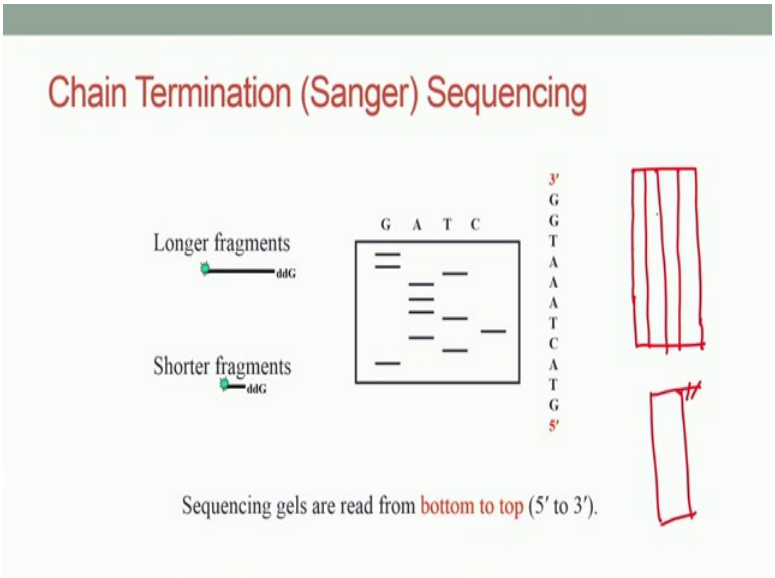
Organic Chemistry In Biology And Drug Development
Prof. Amit Basak
Department of Chemistry
Indian Institute of Technology, Kharagpur

Lecture - 26
DNA Sequencing Method (Contd.)

Let us come back to this session and let me just remind you, where we stopped lat time.

(Refer Slide Time: 00:27)



We discussed the chain termination method by Sanger's technique. Now Sanger's technique is really beautiful, because it is based on a principle that you can terminate a growing DNA chain by using a nucleoside which lacks the 3'-OHgroup.

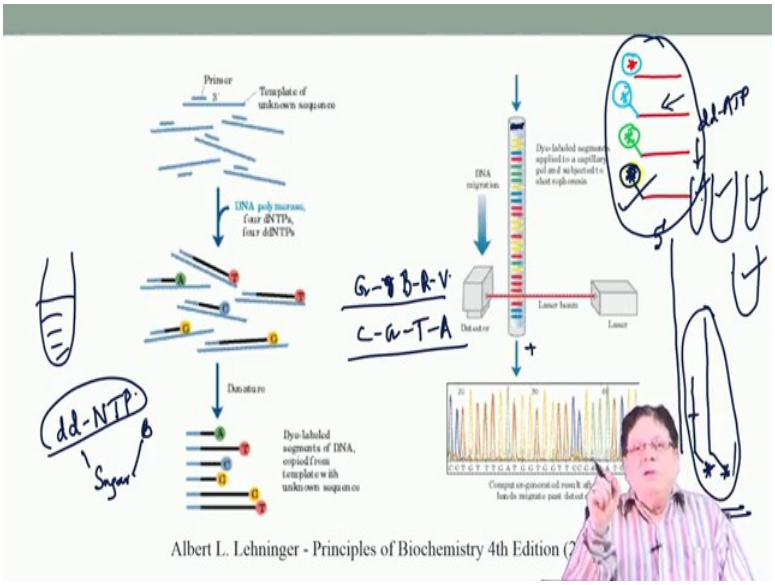
Now the problem with this type of technique is that it is not very quick or rapid. Because, first of all you need to have four lanes in the gel and then in one go, you can determine up to 100 base sequences. Now, suppose you have millions of base pairs in the genome of an organism.

So, it will a huge amount of time. It will take years, before you can really determine the whole sequence of the genome. By the way genome is basically the complete DNA that is present in an organism. Thus determination of the sequence of the entire genome is very difficult using Sanger's technique; because of several limitations. The first limitation is that it can go up to only 100 (maximum) base sequences; and then the

second one is more important; that is that after everything is done, you have to run four lanes for the electrophoresis. These are the two limitations there.

If you run four lanes; your gel size will be more, your applied voltage has to be more. So, it may not be very economic. Thus, apart from time constrain, economy is also important. So, people started looking at other ways using the same principle of dideoxynucleoside triphosphate, but trying to do something which can work in one lane, so that you do not have to do the four lane gels. So, four lane gels need to be replaced by a single lane.

