far is that there are at least three downstream core promoter elements: DCE, MTE and DPE (Refer Slide Time: 30:18); MTE, DPE and DCE are all the core promoter elements, which are actually present downstream of the transcription start site.

(Refer Slide Time: 30:27)



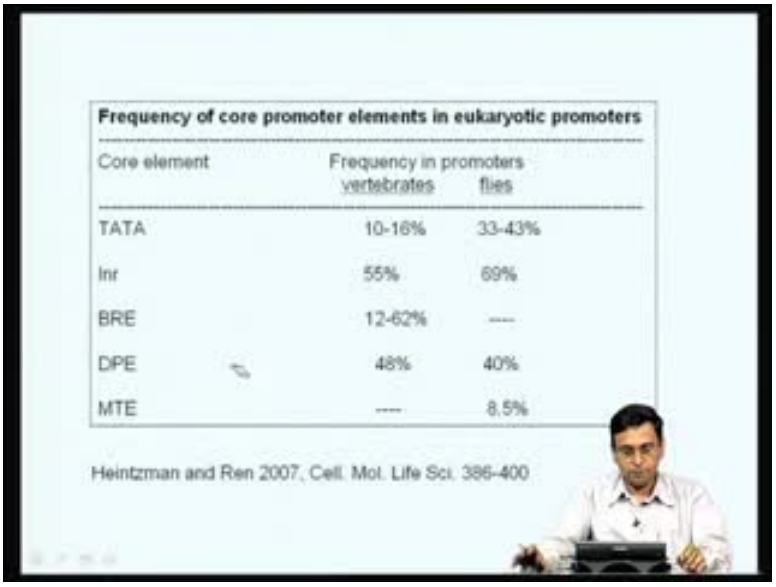
Now, the other important core promoter element, which has been identified in some of the promoters, is what is called as XCPE1, which stands for X core promoter element. XCPE 1 was actually present from minus 8 to plus 2 relative to the transcription start site, and actually functions as a positive cis-acting element in presence of certain sequence-specific transcription factors that bind to far upstream sequences in the promoter. Now, as I told you, all the way of studying only the core promoter sequences, which are present between minus 30 upstream or minus 30 downstream of the transcription session site, do not think that these core promoter sequences function in isolation. This core promoter sequence is almost like a cricket pitch; that is where the action is. But, there is also what is called as an outfield; then, you have a stadium; then, you have people. So, all these things in conjunction together, collectively, actually add to the media of the game. Then only, the cricket becomes lively.

The same way, what we are now talking about eukaryotic gene expression or the eukaryotic promoter sequences, is only about the cricket pitch. That is the core promoter sequence, where the general transcription factors come and assemble. But then, you also have the outfield, which consists of proximal promoter sequences, digital promoter sequences, enhancers, silencers and so on, which again act in conjunction with the core promoter sequences. And then, we have various other protein factors, which actually interact with all these promoter sequences, and it is the collective action of all these things that ultimately results in regulation of gene expression eukaryote. And, this

XCPE1 is one such core promoter element, which actually plays a very important role in certain promoters, because it acts in conjunction with some of the protein factors, which actually bound to many upstream regions of these promoter sequences.

Now, about 1 percent of the human core promoters contain an XCPE1 motif and is again present in the TATA-less promoters. So, as you can see, although the TATA box was the original core promoter element, which was identified, as people started analyzing more and more eukaryotic promoters, it became very clear that this is not the only sequence, which is present; and, as people started identifying many promoters, which do not contain TATA box, they started identifying other core promoter sequences; many of which I have now discussed here. The consensus sequence for XCPE1 is shown here. Again, I have mentioned what each one of these letters stand as per the IUPSC code.

