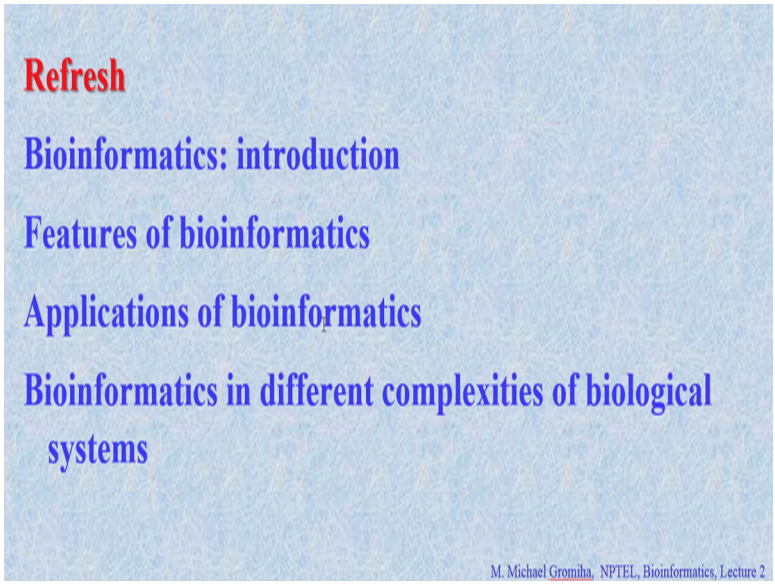
Bioinformatics
Prof. M. Michael Gromiha
Department of Biotechnology
Indian Institute of Technology, Madras

Lecture - 2a
DNA Sequence analysis

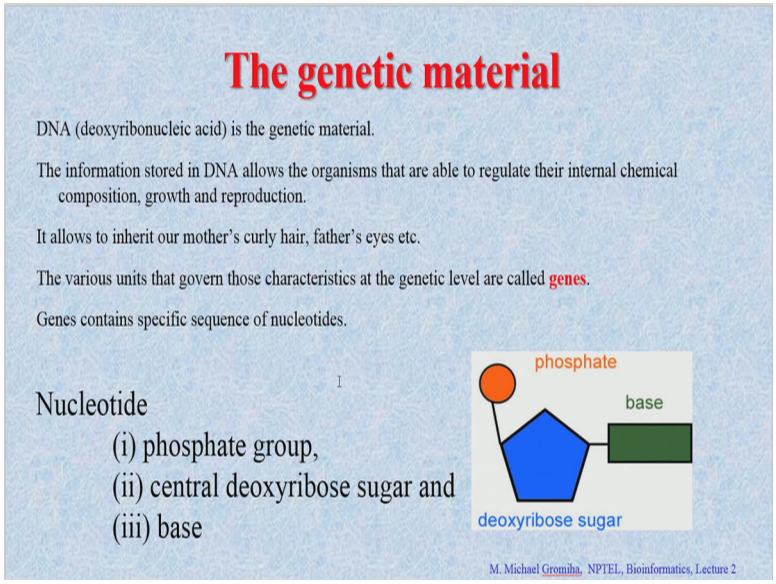
(Refer Slide Time: 00:20)



In the first lecture just for refreshing we discussed about the basics of bioinformatics with few examples, and the different features of bioinformatics; for example development of databases, algorithms and hypotheses, structure based (Refer Time: 00:26) design and next in the sequencing.

And we discussed about the applications of bioinformatics, on different complexities of biological systems.

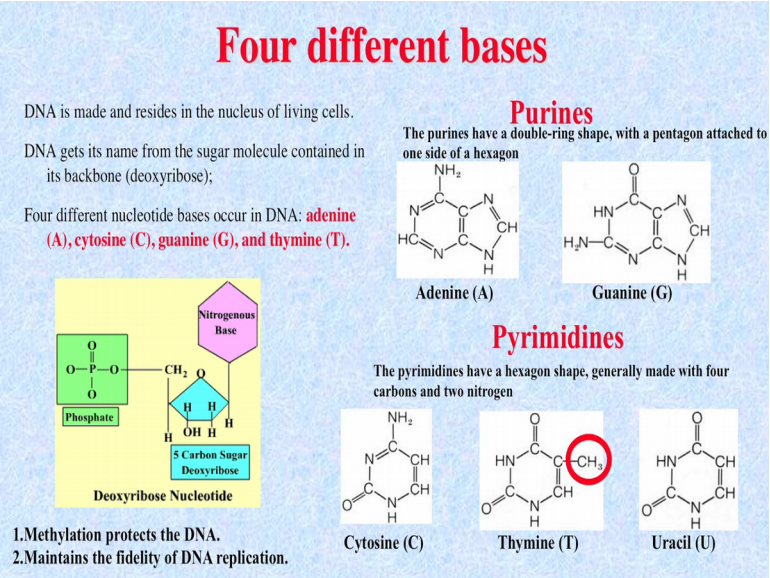
(Refer Slide Time: 00:42)