(Refer Slide Time: 03:33)



Now initially what was done? Initially the change that was brought in Sanger's technique was that the primer; it was realized that radioactivity is one issue that people try to avoid; Both Maxam-Gilbert and Sanger's technique used radioactive phosphorus. For Sanger's method the primer had the radioactivity.

So people thought that let us take a primer, which has got a fluorescent label at this end and these fluorescent labels have different colors. Suppose one fluorescent label is red, another is blue, another is green and the fourth one is deep violet. So, these are the four fluorescent labels that are put at the end of the primer. Now what you do? You take that strand for which, you want to know the sequence. Earlier there were four test tubes, now also there are four test tubes. In one test tube, you add this primer. So, this primer will be attached here. And suppose in your first test tube, you are adding ddATP, along with the

all the dNTPs and the DNA polymerase and magnesium. So, what will happen now? Since you are adding ddATP, so the chain will terminate wherever there is requirement of A.

Suppose the chain will terminate here and another chain will terminate here; but the interesting point is that whatever you, whatever truncated oligonucleotides are made when there is requirement of dATP, and instead of that dATP, ddATP is taken; they will all have the same primer with this, dark violet fluorescent label.

So, in the first test tube, you added this primer you added ddATP. So, all the truncated oligonucleotides will show this dark violet fluorescence. Now in another test tube what you do? You do a very similar experiment. You take the primer which is having a green fluorescent tag. In the second test tube, you are adding ddCTP.

So, then all the truncated oligonucleotides will have only green fluorescence containing label, where ever there is a requirement of C. So, all that truncated pieces will all have green fluorescence. Then in this third case, you take the blue fluorescent labeled primer in another eppendorf and now you add ddGTP.

So, you know that for blue fluorescent labeled truncated oligonucleotides, showing blue fluorescence, there must be G that is required at that point, because you are adding dideoxy GTP there, and the fourth one is the red fluorescence, you add the ddTTP; you add the red fluorescent primer and then add the ddTTP.

So what is the outline? After doing all these, that in one test tube, wherever there is truncation of A that piece will show this dark violet fluorescence; wherever there is truncation of C, that will show the green fluorescence; wherever there is truncation of G, that will show blue fluorescence and finally, wherever there is truncation of T, that we show the red fluorescence.

So, after doing all these reactions in four test tubes, you mix all these four contents together. Now, it will have pieces which will show dark violet fluorescence, green fluorescence, blue fluorescence, red fluorescence. Suppose you now put everything together on a gel and run the electrophoresis. This is an electrophoresis gel in a column, you can load in a glass column and then you apply voltage.

So, all these pieces will now come down, because this will be the positive side. So, everything will come down and there is a laser camera which is aimed at this point and then whatever color comes out, it will tell to the detector. This is the detector; when the laser hits a green fluorescent oligonucleotide, it records that colored fluorescent band. So, now as we apply voltage, they will be slowly separated. How they will be separated? That will depend on the length of the truncated oligonucleotide. Suppose you initially get a green fluorescent band, followed by a blue, then suppose red and then suppose the dark violet.

So, so you know the sequence of colors; now you continue the gel electrophoresis. So as the first band comes here, the detector records the color that goes away; the second band comes here, it records the color and then the third one comes and it records the color. So, you do not have to move the detector; only these bands are moving and finally they come out. Only thing that you should do is to detect the sequence of these colors and depending on the sequence of these colors, you conclude the base sequence.

So, this is the actual picture that you will get; that is the chromatogram that you will see. And then from the color of these peaks, you can tell the sequence. Whenever you see a green sequence, you know that must be coming from the test tube where you have used the green primer. So, where you have used green primer? It has been used where you have added the ddCTP. So, there must be a C here then.

Then for the blue primer, you have used a G, so that oligonucleotide must be having a G and then you have a red. The red one implies TTP. So, you have a T and then for the dark violet, you have used the ATP. So, like that, the sequence will be according to the color, color sequence your base sequences will be determined. I hope this is clear.