(Refer Slide Time: 32:46)

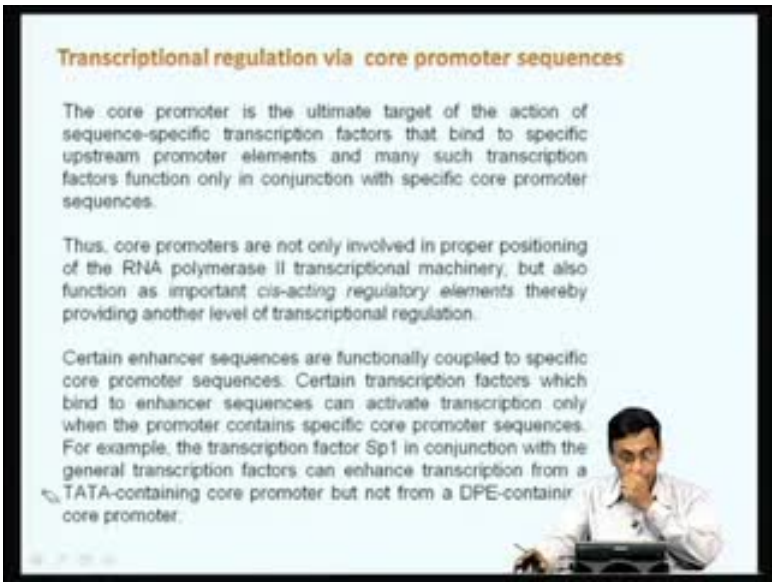


The summary of what I am going to tell you today is what this is. There are number of variance of the core promoter elements. Do not think TATA box is the only core promoter element, which is present on eukaryotes. TATA box is actually present only in about 10 to 16 percent of the vertebrate promoters and about 33 to 40 percent of the promoters of flies like Drosophila. So, there are many promoters, which do not contain a TATA box. Similarly, if you look at the initiator motif, which is also a core promoter element, it is present in about 55 percent of the vertebrate promoters, and about 69 percent of the fly promoters. The BRE is present in about 12 to 62 percent of the

vertebrate promoters; whereas, the DPE, that is, the distinct promoter element is actually present in about 48 percent of the vertebrate promoters and 40 percent of the fly promoters. The MTE is present in 8.5 percent of the Drosophila promoters. So, the crux of what I have told you here is that there is diversity among core promoter elements, and also, TATA box was the first core promoter element I identified. Subsequent studies indicated that the remaining promoters, which do not contain a TATA box. And, in such cases, many other core promoter elements actually function as either TFIID binding elements or TFIIB binding elements, and they also play a very important role in the general transcription factors. That is just like a cricket pitch.

When you say cricket pitch, it is not just all cricket pitches behave the same way. There are some pitches, which are flat; some pitches, which are bouncy; some pitches, which are spin bowlers; some is very good for batsman. The same way, although the core promoter elements' basic job is to assemble RNA polymerase II for initiation of transcription, there are many variations of the core promoter elements; some of them contain TATA boxes; some of them contain BREs, GPEs, MTEs, and so on and so forth. And, this variation in the pitch, just like in cricket, the game becomes more lively, because if the result becomes unpredictable, in the same way, variations in the core promoters also add to differential gene regulation.

(Refer Slide Time: 34:48)



Let us now spend some time to understand how actually the diversity in the core promoter sequences actually contributes to differential gene regulation. The core promoter is the ultimate target of the action of the sequence-specific transcription factors that bind to specific upstream elements and many such transcription factors function only in conjunction with specific core promoter elements. This is very important. As I said, just as the action in a cricket field is concentrated on the cricket pitch, the core promoter sequences is the place where the action actually takes place, because this is the place where the RNA polymerase has to bind and initiate transcription. But, many of the protein factors like upstream activator proteins, proximal promoter sequences, digital promoter sequences, enhancer binding proteins, repressors; ultimately, their job is to modulate RNA polymerase assembly in the preinitiation complex. So, this core promoter sequence has a very important role.