If you talk about the DNA it is genetic material, on the information stored at DNA allow the organisms that are able to regulate their internal chemical composition growth as well as reproduction. For example, it allows mothers curly hair and fathers eyes and so on that we know that they getting the genetic information. So, various units that governs this characteristics and genetic level are called as genes.

So, there are several genes and correctly form a genome sequence, it contains various genes. So, if you look at the genes they contain specific sequence of nucleotides for example, ATCG right adenine guanine cytosine and we (Refer Time: 01:27) nucleotide. So, this is an example, it contains a phosphate group here and here this is the central deoxyribose sugar and here is a base here there are 4 different base for the DNA and the 4 different base for RNA.

(Refer Slide Time: 01:45)



Three are the common and one is different thymine replaced to uracil. So, if you look into this DNA as well as the RNA. So, the name of DNA that sense for deoxyribose nucleic acid, come from the sugar molecule contained in the backbone. So, if you see this is the base and this is the phosphate and here this is sugar. So, here if you see this is starting form 1 here, 2 3 4 and 5. So, in the two prime here it is h this is deoxy.

Hence, the name deoxyribose nucleic acid this DNA for a case of RNA, so this is OH because this is oxyribose; so is the (Refer Time: 02:24) nucleic acid. So, there are 4 different bases in the case of DNA, we have adenine cytosine guanine and thymine and if you look into this 4 bases, they are classified in two groups one is called purines and they are double ring shape, here is one ring this is another ring shape with a pentagon attached to one side of a hexagon. So, which is the hexagon and this is another pentagon attached to this hexagon.

So, in pyrimidines they have hexagon shape here right. So, made generally with the 4 carbons and 2 nitrogen, if you see 1 2 3 4 carbons and 2 nitrogens. So, cytosine is both the DNA and RNA and the difference between the thymine and uracil; thymine you can see in the DNA and the uracil you can see in RNA. So, here this is the one CH3 group that is the made difference between the DNA and RNA. So, difference comes on two difference aspects right what are two different aspects DNA and RNA.

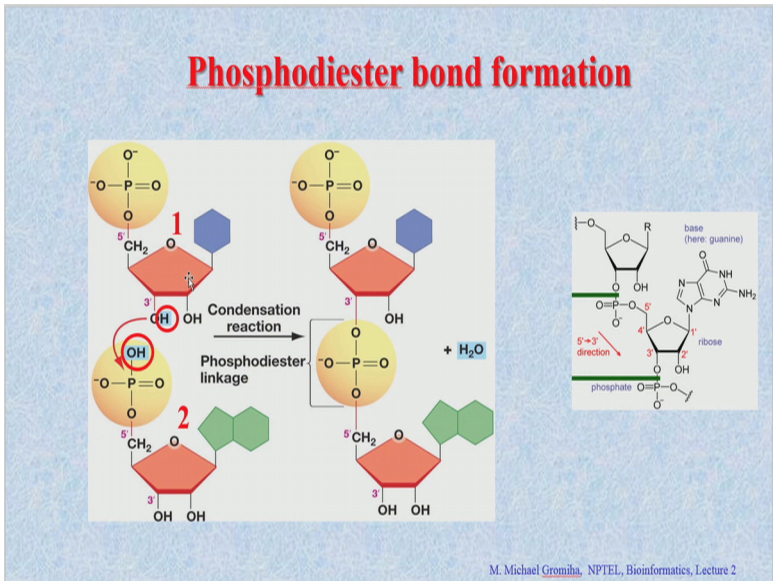
Student: Deoxyribose and (Refer Time: 03:18).

Right what is at the sugar level?

Student: (Refer Time: 03:21).