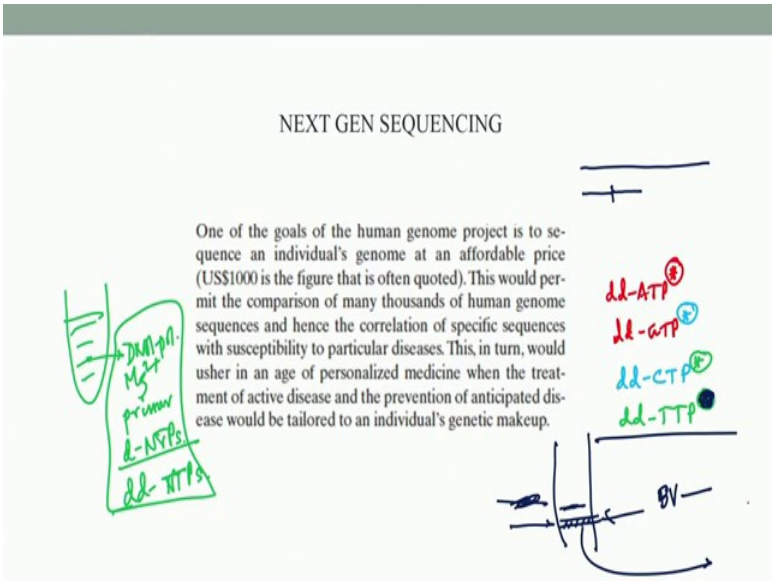
To summarize this technique, you have to use four primers with different fluorescent colors and then you have to do the reactions in different test tubes and you have to add that ddATP like previously. In previous cases, Sanger used the same primer in all the cases, because all are radioactive. Here the primers have different colors, the base sequence in primers are same only the fluorescent, the 5’ OH is attached to a fluorescent labels which give different colors. And then the steps are same; you add ddATP in addition to whatever else is required all the time; and then ddATP, ddCTP. ddGTP and ddTTP.

What is the difference between Sanger's method and this technique? In this method, you mix all these and you can run only a single lane gel electrophoresis. And then you detect the color sequence, and depending on the fluorescent color sequence, you predict the base sequence. In this method, light is required; the laser detects fluorescence as it hits the band and then this records that what is the color that is coming as the truncated oligonucleotides migrate from here to there.

So, this is the next development after Sanger's method. But then people again questioned it that can we simplify it even further? Because one drawback of this method is that you are doing the reactions in four test tubes or four eppendorfs. So, can we do the reaction in one container and then do the sequence technique?

So, the next challenge is to do the reaction in only single eppendorf and then do a similar kind of assay; but if you use the primer of different colors, then you have to use different test tubes or different eppendorfs. So, then somebody thought that the best way to do these reactions in one test tube or one eppendorf is to you use the ddNTPs (dideoxynucleoside triphosphate) having fluorescent labels NTP has a sugar and base with lot of reactive nitrogens. So, what you can do? You can put fluorescent label in the base of the dideoxynucleotide triphosphate and these fluorescent labels are different for different bases.

(Refer Slide Time: 15:35)



That means now instead of the primer being colored differently, which means fluorescently label differently, now the primer is same, but your dideoxy bases have different colors. Suppose ddATP is colored with a red fluorescence. Then ddGTP, has got a blue fluorescence. I am just arbitrarily giving some colors. Then you have ddCTP has the green and ddTTP has the dark violet fluorescence. So, instead of putting the color on the primer, now you put the fluorescent labels on the dideoxy nucleoside triphosphates; thus these labels are in the bases.

Fortunately the DNA polymerase does not discriminate; it accepts the fluorescent tagged ddNTP as a substrate, although the size is bigger because some of these fluorescent labels are quite big, but the DNA polymerase is not that selective, it accepts even if the base has some handle which is a fluorophoric handle. Now you can do the all the reactions in the same test tube.

So, you have the primary sequence of DNA and you do the reaction; that means, you are adding all the dideoxy nucleotide triphosphates. So, this contains your DNA polymerase, magnesium, the regular primer without the label, all the dNTPs and finally it contains all the ddNTPs.

When you do the reaction, what will be the outcome? Now you will again have truncation; the oligonucleotide upon picking up a ddATP will be truncated. If the truncated portion is red; that means, there was the requirement of A at that time. When the truncated piece is showing a blue fluorescence, that means, there was a requirement of G at that time.

And then the green fluorescence means there was a requirement of C and then this violet fluorescence indicate that a T is required. So, now, what will happen? You can do all the reactions in the same test tube and again run the same single lane gel. Then you just check the color of the bands that are coming through the column containing your agarose gel.

So, by just checking the color sequences, you can immediately write the base sequence. Now all these things are computerized. So, the computer will see the colors that are constantly fed to the computer and the computer already knows that this color means this base, so it will immediately write all the base sequences.