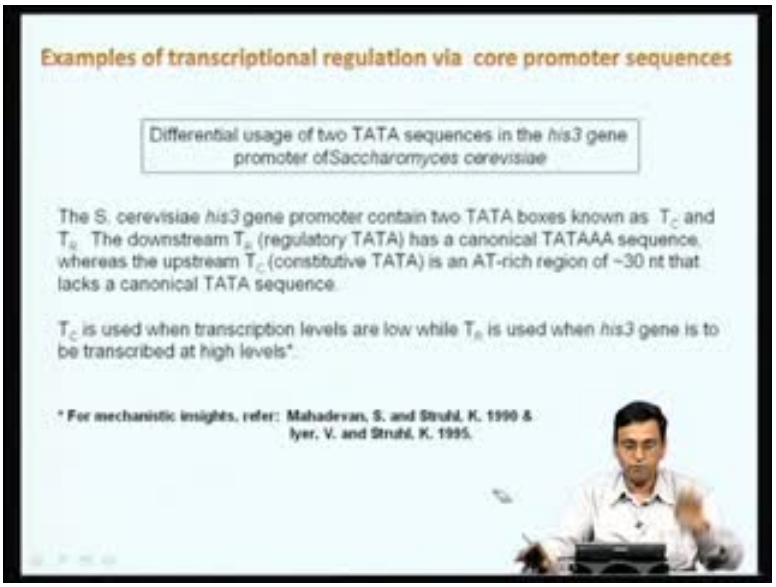
The core promoters are not only involved in proper positioning of the RNA polymerase II transcriptional machinery, but they also function as important cis-acting regulatory elements, thereby providing another level of transcription regulation. The reason I am spending this much time to emphasize the importance of core promoter elements is that most of you would have studied in text books that the only job of this minus 10 or minus 30 in the case of prokaryotes and the TATA box in the case of eukaryotes is to serve as a recognition site for RNA polymerase and general transcription to come and bind. Already, I told you in my first class, this minus 10 and minus 30 sequences in prokaryotes, by variations, can also lead to what is called as sigma-factor switching, and variations in the minus 10 and minus 30 sequences can actually lead to differential gene regulation; that is what I have told you in the last class. In the same way, variations in the core promoter sequences and variations in general transcription factors binding to it, can also contribute to differential gene regulation. Therefore, the function of the core promoter element is not just to serve as a pitch for bowling. There are variations in the pitch, which can actually some time as a spin bowlers; some time as a fast bowlers; some time as a batsman. Similar way, there are number of variations in the cis-acting regulatory elements of the core promoter elements, which actually contributes to differential gene regulation.

In fact, this is very important – certain enhancer sequences are functionally coupled to specific core promoter sequences. You can see how important the core promoter

sequence is. For example, there are many DNA sequences, which are actually present further upstream in the promoter region and many of these further upstream elements actually can function only when a specific core promoter sequence is present. For example, some of these upstream sequences can present only when a TATA box is present downstream. If you replace the TATA box with the initiator, this upstream sequence can no longer function as an enhancer; or, it can no longer function as any activation sequence. So, there is a link between the core promoter sequence and the sequences, which are further upstream.

Certain transcription factors, which bind to enhancer sequences, can activate transcription only when the promoter contains specific core promoter sequences. That is what I just mentioned. For example, the transcription factor called SP1, which actually finds to further upstream sequences, which we are going to discuss much later, called the digital promoter sequences or proximal promoter sequences. This SP1 can actually bind to those sequences and function as a transcription activator only from a TATA-containing promoter. But, if you replace the TATA box with a DPE, SP1 can no longer activate transcription. So, whether SP1 can activate transcription or not from a digital promoter element, depends on whether TATA box is the core promoter sequence or not. When TATA box is not present, SP1 cannot activate transcription although it binds to the upstream action sequences. So, the core promoter sequences actually modulate the function of protein factors and cis-acting elements present further upstream, which is very important.

(Refer Slide Time: 38:26)



Somewhat to give you some example to tell you how actually these core promoter elements can contribute to f differential gene regulation; for example, differential usage of two TATA sequences in the his3 gene promoter of Saccharomyces cerevisiae. Saccharomyces cerevisiae is the (( )) and initially, number of studies have been carried out to understand basic aspects of eukaryote gene regulation using simple eukaryotes like Saccharomyces cerevisiae, Drosophila, and so on and so forth, because they are much easy to grow and manipulate compared to complex eukaryotes like mouse, man and so on and so forth.

