

**NMR spectroscopy for Structural Biology NS**  
**Prof. Ashutosh Kumar and Prof. Ramkrishna Hosur**  
**Department of Chemistry**  
**Indian Institute of Technology - Bombay**

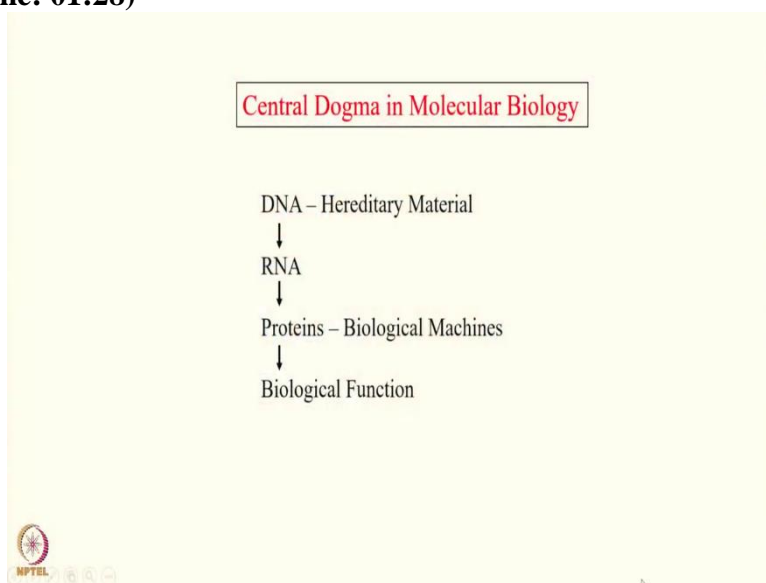
**Lecture: 24**

**Application of NMR in the Area of Structural Biology: Structure of DNA and RNA 1**

For so long we talked about NMR methods various kinds of techniques which are developed over the years and now we are ready to go into the applications of these methods into the area of structural biology which is the main theme of this of this course NMR in structural biology. So, we have prepared the ground with methods. So, now we are going to see how they can be applied in structural biology.

So, therefore we will have to prepare a little bit of ground for the biology itself what do we mean by structural biology what sort of molecules we are going to look at. So, that will be the theme now onwards. So, we are going to spend a little time on introducing the different concepts and different biological molecules. And everybody knows that especially the biologists they will know that the central dogma in molecular biology is described in this particular manner.

**(Refer Slide Time: 01:28)**



Every cell has a DNA which is their inside the nucleus and this is the hereditary material. So, different individuals have different DNAs. So, therefore the hereditary comes because of this molecule. So, the offsprings will be similar to the parents and this is because of the similarity of the DNAs and therefore it is in fact used to identify people on the basis of their DNAs.

And then DNA is then transcribed into another molecule the DNA stands for deoxyribonucleic acid. So, nucleic acids are the other kind of polymers and then these ones are transcribed into another kind of a nucleic acid which is called as RNA this is the ribonucleic acid. And this actually is the process called transcription and different segments of the DNA are transcribed into a certain molecules RNAs we will describe that a little bit later.

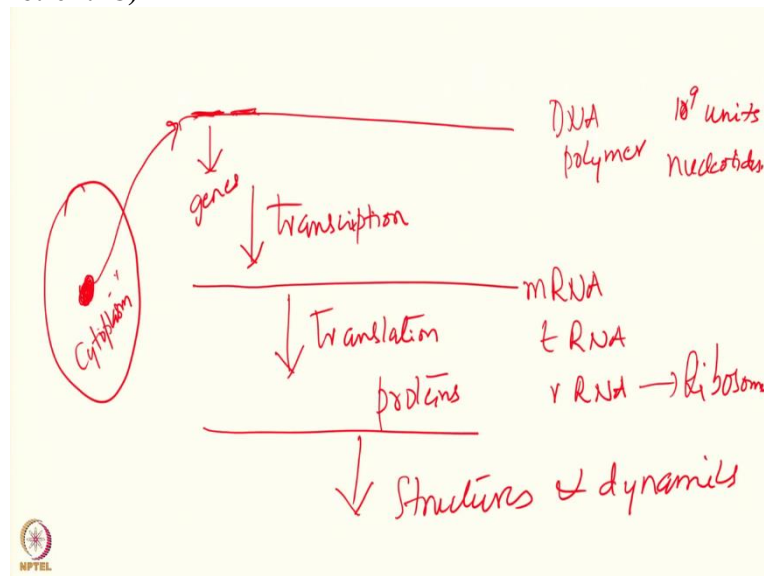
And those segments which are transcribed are called as the genes in the DNA and then they form RNA and this comes as what is called as the messenger RNA this process is known as

transcription. So, this process is known as transcription and from the messenger RNA you have what are called the proteins. So, it is there is a kind of a code in this which is called as the genetic code and this genetic code describes what kind of a protein has to be produced.

And you know the proteins are made up of amino acids there are 20 different types of amino acids and a set of 3 units of RNA codes for one particular amino acid and the of the 20 different amino acids are coded by different codes these are called as the genetic codes. And then this process of expressing the information in the RNA into the proteins this is a process which is known as translation.