And the second is another base level. So, you can see the thymine and for the DNA and the uracil in the case of the RNA. So, how they form? So, there are the 4 different bases attached with a ribose sugar and the phosphate, and how they form a chain of this DNA sequence.

(Refer Slide Time: 03:35)

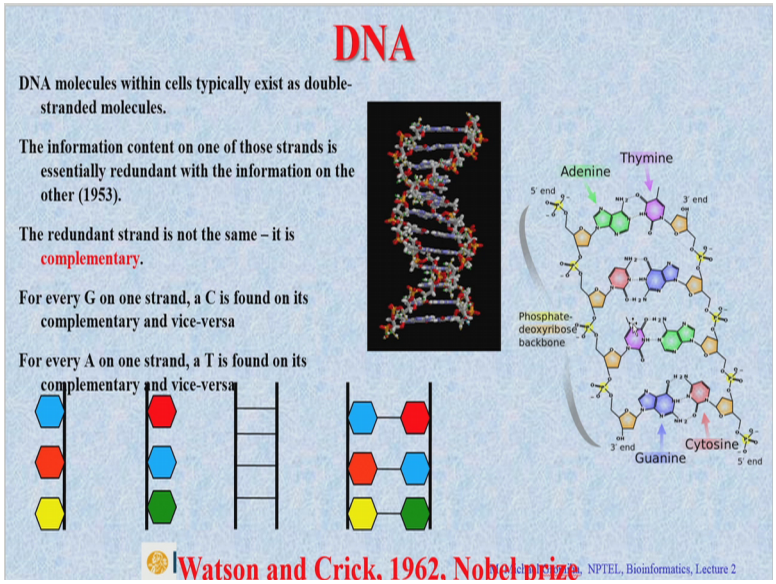


So, here these for example take molecule 1 this is a molecule 2; so in this 3 prime. So, here is OH right. So, here is the phosphate attached with OH.

So, this H and this OH for the condensation reaction; so eliminate water molecule, and then you can see there is a linkage phosphodiester linkage because phosphodiester two esters right one with this one and this one. So, we can form the diester linkage; so again this molecule. So, if you this is the first one right and here this is the second one, here this is the linkage with the condensation of this water molecule right.

So, when you have this phosphodiester linkage here I show the continuous ones. So, the one nucleator contains form here this line to this line. So, you can see this is the phosphate attached here, and this is the ribose, and here this is the base right and here to this one you can see the one nucleotide.

(Refer Slide Time: 04:36)



So, now if you look into this DNA, DNA molecules within the cells typically of two double strand. So, you can this is the one strand and this is another strand right in the information content in the one strand here right essentially redundant on the information on the other.

If you see this is redundancy is not the same, because the direction is from see here the direction from 5 prime 3 prime right. So, here this is one is 5 prime 3 prime and the other one is and the 3 prime 5 prime, and then you can see the complementary a of this basis. So, every G on one strand for example, right and C is found on the complementary strand and the vice versa.

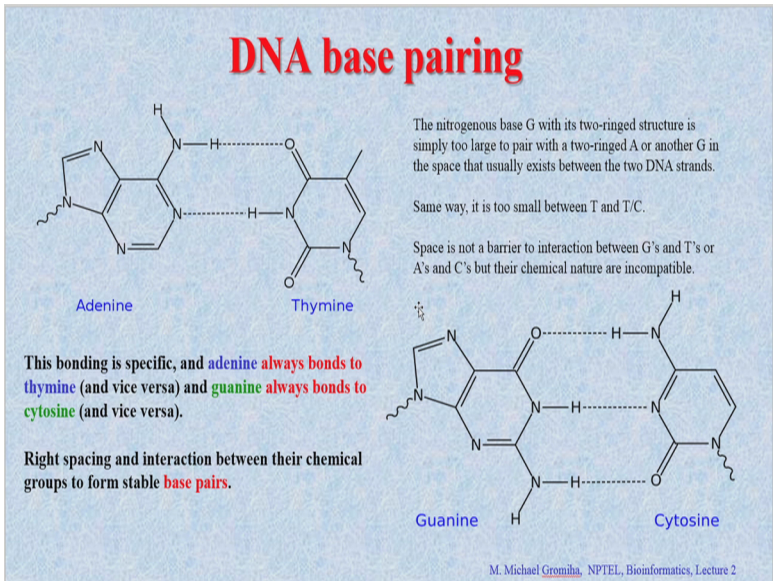
So, here I show the example right. So, here this is the green adenine and this red one is cytosine, and the blue is a guanine and the emergent the iso thymine right. So, adenine spread with the thymine right here this is cytosine is spread with the guanine, and here this thymine is spread with adenine and the guanine is spread with cytosine. So, when they form make this model of this DNA, first they try to put that to strands here this is the backbone net past by backbone here.