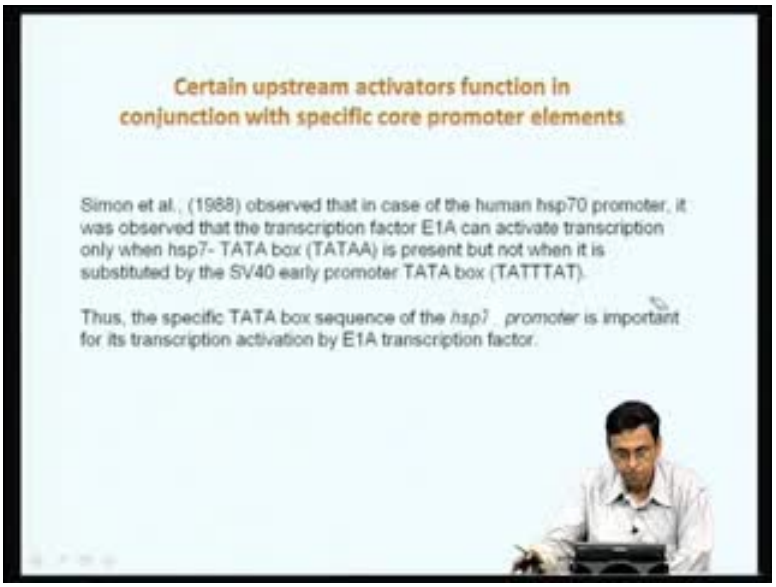
In the case of Saccharomyces cerevisiae, very interestingly, there is a gene called his3 gene, which is actually involved in the biosynthesis of his3. Now, as you know yeast cells, if you grow the yeast cells in a rich medium, where which contains all the amino acids like what we call as YPD, which contains yeast hexatone, peptone, dextrose, etcetera. Then, the yeast cells need not synthesize amino acids, because they are already present in the medium. So, there is no point in expressing many of these genes, which are actually involved in amino acid biosynthesis. So, there is something called a general amino acid control, wherein when the amino acids are already present in the medium, transcription of all the genes involved in amino acid biosynthesis are shut off. This is called a general amino acid control. So, in such cases, these genes are transcribed at very low level or very basal level. But, when you now shift this organism to a medium, which will lack many of the amino acids, then the organism has no choice, but make its own

amino acids. And, in such cases, these genes known as amino acid biosynthesis have to be turned on at very high levels, and therefore, you need a differential regulation. So, what are the conditions? What are the factors of cis-acting elements under which these promoters are transcribed at low level and how these promoters are transcribed at high level? This differential integration, how is it brought about?

What has been show in the case of such one such (( )) his3 is that this promoter actually contains two TATA boxes known as T C and T R. T C stands for constitutive TATA box and T R stands for regulatory TATA box. The downstream the sequence, that is, the regularly TATA box has a typical TATA sequence, that is, the TATAAA sequence. So, there are two TATA sequences: one is called T C, another is called T R. And, the T R, which is the regulatory TATA box, actually resembles the actual TATA box sequence. Whereas, the upstream, that is, the one which is present about 30 base pairs upstream is called as the constitutive TATA. This is actually an AT-rich region of about 30 nucleotides that does not really resemble a typical TATA box sequence, but kind of resembles the TATA sequence.

Now, what is very interesting is that when the yeast cells do not have to make amino acids and when the his3 gene need not be transcribed at every levels, this particular TATA box is used; namely, the T C is used when this gene has to be transcribed at very low levels. Whereas, the T R or the regulatory TATA sequence is actually used when his3 gene has to be transcribed at very high levels. These are the two original papers in which this particular study was carried out, but simply telling, I gave this example just to tell you that by using two different TATA sequences in the same promoter, how differential gene regulation can be brought about. So, you can see all these core promoter sequences are not just sitting there to assemble RNA polymerase; variations in their number and their sequences, minor variations in sequences itself can contribute to differential gene regulation.