And then proteins are actually the biological machines actually for a long it was thought that proteins are the biological machines. But today we also know that sometimes RNAs can also carryout function biological functions but the general belief has been for men which is true that that proteins are the real biological machines and they carry out the biological functions in inside the cell.

(Refer Slide Time: 04:28)



So, now to explain this little bit more. So, you have the DNA which is a long polymer which is a long polymer it has more than a billion molecular which is a huge molecule DNA this is the polymer and it is very long. So, it is approximately  $10^9$  units and these are called as nucleotides. So, on this DNA there are certain segments small segments and these are called as the genes.

There is a lot of other things also inside the DNA and these are the genes and from the genes you get what is called as the RNA and this is called as the messenger RNA. There are 3 different type of RNAs they are messenger RNA then the tRNA and then you have the ribosomal RNA. So, if I went to draw the cell here you have the nucleus and this contains the DNA.

And the mRNA and these things that actually happen inside this other thing which is called as the cytoplasm here. So, the genes are expressed in the form of mRNA this that is a particular mRNA there are other kinds of RNAs which are called tRNA and ribosomal RNA and tRNA is involved in the translation process from here. Here this is then it goes to the proteins here this is the process of translation.

And from here to here this is the gene this is the transcription and the translation generates the proteins. So, now ultimately our objective will be to understand the structures and how these processes happen. Of course this is very complex there will be many different kinds of proteins involved in every step and it is an extremely complex process. Clearly there is a kind of synergy in all these actions of these various proteins and the nucleic acids.

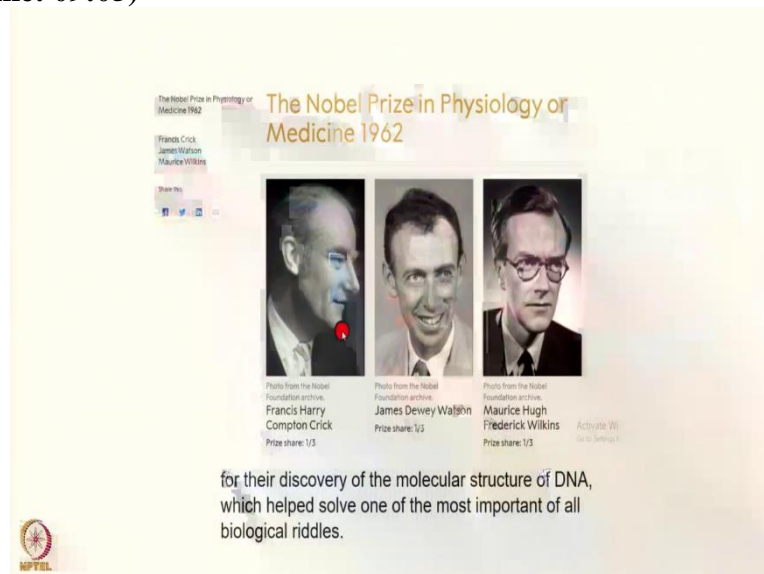
And one needs to understand this energy in order to understand that we have to understand the structures, and the dynamics of these molecules. So, this actually constitutes the main theme of structural biology and protein there are also other molecules inside your cell and these are small molecules which are the source of energy or metabolites. So, the food you eat is actually degraded into various kinds of small molecules and these are the markers you have the ions and things like that.

So, but biologically we will be concerned with these macromolecules which are the DNA segments then you have the RNA segments then you have the proteins. The ribosomal RNA is present inside the what are called as the ribosome's and these are the machines where the actually the proteins are synthesized. So, these are the areas these are the kind of powerhouse where the proteins are synthesized mRNA comes inside the ribosome's and then there is a tRNA which also plays a role inside there.

And the ribosome also contains many proteins therefore the whole thing is a big complex is a big assembly and the structure of this ribosome is also an important aspect to study. So, individually one looks at the molecules of the nucleic acid segments mRNA segments the tRNA segments and the proteins. Now this idea of this process of understanding the structures was of course an important topic and several people had worked on it.

And there are various kinds of empirical rules which had evolved and on the basis of what is called as a fibre diffraction, fibre diffraction data.

**(Refer Slide Time: 09:05)**



These 3 people these 3 people brought out a model of the DNA the structure of the structure they proposed a model of the DNA structure and they got the Nobel prize for that in physiology or medicine in 1962. See for the discovery of the molecular structure of DNA which help to solve one of the most important of all biological riddles that is the replication process. Why is it a hereditary molecule how is it that the offsprings have similar features as the as the parents.

So, why is the DNA so important what is the structure how is the information transferred from one from the parent to the offspring and that is retained in the same manner. So, that was an important riddle in biology and these people gave an answer to that in the form of the structure of the DNA molecule. So, what did they propose they showed that the DNA is a on the basis of the fibre diffraction data.