So Sanger used the radioactive primer, then the improvement was that using a primer containing the fluorescent label, but that required the reactions to be done in different test tubes or eppendorfs and then the subsequent development was doing the reaction in one container (either the test tube or eppendorf), but using the deoxynucleoside triphosphates which are differently labeled with fluorescent markers.

Then what you can do? You can quicken the process. , One lane gel will be much more rapid; more number of bases can be sequenced; if you have radioactivity then you have to take the photograph of that, a photographic plate has to be put on top of the gel and the next day you have to take the print of that, you have to develop that.

But when you have these fluorescent labels, you do not care whether something has passed through. Suppose this color has passed through, that goes away into the solution, you do not need to know what goes away in the solution; what you need to know is that what comes after what (basically the sequence of colors). If this blue violet band comes first, then the second one which was following it was yellow, then that will come here and then the detector records the color and then all these go away.

So, you can actually sequence much greater number of bases in one attempt. About 600-700 base sequences can be done by this technique. Now why this became important? As I told you that the genome mapping of different organisms of different living species was taken up; scientist wanted to know that why a cat is different from a dog; or why a dog is different from a man; there must be differences in their genome sequence, because everything is ultimately dependent on the base sequence of the DNA that are present in the cells.

So, that is why this became a very important issue and a big program was taken which was called the 'Human Genome Project'. So, they attempted to map the entire human genome. Now how many base pairs are there in the human genome? Around 3.2 billion base pairs.

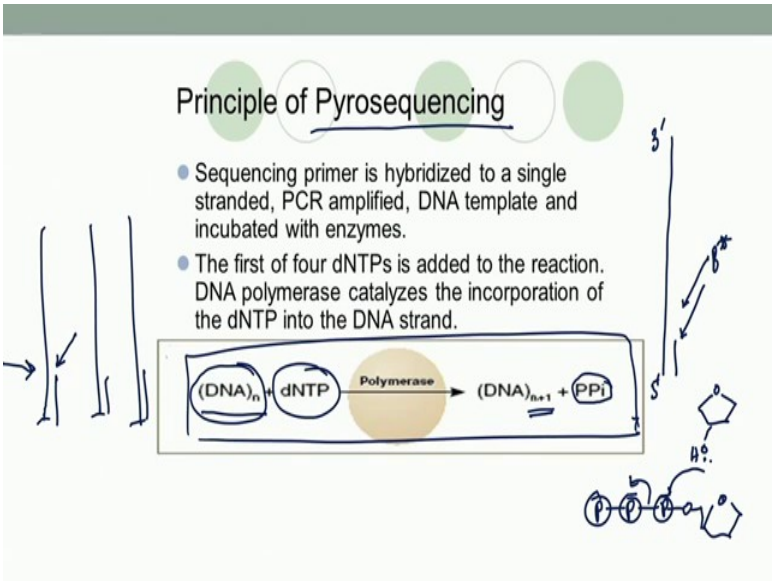
So, if the Sanger's method would have been followed to sequence the entire human genome, it would have taken 10-12 years to complete that or maybe more. However, after all these changes, this is one project, which was finished much earlier than the predicted date. This could happen because of these new developments that took place in the late 20^{th} century.

Why this became very important? We know that the same drug does not work equally for all of us. A drug may cause acidity to some person, whereas the same drug works very well for another person. What does it signify? That means, every person has some differences somewhere in their gene with respect to the other person. If you know those differences, then you can develop, what is called the personalized medicine.

So, today we are in the era of personalized medicine; that means, if your gene sequence is known and then the doctor can say that this drug is going to work for you or this drug may not work for you. So, now, this is person specific, although we are not right there, but now in the in the western world there are cases where personalized medicine is in use. Specially when they treat cancer, they actually adapt this strategy; they determine the gene sequence of that person very quickly, and then compare that with a healthy individual and then immediately they can find where the problem is.

So, this type of medicine has to be given to that person. This is called personalized medicine and that is only possible after the development of the human genome project.

(Refer Slide Time: 24:37)



All these methods that I have told you so far, the fluorescent based methods, that also take took 3-4 years to complete the process of DNA sequencing, I give you some statistics here. For human genome project, the goal was to sequence the individual genome at an affordable price.

If you want to take benefit of this personalized medicine, you have to analyze the whole gene sequence of your body, but that should be at an affordable price; if it is really very expensive, then it is very difficult. US dollar 1000 is the figure that is often quoted. This would permit comparison of many thousands of human genome sequence and hence the correlation of specific sequence with susceptibility to particular disease. I said that different diseases have different gene sequence.