(Refer Slide Time: 41:45)

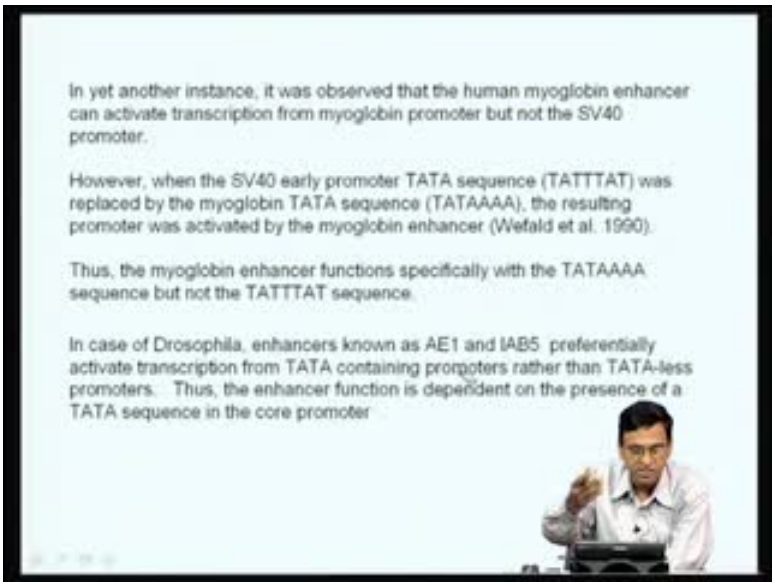


The other example, which I told you, in the case of SP1, which I told you in the previous slide, there are also many upstream activators, which actually function only when specific core promoter sequences are present. For example, Simon et al., observed that in the case of human hsp70 promoter, the transcription of factor E1A, which actually binds to many upstream sequences, further upstream of the core promoter sequences, this E1A protein, which is a transcription factor, can bind and activate transcription only when the hsp7-TATA sequence is present. So, you have an hsp70 promoter, which has a TATA box. You also have a binding site for a transcription factor called E1A. Now, the E1A will bind to this upstream sequence and activate transcription only when you have the TATA sequence of the hsp promoter.

Now, if you replace this TATA box hsp70 promoter from related sequence, that is, TATA sequence from the SV40 early promoter – SV40 is actually a virus, eukaryotic virus, which in fact, eukaryote cells, they actually contain eukaryotic promoter like sequences. Therefore if you now take the TATA box, you can see there is a difference. The hsp70 TATA box is TATAA; whereas, the SV40 TATA box is TATTAT. Both of them are TATA box; they bind to TFIID and so on and so forth; they are involved in (( )) But, if you replace this sequence in the hsp promoter (Refer Slide Time: 43:09) with this sequence, the E1A protein can no longer activate transcription. So, very clearly telling you that these core promoter elements are not sitting there to assemble RNA polymerase in and around transcription start site, but they also function in conjunction with many

upstream activators. And, certain TATA sequences are actually required for the function of certain upstream activators, which bind to far upstream sequences. Therefore, specific TATA box sequence of the hsp promoter is important for transcription activation by the E1A transcription factor. So, you can see variations in core promoter sequences actually determine which kind of transcription factors binding upstream can either activate or repress transcription.

(Refer Slide Time: 43:47)