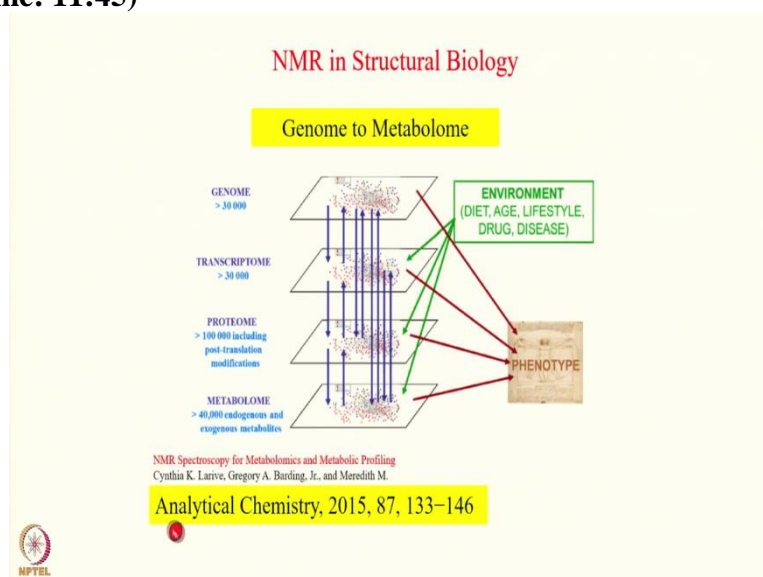
So, they showed that the DNA is actually a double helix. So, is wound like this then you have another one which is also going parallel I cannot draw this properly here but this is a double helical structure which they proposed and that actually became an important contribution for understanding double helix. There are certain rules as to how the double helix is formed we are going to look at all of those ones.

But they based their arguments on the basis of the fibre diffraction data. Fibre diffraction does not give you such very high resolution at the atomic resolution it does not give that information it only gives the kind of a symmetry elements what are the kind of symmetry elements in the in the molecular structure. So, either in any proper symmetrical spacings of certain items pattern spacing of certain groups what kind of structural elements may be present inside the big polymer.

Because you remember this is a very big polymer and in this we have these 2 people Francis Crick and Morris Wilkins these were the crystallographers the diffraction the ones we interpreted the fibre diffraction data and Watson was a chemist who actually built the model of this using ball and stick models. So, the and then actually turned out to be such an important contribution that they got the Nobel prize in 1962.

So, now we have to go into understanding the structures of these molecules in greater detail. So, before we actually go to that let us look at this various biological function in a light little bit more detail and that is indicated here.

**(Refer Slide Time: 11:45)**



So, here now here also it makes an introduction to NMR because there is an article written here in analytical chemistry for metabolomics and metabolic profiling and it also has describes what the various elements in structural biology. So, you have the genome to metabolome. So, this is

the genome the DNA that is the DNA which is also called as the genome here and there are typically about thirty thousand genes inside the DNA.

In a particular typically in a living cell there are about 30000 genes. Now these 30000 genes are they are characteristic of the individuals that define what is called as the phenotype and then you have of course phenotype all the things also contribute to the phenotype. Now this genome is transcribed into what is called as a transcriptome. So, you have 30 000 genes which are expressed and produced 30000 mRNAs.

So, these are 30000 mRNA this is called as a transcriptome those this is very characteristic of every individual every cell has its own transcriptome therefore this becomes an important thing to study what is the if process here from here to here and then from these 30000 transcriptomes you produce nearly 100000 proteins. So, which means that one mRNA is producing more than one protein in other words there is what is called as a mRNA splicing.

So, therefore different combinations of the mRNAs are made here and not only they produce individual proteins but some kind of alterations in the mRNAs and join in cutting them joining them together things like that and that results in producing nearly 100000 proteins in their particular cell and this is called as the proteome. And also certain changes happen after the translation some modifications happen and those are called the post translational modifications.

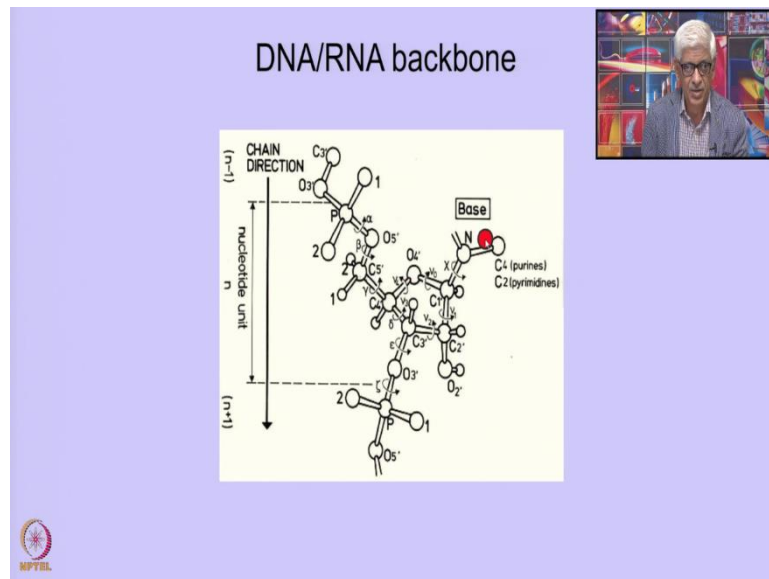
These are important events that happen in biology and they are required for particular processes in the biological function. And then from here you have this metabolome the proteome deals with the metabolome that is the various kinds of molecules small molecules which are interacting with this. So, the metabolites are those which are the food you eat which is degraded into small molecules.