And then first will try to put the bases on the two other sides, then they could not get the compatible structure. So, then they think about the ladder type structures. When they (Refer Time: 05:57) they play lagos right they like to link this lagos each other right. So, likewise take they ladle (Refer Time: 06:04) structure and they put the two backbones

here, and they made the side chains the bases in between the ladders, then they found it exactly matched. They could see the matching (Refer Time: 06:17) bases right with respect to the chemical groups as well as the space right.

And then they found the hydrogen wanting pattern also they could exactly fit with the favorable energy. Now, for this one is the Watson crick the proposed in model for the (Refer Time: 06:33) DNA and they got the Nobel Prize in 1962 for this (Refer Time: 06:37) structures.

(Refer Slide Time: 06:38)



So, now how this base pairing happens; so if you have adenine and this thymine right it forms two hydrogen bonds rights it is pairing, and the guanine and cytosine they are paired with 3 hydrogen bonds. Why this pairing is specific why not with the other basis, why a start pairing with a G and the T is pairing with the C.

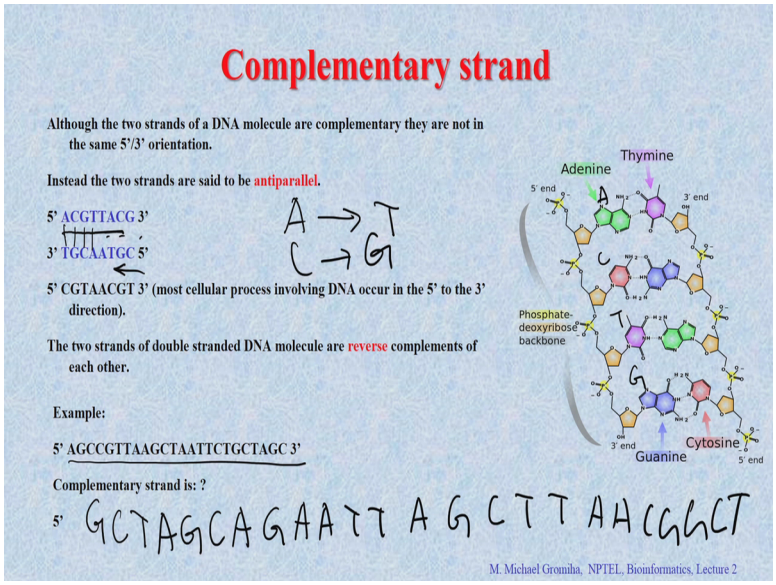
Student: Because of (Refer Time: 07:00) some because one is purine and other is pyrimidines. So, that will be uniform in space.

So, because if you yeah you are right. So, if one is purine and what is the pyrimidine, if you have try to have this type of 3 two pyra pyrimidine parents are two pyrimidines, either they have excess space which also has loss of energy and also they are very crowded. So, they did historic interaction they are not able to pair with each other. So, two different ways one is the static hindrance and the second one is a chemical group.

So, because of the two reasons adenine is always pairing thymine and the guanine is always pairing with cytosine.

So, if you know one strand, then we can know the complementary strand because A always pair with the T and the G always pair with C.

(Refer Slide Time: 07:51)



So, if you see this one as I discussed now, the two strands of DNA molecule are complementary right, but they are not in the same 5 prime 3 prime direction. One is in the 5 prime 3 prime direction, one is in the 3 prime 5 prime direction right, because most of cellular process they involve in the DNA occur in the 5 prime 3 prime, this is why the right in the sequences in the 5 prime 3 prime directions right ok.

So, now if you have one sequence for example, ACGTTAVG, we say sequence and what is the complementary sequences? If this is a 5 prime 3 prime sequence right when you go to the just complementary, then you can get the pairing you can write this complementary sequence, but that will be in the 3 prime 5 prime direction right. So, for the A what is the complementary for A?