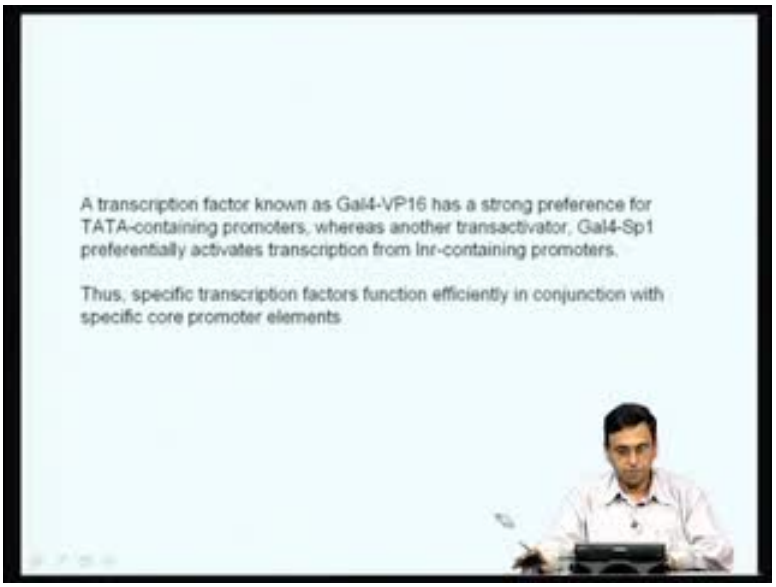
There is one more example. In the human myoglobin enhancer for example, it can activate transcription only from myoglobin promoter, but not SV40 promoter. Now, as we discussed more and (( )), these examples will become more clear, because as I told you, regulation in gene expression consists of a number of promoter elements. On what we are discussing now is only about a very small aspect, namely, what is happening in and around the transcription start site; we are only talking about plus 30 to minus 30 region in and around the promoter, but there are many other sequences and many other protein factors, which bind further upstream, and they contribute about… What I am now trying to tell you is that for these protein factors and upstream sequences to bind and activate transcription from these digital elements, only certain core promoter sequences have to be present. So, there is a link between what is happening in and around the promoter region and what is happening further upstream. So, when the SV40 early promoters TATA sequence was replaced by myoglobin TATA sequence, the resulting promoter was activated by the myoglobin enhancer. I am clearly saying that variation in

TATA sequence is determined; what kind of sequences can function as enhancer sequences, which are present further upstream? So, the myoglobin enhancers function specifically with the TATAA sequence, but not with the TATTAT sequence. These are all very important for you to understand, because when I said TATA box, you assumed that all TATA boxes have only one job; their job is to bind to TFIID and initiate transcription.

But, what I am telling you now is that there are variations in the TATA sequences and these variations in the TATA sequence are very important for the activity of many other protein factors, which are actually present in further upstream. That means there are interactions taking place between proteins, which are binding to the upstream enhancer sequences and these general transcription factors, which are binding to specific TATA sequences. And, to specific TATA sequences, specific general transcription factors are binding. And, what kind of TATA sequences are present actually determined whether a sequence further upstream can actually function as an enhancer sequence or not? So, in this particular case, in the myoglobin enhancer, this digital upstream sequence and an enhancer sequence can actually add to a transcription only when this TATA sequence is present. But, if you replace this TATA sequence with this TATA sequence, it can no longer function as an enhancer sequence.

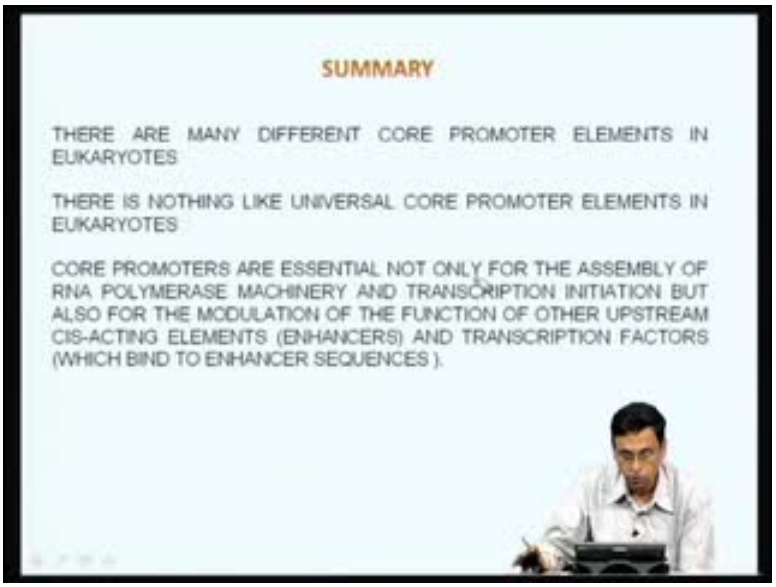
Similarly, in the case of Drosophila, enhancers known as AE1 and IAB5 do not bother about all these terminologies; just, different names have given to identify or study different enhancer sequences, which actually serve as binding stage for many other transcription factors, which bind further away from the promoter sequences. And, these two enhancer sequences preferentially activate transcription from TATA containing promoters rather than TATA-less promoters. So, you can see these enhancer sequences can function as enhancer sequences only from TATA promoters, which contain TATA box. But, when we now remove the TATA box, then no longer function as an enhancer sequence. So, the enhancer function of these sequences is dependent on the presence of a TATA sequence in the core promoter. So, there is link between core promoter sequences and digital promoter sequences, and function of many of the other promoter sequences, which are further upstream from the transcription start site, are actually dependent on the core promoter sequences.