There are ions and there are nearly 40000 small molecules which are actually present in a cell and is these are upregulated these expression levels are up regulated down regulated and things like that depending upon the disease conditions depending upon the nature of the individuals it depends upon the environment diet age lifestyle drug disease and therefore one has to target various molecules in this category.

So, you can sometimes you may have to target here sometimes here sometimes here sometimes here and therefore all this contributes to the research in drug development disease control and these are dependent on the environment the lifestyle age etc. So, therefore it is a huge topic to study all of these ones there are many challenges. And of course now the NMR has come of age to provide solutions to many of these situations.

So, in order to see how one can apply NMR we will have to go into the details of the individual molecular structures.

**(Refer Slide Time: 15:05)**



So, let us begin with the RNA structure of the DNA structure. As I said their DNA is a polymer it is made up for nearly 1 billion units and what are these units these are called as the nucleotides. So, the nucleotide is described in this manner it is a phosphodiester linkage here this has the sugar ring here this is the sugar ring this is a 5 membered sugar ring as you can see this is a and then you have a certain kind of a base here.

There are 4 different types of bases which we will describe soon sometimes 5 and if in the DNA has a particular kind of a base called the thymine whereas the corresponding one in RNA is called uracil it is and then you have the phosphate group here and this phosphate group this continuous this is the monomer unit of the DNA and this repeats itself this repeats itself through the polymer what changes in this the only thing that changes is this base.

From one unit to another unit there is a change in the base nature of the base and that is called as the DNA sequence. As you go from one unit to another unit DNA base changes and that is the one which is the information carrier. The sequence of the bases that are present in the DNA that is the information carrier and the div this has a certain chain direction here. So, we represent this as the chain direction going from  $n - 1$  and  $n$  and  $n + 1$  it goes in this direction. So, therefore we write it as 3 prime.

So, this is the 3 prime and this is the 5 prime end here see these are atoms are labelled in this manner the sugar ring sugar ring is labelled in this manner this atom is called the C1' the C2' C3', C4' and C5'. And then the C5' of this unit is attached to the phosphate group here this phosphate group and this phosphate group is attached to the C3' end here at this point.

Now you see there is an oxygen here all of them have an oxygen and this C3' attaches with through the phosphate group to the 5' end of the next nucleotide and therefore this is the way sequence goes on. So, this is the 5' one end is the 5' end other end is the 3' end the chain goes in this direction 5' into 3' end. So, the nucleotide one particular nuclear unit this is N this is in this is from one 5' end to the other 3' end.

So, include the phosphate group of this into this unit and the next one of this goes to the next unit. So, therefore this is the 5' to 3' sequence. Typically when you write the DNA sequence you write in this manner. So, in go for going from 5' to 3' what is the DNA sequence you write

the various bases which are present in this nucleotide unit and that is called as the DNA sequence.

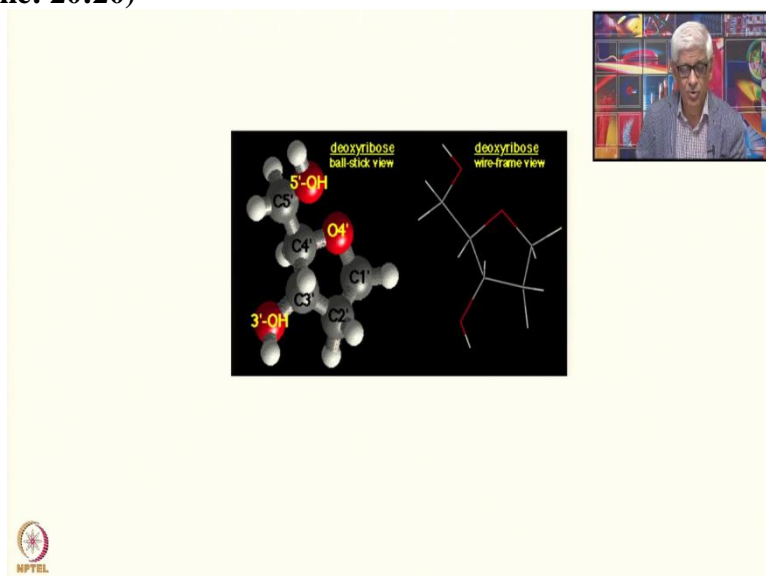
So, this is the monomer of the of the entire polymer there are about 1 billion of this kind of things. Now you can imagine how much how long this should be. So, if you consider the bonds here number of bonds here. So, from here to here, how many bonds are there. So, this is 1, 2, 3, 4, 5, 6, 7. So, these many bonds are there so typically length if you consider the length of each bond is approximately let us say 1.5 angstrom then the total will be approximately 10 angstroms from here to here it will be really 10 angstroms.