Student: T.

T or the C.

Student: G.

It is g right. So, we write like this A to T, C to G, and G to C, T to A and so on. This is the 5 prime 3 prime direction and here this is 3 prime 5 prime direction. So, we need to the complementary T. So, we had to reverse direction. So, if you write the 5 prime 3 prime, then you reverse this direction right. So, then start from CGTAACGT right this is how to write the complementary sequence. Here a few example which is A, this is C and here this is T and here this is G right.

So, this is here if you write 5 prime 3 prime, this is 3 prime 5 prime this is why just you can see the complementary here. So, and I give this example with this a sequence, what is a complementary strand from 5 prime 3 prime direction.

Student: G.

G.

Student: C.

C.

Student: T.

T.

Student: A.

A.

Student: G.

G.

Student: C.

C.

Student: A.

A.

Student: G.

G.

Student: A a.

A a.

Student: T t.

T t.

Student: A.

A.

Student: G.

G.

Student: C.

C.

Student: T t.

T t.

Student: A a.

A a

Student: C.

C g.

Student: G.

G.

Student: C.

C t right; so we have to get the pair and this reverse the direction. So, write this is the complementary strand. Now in the bioinformatics you can write algorithms we have

were any given sequence you can write the algorithm to get the complementary strand right how to write the algorithm? First what we have to do.

Student: We have to take this sequence (Refer Time: 10:04).

Take this sequence and put.

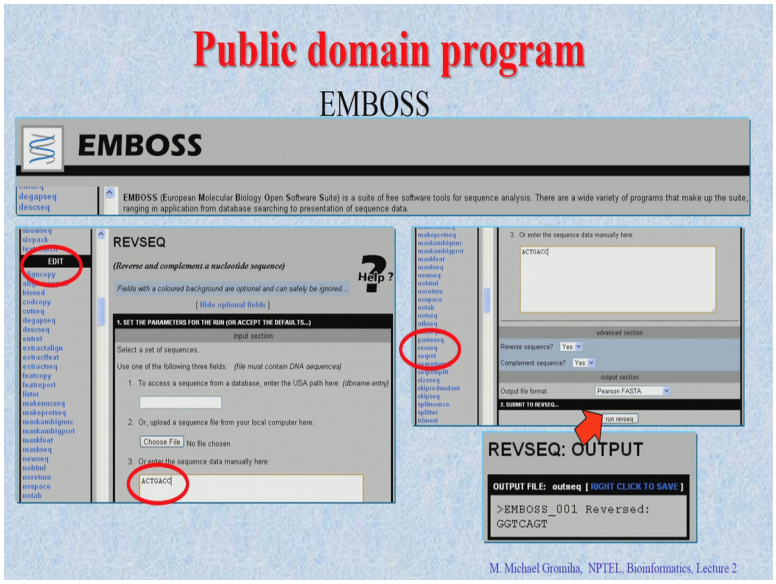
Student: Reverse it.

Reverse it and the complementary.

Student: (Refer Time: 10:09).

Or you can get the complementary and reverse it, right.

(Refer Slide Time: 10:11)



So, you can do that likewise there are various programs available in the literature right because they collected various programs to analyze the DNA sequences put it together as kind of a package.

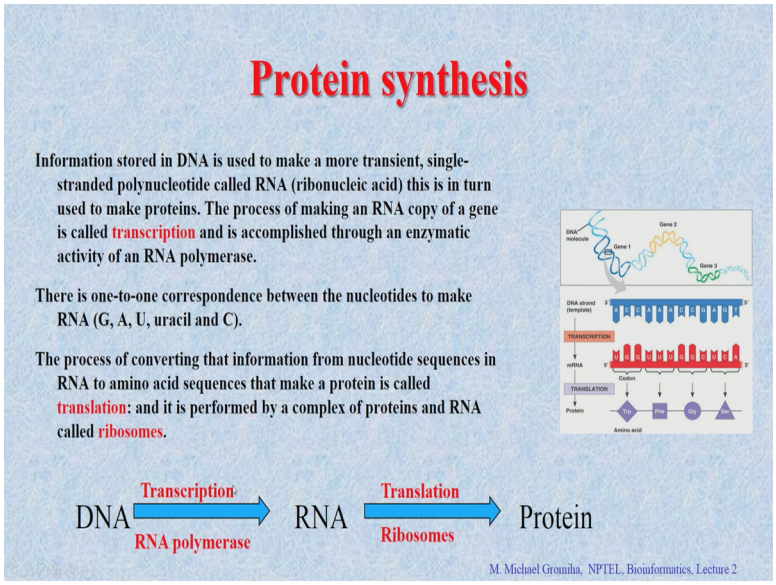
So, emboss is such a package emboss stands for European molecular biology open software suite right it is a compilation of several programs you can see the several programs at the end at the side here write the menu. So, for a sequence analysis; say behave DNA sequence we can use this sequence and do way carry out various analysis right using this specific a software called emboss, right.