(Refer Slide Time: 46:59)



Similarly, another example – a transcription factor known as Gal4-VP16 has a strong preference for TATA-containing promoters; whereas, another transactivator, Gal4-Sp1 preferential activates transcription from the initiator containing promoters. So, I am giving all these examples just to emphasize the fact that the core promoter sequences are not just sitting there to invite RNA polymerase to come and bind; there whether an initiator is present or whether TATA box is present, this actually determines what kind of upstream activation sequences, what kind of other transcription factors can regulate transcription or not. So, when you have a TATA-containing promoter, transcription factors like Gal4-VP16 preferentially activate; whereas, when you have an initiator motif instead of TATA box, certain other transcription factors called SP1 can activate transcription much better. So, specific transcription factors, which bind to further upstream sequences, efficiently function in conjunction with specific-core promoter elements. This is the message I want to give.

(Refer Slide Time: 47:56)



I would like to summarize so far what I have told you. The crux of what I have told you is that there are many different core promoter elements in eukaryotes. So, do not just think that when somebody talks about transcription initiation in eukaryotes, it is not just the TATA sequence. There are many variations of TATA sequence; TATA box in fact is present only in a small subset of eukaryotic promoters. There are many promoters, which are called as TATA-less promoters. In such TATA-plus promoters, many other core promoter elements like the BRE, initiator, DRE, and so on and so forth take over the function of TATA box. So, there are many different core promoter elements in eukaryotes and TATA box is just one of them.

There is nothing like a universal core promoter elements in eukaryotes. So, do not assume that all the eukaryotic promoters contain only a TATA box, which is not correct. TATA box in fact present only in very small fraction of the eukaryotic promoters; many other variant core promoter elements actually are present in many other eukaryotic promoters. So, core promoters are essential not only for the assembly of RNA polymerase machinery and transcription initiation, but they are also necessary for the modulation of the function of other upstream cis-acting elements or enhancers and transcription factors, which bind to the enhancer sequences actually play a very important role only depending upon what kind of core promoter sequence are actually present.

I gave you a number of examples, which actually tell you that the kind of core promoter sequence, what kind of TATA box sequence is present; whether it is TATAAT or TATTAT; both of them serve as a TATA box. And, a preinitiation complex can be assembled there, but only certain TATA box sequences are essential for the activation of certain transcription factors. And, when you replace by a TATA box to the related TATA box from some other promoter, these transcription factors can no longer activate transcription. So, the two important messages I want to give you in this particular talk. One is core promoter elements are diverse; there are many variants of core promoter sequences and the function of core promoter elements is not just to be involved in transcription initiation by RNA polymerase II and general transcription factors. They also play a very important role. In general, the overall activation of transcription, the variations in the core promoter sequences itself contribute to differential gene regulation. In addition, these different variants of the core promoter elements in conjunction with protein factors, which bind to upstream, actually contribute to differential gene regulation in eukaryotes.