Now the 10 angstrom into 1 billion so, this is  $10^{10}$  angstroms then you can imagine how long this will be if you take when  $10^{10}$  angstroms and how much is one angstrom one angstrom is  $10^{-8}$  centimeters. So, this will be  $10^2$  centimeters that is about one meter  $10^2$  to about 1 meter right. So, the DNA length will be approximately  $10^2$  centimeters if it is completely near linear it goes like that it is completely if it is around one meter.

However you see the size of your cell is just about a micron  $10^{-6}$  centimeters. So, then and then the nucleus is even smaller than that now how is this entire DNA packed inside that small space inside the nucleus. So, which obviously means that the molecule adopts a very folded structure all kinds of folding things happen there and it is completely wrapped around in various ways and so, that it is packed inside the nucleus.

And this has to be unpacked when it has to be expressed when the protein has to be and the mRNA has to be produced it has to be expressed. So, it has to be unpacked and then the proteins had to be produced. Therefore you can imagine how complex this whole process is and a lot of this and lot of research therefore goes on in trying to understand all of these. And how can NMR contribute in this is what we are going to see.

**(Refer Slide Time: 20:20)**



Now this is a little bit of an explanation in somewhat more detail at atomic level. So, you can see here the atoms are listed here where one prime tube these are the carbon side this is the ribose ring this is the ribose ring here see the 1', 2', 3', 4', C4' and then the O4' this is an oxygen here. And then at this point you have a 5' carbon, carbon is added to a phosphate group here.



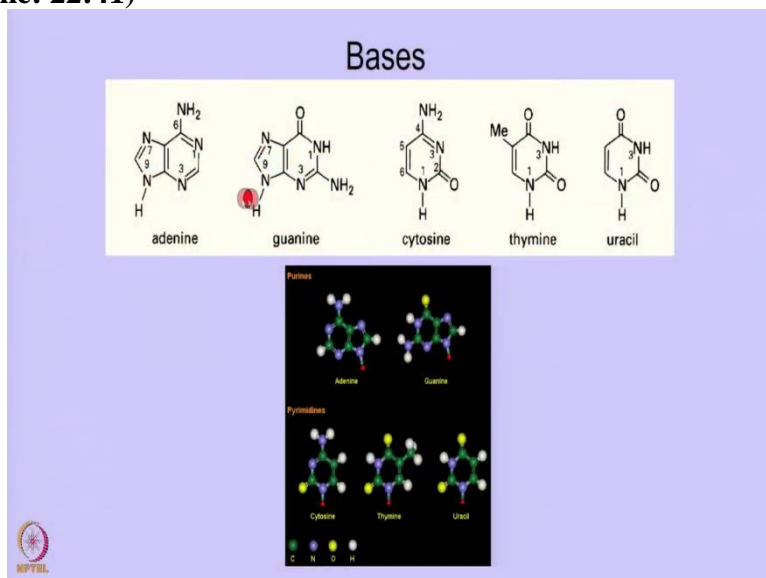
So, they have a 5'OH group and that current corresponds to the phosphorus and in the deoxy this is a deoxyribose. The deoxyribose means that at the C2' position here there are 2 hydrogens these 2 are hydrogens here therefore this is the CH<sub>2</sub> group the CH<sub>2</sub> group is there is a hydrogen and the C1 prime this is actually connected to this oxygen and this C4 is also connected to this oxygen and CO' C1' here is connected to the base.

From here the base goes and then the C3' is connected to an oxygen and in the case of the RNA there will be small difference here in the case of RNA you will see that the 2' position here also has an oxygen. So, there is one hydrogen and one oxygen in the case of RNA that is a ribose nucleic acid. So, therefore here in the C2' prime in the case of DNA there are 2 hydrogens these are labelled as H2', H2'' and in the case of RNA you only have H 2 prime you do not have H2''.

And all of these also have a proton on this there is also a CH<sub>3</sub> on this CH<sub>3</sub> there is a proton on the C3 there is a proton which is called H3' and on the C1' also there is a hydrogen which is called H1' and likewise on the C4' there is a hydrogen this is called H4' and on the C5' there are 2 hydrogens which are indicated here and these are the two H5' and double prime that is better seen in this picture here.

So, H5' and H5''. So, therefore there are a number of protons. So, therefore one can study using proton NMR what is the kind of structure of these individual nucleotides are. Now what are the bases? As I said the one prime is connected to the base.

**(Refer Slide Time: 22:41)**



Now there are 4 different kinds of bases or there are other 5 here see this is adenine, guanine cytosine and thymine these are the 4 which are present in DNA. In RNA thymine is not present it is replaced by a uracil. Now you see these are all these 2 are called purine rings. So, this is the purine structure here you have a 5 membered ring in a 6 membrane joint here this is the purine structure.