But now they want that the human genome sequence should be known very quickly, because when somebody is suffering from some terminal illness, you need to know the sequence very rapidly, so that the proper medicine can be given and also it should be at an affordable price. It cannot cost thousands and thousands of dollars; that is number 1 and number 2 is that it cannot take months and months or years to know the sequence, because somebody who is terminally ill, he might have only 2 to 3 months.

So, by that time you have to know the gene sequence and start giving the proper medicine. So, now, it is the era of Next Gen Sequencing (next generation sequencing). Now, we have more rapid method of sequencing, even within 4-5 days, the mapping the base sequence can be completed. In Next Gen Sequencing, there are different methods. I will not describe all the methods. I will just take one of the methods of that next Gen Sequencing.

There is a method which is called the Illumina method that is developed in Cambridge. In earlier methods, whatever we have said, they are not real time determination of the base sequence. What is real time determination of base sequence? When the dideoxynucleoside triphosphate is added, or any other correct base is added, at that time you cannot determine the result. You have to wait till all the reactions are over and then compile them and do the electrophoresis. That is not real time because the you first do the reactions, then you do the electrophoresis and then come to the result.

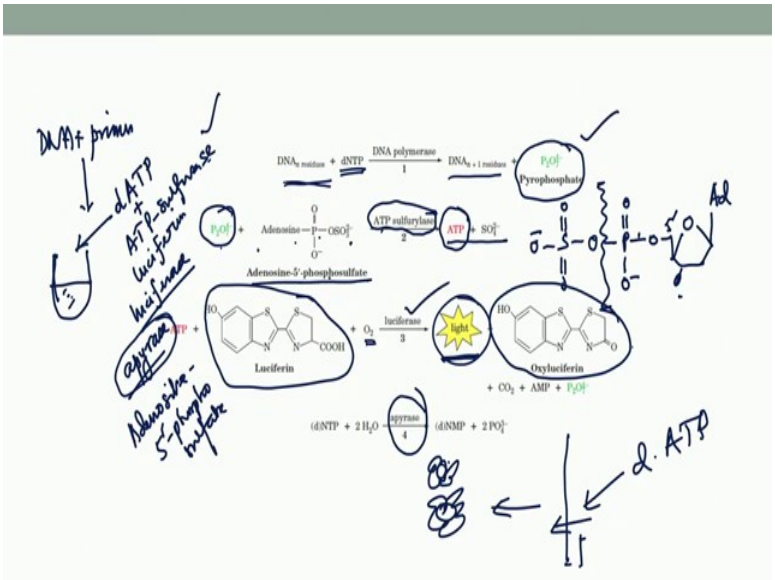
Under the real time analysis, as soon as the base is taken, you get to know which base has been taken. What I am saying is that under real time analysis, whenever a base is added, you immediately know that which base is being taken and whenever the next base is added you immediately get to know what is added. This is called real time determination.

This appeared to be very difficult at the beginning. The base is having this fluorescent marker and you have to amplify these strands. So, you have many strands suppose and the base has a fluorescent marker. Suppose the there is a requirement of A. So, you do not add the dideoxy NTP here; if there is requirement of A you attach a fluorescent label to A and you should have a detector which is extremely powerful that whenever A is added there is a constant shining of the light at this point. So, A is added and you know that what type of fluorescence you are getting and from that color of the fluorescence, you can determine the sequence.

But today I am just concentrating on another method which is called pyrosequencing. Pyrosequencing is a method based on this simple chemistry. I told you that the DNA polymerase joins the growing oligonucleotide to a new nucleoside triphosphate. So, for DNA synthesis, what you need is the oligo nucleotide which should have a 3' OH and which should have a triphosphate.

So, this is your the chemistry that is happening. This 3' OH is attacking this phosphate and pyrophosphate (PP_i) goes out. A diphosphate or a pyrophosphate (PP_i) is an inorganic phosphate, no organic, no carbon is there. So, an inorganic diphosphate comes out. If the DNA had n number of nucleotides earlier, after this reaction, you have n plus 1 number of nucleotide and one pyrophosphate molecule is generated.

(Refer Slide Time: 31:33)



So, the first reaction is that you have a DNA residue, you added the dNTP. So, a pyrophosphate $P_2O_7^{4-}$(pyrophosphate) is formed. It has been found that there is a compound called adenosine-5'-phosphosulfate; that means, you have adenosine; and you have a phosphosulfate at the 5'-position.