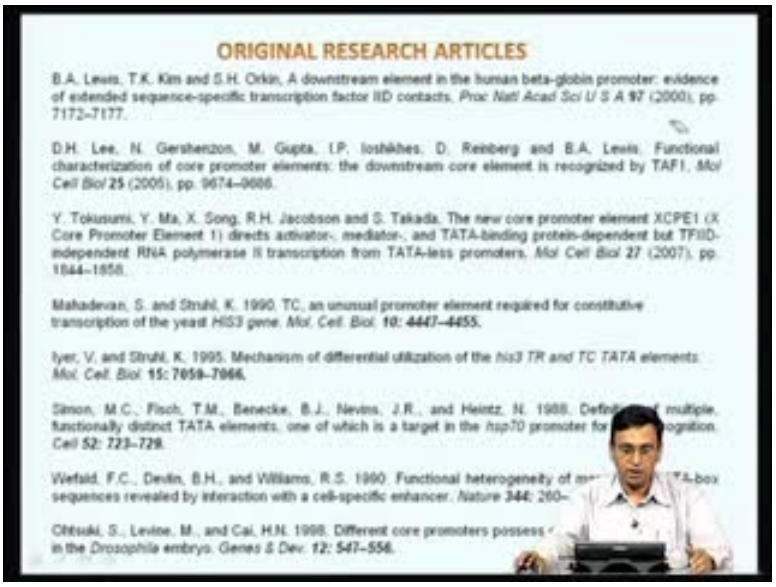
(Refer Slide Time: 50:31)

Now, what I have given in the next two-three slides is actually the original research articles. Wherever I talked about some of these (( )), I have actually given you, for example, I have mentioned you Baltimore's lab discovered this; and, TFIID was discovered by Goldberg and so on and so forth. All these references are actually listed here. So, whenever you want to learn more about some of these actual experiments,

which actually went to the discovery of some of these things, I suggest you go through some of these original research articles, which are listed here. For example, the initiator motif – this is the original research article by a Smaleb and Baltimore, which actually identified the initiator as a transcription control element. So, it is the first report of a identification initiator promoter elements.

Similarly, the discovery of TATA box was actually identified by Goldberg, which is actually part of his Ph.D. thesis. So, these references are arranged in the same order as I told you by discussion.

(Refer Slide Time: 51:29)



Whenever you want to learn more about many of the aspects, which are discussed here, please refer to some of these original research articles, so that you can learn more about how actually, what kind experiments were actually done, what kind of model systems were actually used to study some of these aspects, which we discussed in this…

I think we close now. Of the two lectures we have completed so far now, what I have actually told you, in the first lecture, we actually understood overview of how transcription initiation takes place in the eukaryotes; we discussed how a bunch of protein factors called TFIIA to TFIIH in combination with RNA polymerase II go and bind to the promoter region, and help the RNA polymerase to actually come and bind, and initiate transcription. We also briefly discussed how the RNA polymerase from a

single form of RNA polymerase in prokaryotes evolved the three different RNA polymerases. And also, I told you how a fourth RNA polymerase was discovered recently in plants.

And, in this lecture, I actually told you that these sequences, the core promoter sequences to which the basal transcription factors go and actually bind, are not just present there to form a preinitiation complex and initiate transcription, but there is lot of diversity in the core promoter elements. And, diversity in these core promoter elements actually contributes to differential gene regulation. And, more importantly, these core promoter sequences also determine what kind of transcription factors go and bind to upstream sequences. And, there is a talk or a cross talk between core promoter elements and the upstream activation sequences and coding factors binding to it. And, certain transcription factors function only when specific core promoter sequences are actually present. So, what we will do in the next class, used to actually talk little bit about diversity in the general transcription factors; so far, I have told you how diversity in this cis-acting elements in the core promoter. Diversity in the core promoter sequences actually contribute to gene regulation.

In the next class, I am going to tell you how diversity in general transcription factors, which bind to the core promoter elements, also contribute to differential gene regulation. Do not assume that there is only one TFIID; do not assume that there is only one TBP; do not assume that there is only one TATA binding protein. And, there are variants of these general transcription factors and that also can contribute to differential gene regulation in eukaryotes.

I think I will stop here.