And these are only the 6 membered ring these are called as pyrimidines. So, the basic energy in organic chemistry this is the pyrimidine here and these are 2 purines and the uracil is also a pyrimidine and it has 2 oxygens here and same is true with the thymine. The difference between



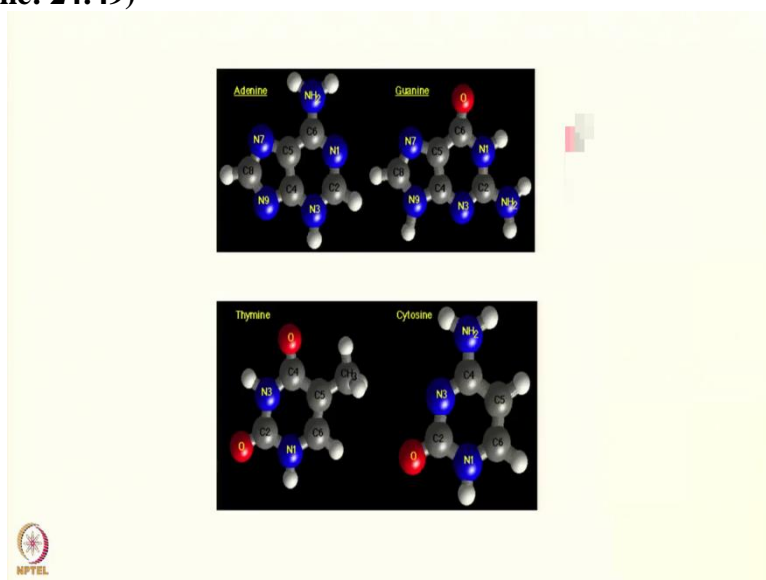
thymine and uracil is that you have a methyl group in the case of a thymine whereas in uracil there is a hydrogen here at this point there is a hydrogen.

And all of these there are hydrogens here in this point there is a hydrogen here there is a hydrogen here there is a hydrogen and so, on so, forth. So, these are the structures of the bases and those are indicated in a colour coded man here. So, all the nitrogens are blue here and the oxygens are in red and the hydrogens are in white and the carbons are in green. Notice here there are  $\text{NH}_2$  groups here the cytosine has  $\text{NH}_2$  guanine has  $\text{NH}_2$  and adenine has  $\text{NH}_2$  all of these are hydrogens and it is 2 groups.

And they also have iminoprotons the guanine has an amino proton that is the  $\text{NH}$  group here and the thymine also has a  $\text{NH}$  uracil also has a  $\text{NH}$  therefore. So, far the nitrogen is concerned there are amino groups and the iminogroups and there are also some nitrogens which do not have a proton attached. So, the cytosine has it is at this point actually these ones are attached to the carbon these  $\text{NH}$  is actually will get this hydrogen will disappear and it will be attached to the sugar ring.

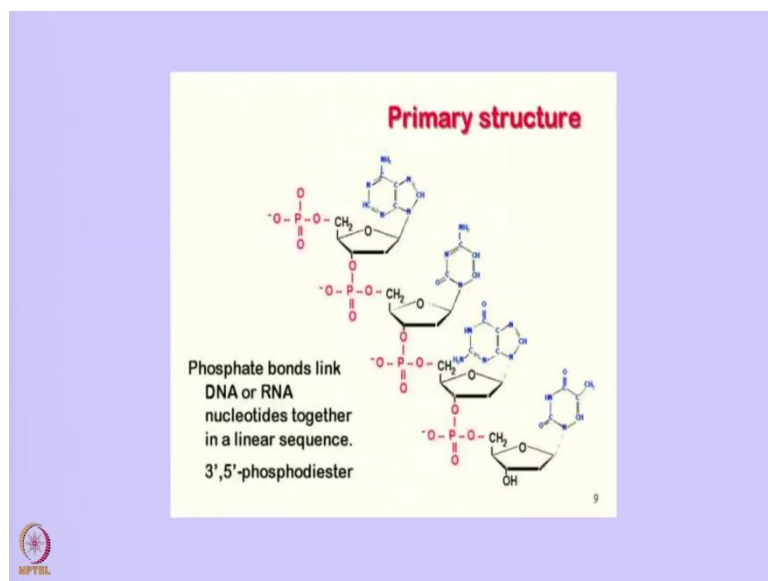
Similarly and this will be attached to the sugar ring base will be attached here and the sugar will be attached here and like that here. So, this is how the structure of the DNA and the RNAs are built.

**(Refer Slide Time: 24:49)**



So, this is the same molecular structure of the basis given in a more explicit manner.