. There is an enzyme called ATP sulfurylase. What is sulfurylase? It breaks this P-O bond and puts the pyrophosphate here. If you put the pyrophosphate that means this becomes triphosphate and the sulfate comes out.

So, the reaction is pyrophosphate plus adenosine phosphosulfate in presence of the enzyme ATP sulfurylase gives ATP plus sulphate. So, a molecule of high energy is generated which is known as ATP. Now this ATP reacts with a compound called luciferin. Luciferin is a compound which is present in fireflies, which gives light, light comes out of the insect which is called the firefly.

So, this luciferin in presence of molecular oxygen and in presence of this enzyme which is called luciferase, it reacts with ATP and then forms a compound which is called oxyluciferin; this is its structure and it generates light. So, in the pyrosequencing, you take your DNA strand, you add the primer, that is requirement.Suppose you have added dATP and then you have added ATP sulfurylase and you have added luciferin; and you also add luciferase that is the enzyme which generates light when luciferin reacts with the ATP in presence of oxygen. So, light comes out.

There is another enzyme that you have to add which is known as apyrase,. Now I have told you what is the function of this ATP sulfurylase; the function of ATP sulfurylase is to generate ATP from pyrophosphate. Actually you have to add that adenosine-5'-phosphosulfate in the same test tube.

So, you have added all this. So, first the $P_2O_7^{4-}$ (diphosphate) is generated then it reacts with this adenosine-5'-phosphosulfate, in presence of ATP sulfurylaseto generate ATP. As soon as ATP is generated, luciferin reacts with the ATP. In presence of oxygen and luciferase, it generates light and then there is an enzyme apyrase. Now you have added dATP to start with. Suppose your DNA that you want to sequence does not require dATP to start with, maybe it requires a dGTP to start with, but you do not know which one is the first one.

So, you added first dATP; suppose it does not react; if it does not react; that means, there will be no generation of pyrophosphate. Now before you add the dGTP, you have to break down this dATP. So, this apyrase is an enzyme which has the ability to break down these dNTPs. That means, if unused, then these dATP or dGTP or dCTP or dTTP will be broken down into nucleoside monophosphate (harmless products) by apyrase.

So, basically what happens now as you have added dATP, if it reacts then you will get some light; if it does not react then what will happen? It will be broken down by apyrase into some harmless thing. So, after some time, you add the dGTP, if that reacts, then you will get light. If that does not react, that will be broken down by the apyrase; and then if you add CTP, if it does not react that will be broken down.

So, you will get no light in case of C and in case of the next one say TTP, either you will get light or you may not get some light. So, basically the whole instrument measures the light. It actually measures how much light is generated, when you are adding deoxy ATP, deoxy GTP, deoxy CTP or when you are adding deoxy TTP.

(Refer Slide Time: 39:13)

So, if there is no light; that means, that base is not required at that moment. If there is light; that means, there is a requirement of that base at that moment. If there are two consecutive Gs which are required, as you have added your dGTP, both will be incorporated one after another, and you will get a light which will be having twice the intensity as that obtained for the requirement of only one GTP. Ultimately, you have this

kind of a graph. In this direction (along X-axis), you have the bases that you are adding one after another and on this Y-axis, you have the intensity of the light.

So, how to know what is the sequence? You have added G here and that is the light intensity that you got. Remember the light comes from the generation of ATP, which reacts with luciferin to give oxyluciferin, and that happens in presence of luciferase enzyme to gives light. The intensity of the light will depend on how many Gs or how many Cs are required. So, now, after G you add that dTTP and you see that the light intensity is double with respect to C; this means two Ts are there. Then you have added C with practically no light intensity. That means we do not have a C after T. Then addition of dATP yielded double intensity light, that means two As are there.

Light intensity upon addition of dCTP is three times with respect to your base value. Base value is basically the intensity obtained when one of the bases are incorporated. So, now, it is easy to read the sequence of the DNA. This is G then the T has twice the intensity of the light that is emitted when you have added the dGTP.

So, that will be two Ts; that means, there are two Ts because you have double the intensity; then there is no C, because you have not got any light then there are two As; then there is a G, one G because the light is the base value of the light that you get then there is no T, then there are 3C s. So you may have the following sequence: It is G T T A A G C C C A T A A A C C G C C A.

So, that is the way to read this diagram. This is what is called pyrosequencing and I say there are other methods in Next Gen Sequencing. All are direct methods; that means at the time when it is added, you are analyzing the product and then your computer will ultimately tell you what is coming out and finally what is the sequence of the bases.

Thank you.