**(Refer Slide Time: 24:57)**



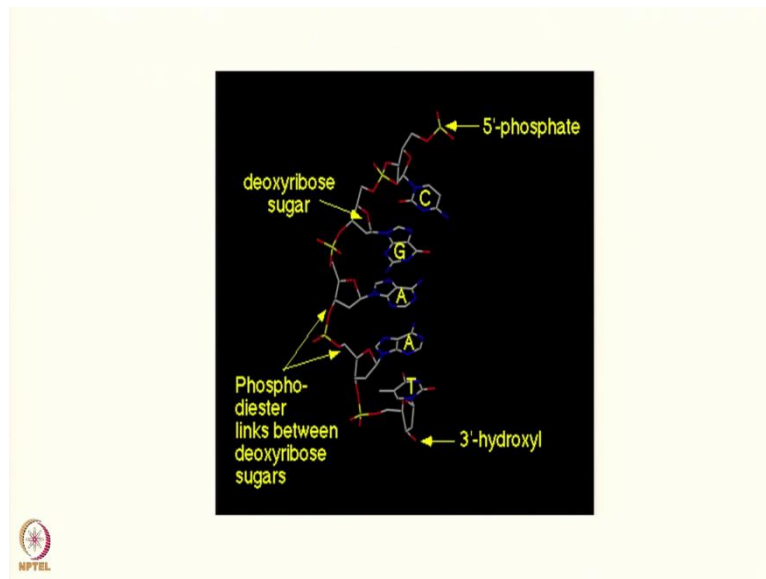
So, therefore now this is a kind of a summary of what I explained for so, long and that is the primary structure of the DNA or the RNA. So, this sequence goes like this. So, as I indicated earlier. So, you have a base here and this is the ribose ring or the deoxyribose ring whatever in the case of only the carbon skeletons are shown the protons only are shown for the CH<sub>2</sub> groups here.

But there are protons here protons here and protons there those ones are not shown. So, you have the phosphate group here. So, therefore this is called the phosphodiester linkage we have a phosphodiester linkage in this is called the phosphodiester linkage because there are 2 ester groups are phosphodiester. So, every nuclear dinucleotide has a particular phosphodiester linkage here.

So, you have the CH<sub>2</sub> this is the 5 prime end and this is the one prime here one prime attached to you see as I told you earlier at this amino at this nitrogen these bases are attached similarly the pyrimidine at this nitrogen the base the sugar rings are attached and likewise. So, this is at the that this nitrogen the sugar is attached and so on so forth. So, in the purines the sugar is attached at to the 5 membered ring and in the pyrimidines it is attached to the 6th member ring at this position.

And to repeat you have this one prime group here this is the 2' in the deoxyribose you have 2 hydrogens here and there is one hydrogen at the 3' and the 4' there is one hydrogen here and this repeats itself in the all along the sequence. So, in a linear sequence this is how the structure is made.

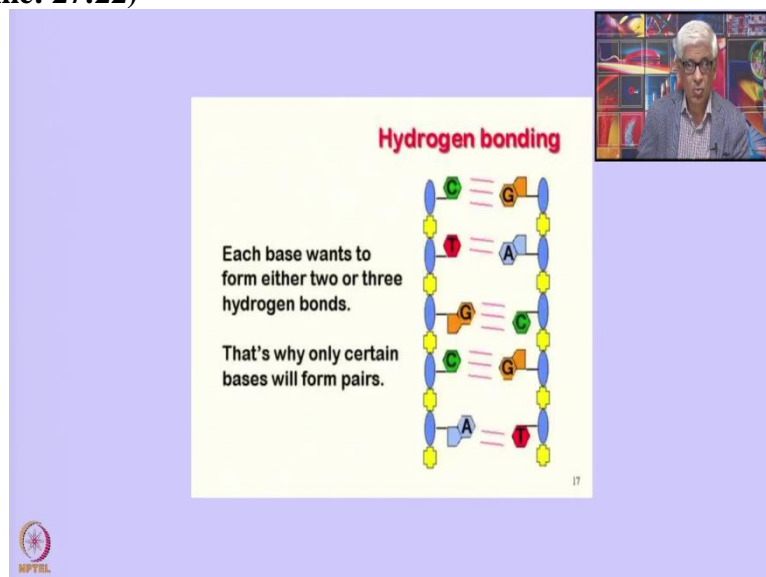
**(Refer Slide Time: 26:39)**



Now this is shown in a colour coded manner here. So, where the DNA sequence is formed you have the 5 prime phosphate group here and then there is the base attached to the sugar ring. So, this is called as the backbone. So, the phosphodiester linkage this is called as the backbone of the DNA you have the deoxyribose ring here ribose ring here. So, all the ribose ring in the phosphodiester linkages they constitute what is called as the DNA backbone.

And the bases are indicated here C G A A T etc. So, this find the sequence of the DNA and this is called as the DNA backbone.

(Refer Slide Time: 27:22)



Now what these people said Watson Crick and Franklin and all of these they proposed this model that the DNA is not a single strand but it is a double strand. And that is how the hereditary information comes hereditary property comes because of this. That is that if you have a DNA and these this molecule exists as a double helix what happens in the double helix there is one strand runs like this from 5' to 3' end and the other strand runs in the opposite direction 5' to 3' end.

And these 2 nucleotide units the 2 strands are held together by so called hydrogen bonding between the bases and these are called as the base pairs. So, you have the CG pair here the TA pair here GC pair here CGP pair here we also observed that C can pair with G and T with A

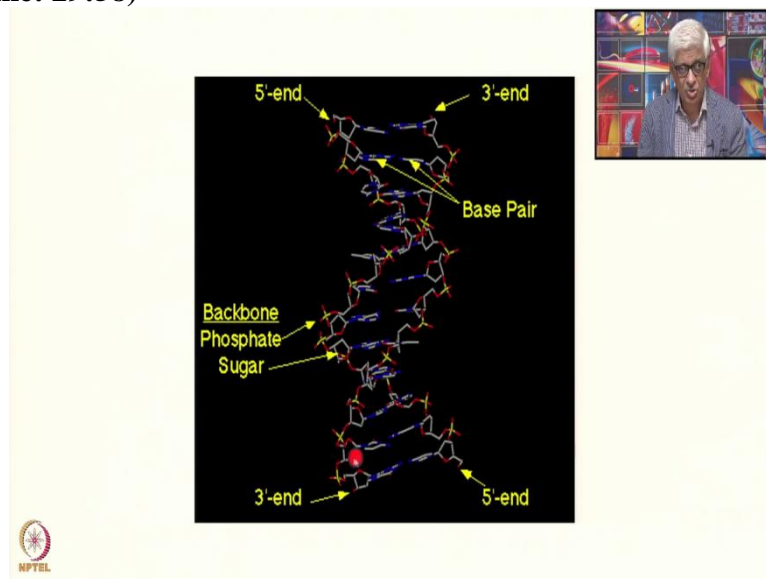
and no else. So, therefore there this is the model which they put up and of course this was based on previous informations inside available to them that the number of C's is equal to the number of G's the number of T's is equal to the number of A's in the DNA.

So, on that basis actually they had put up this model then you have the C and the G are hydrogen are paired together through hydrogen bonds and there are 3 hydrogen bonds between C and G and there are 2 hydrogen bonds between T and A and so on. So, this is how the model goes that the only certain bases will form base. So, now you see if this DNA breaks apart suppose you they have 2 strands separate out.

And in the daughter spring offspring if the new DNA is synthesized on the basis of one DNA strand naturally the other strand which will come will have the complementary sequence. So, if the sequence goes C T G C A the complementary strand will have to have this sequence here that T G C A G it has to have that therefore the information is transferred from one molecule to the other molecule.

If this is the parent molecule the during the process of replication the DNA strands separate out and each one of them actually become the starting point for the new DNA molecules to be synthesized when the new molecules are synthesized to maintain this complementarity of the base sequences and that is how the offsprings get the same kind of DNA information and that is called as the heritage.

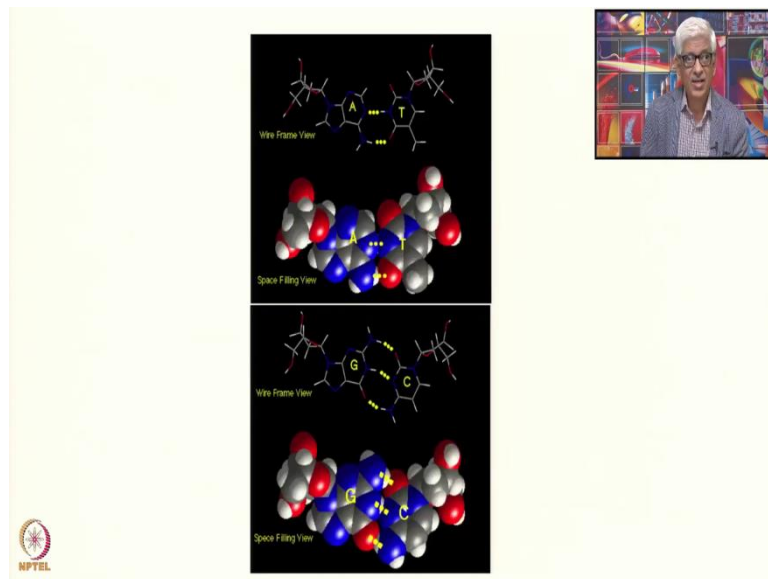
**(Refer Slide Time: 29:58)**



Now this is explicitly shown in a little bit more clear manner the 2 doubles the 2 strands are running in this manner you see you have the 5 prime end of one chain which is going like this one DNA chain is running like this. And the second DNA chain runs like this they are intertwined they are bound together in a kind of an intertwined fashion and then you have this base pairs that are happening between the 2 strands.

You have the phosphate backbone and you have the base pairs which are there between 2 strands which are running in opposite directions. So, this is how the hereditary information comes in the in the DNA molecule.

**(Refer Slide Time: 30:36)**



And to show explicitly what sort of hydrogen bonds are there and this slide shows you the particular way the hydrogen bonds are formed. A and T there are 2 hydrogen bonds formed here this is the thymine iminoprotons and goes here and the oxygen here goes to the NH 2 proton of this of the purine here. The purine has an NH 2 here these are the amino protons and this is the keto group here this is oxygen of the thymine which pairs with this.

And here is the amino group of iminoamino group of thymine pairs with the nitrogen of a and therefore this is how the hydrogen bonds are formed in this. Similarly in this case you have the amino of cytosine going to the oxygen of the guanine and the oxygen of the cytosine goes to the amino of the guanine and then there is an NH and hydrogen bond G amino applies to the nitrogen of the because you see the colour code here.

So, what is the colour code the blue is indicating the nitrogen and the red is indicating the oxygen therefore these actually clearly tell you what sort of hydrogen bonds are formed between the AT and the GC pairs. So, this has been the basic discovery which actually led to the Nobel Prize to Watson Crick and Wilson. So, we will stop here.