How to get this complementary strand from emboss software? So, we go to the emboss website. So, then go the edit. And finally, you can see the model revseq, that is reverse the sequence reverse and complementary a nucleotide sequence. So, either you can give sequence from database or here you can choose a file, if you have your file in your computer or you can give a sequence manual type is sequence.

Here I give a sequence ACTGACC right. Now, you if you click on the run reverse sequence, then you will get the sequence GGTCAGT is it correct right (Refer Time: 11:24) in reverse complementary of this sequence right this is fine. Now you get the complementary strand, now next step is when you have the DNA sequence you can translate this into proteins right you want to two steps involved in protein synthesis.

Student: Transcription.

(Refer Slide Time: 11:43)



One is a transcription, another is the translation right. In the transcription what happens in transcription DNA is converted to.

Student: Messenger.

Messenger transcription by messenger RNA right RNA and the translation?

Student: (Refer Time: 11:52).

Messenger RNA to the proteins right; so two steps, first one we have to have the DNA and it change to RNA right in this case DNA as a 4 different bases ATCG in the case of RNA a.

Student: U.

UUG right; so this way if you see here this is the DNA sequence right contents ATCG right, but is a RNA sequence this case no T, but instead of T have you right. The a complementary A is T instead of T you put U here C is G and this is G and this A is U and A is U and so on. Now if you have the RNA sequence right now the then this RNA is translated into proteins by a ribosomes.

But in a transcription mainly the they RNA polymerize does this transcription right now the ribosomes are responsible to convert this translate the m mrna to proteins right. So, there they are the different nucleotides right they use the codons, there 3 nucleotide together form one codon, so each codon they re code for a specific amino acid. So, there are 4 different nucleotides, but how many amino acid residues 20. So, 120 correspond is not possible. If you take one to one, 4 nucleotides mean code for only the 4 nucli 4 amino acids, but you have 20. If it is combination of two how many combinations?

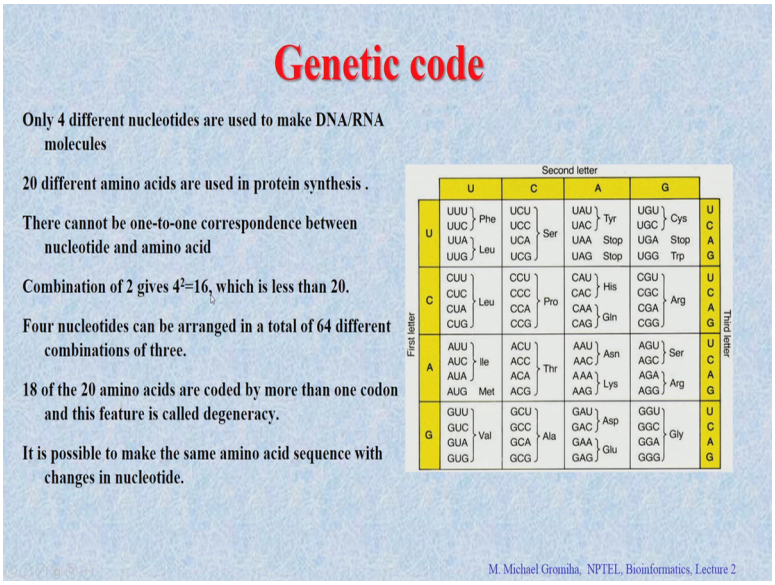
Student: (Refer Time: 13:18).

4 into 4 equal to 16 so that is also not possible because we have 20; so these have the 3 d combinations, 3 means totally how many possibilities?

Student: 64.

64 possibilities right to 3 to 3 to this.

(Refer Slide Time: 13:32)



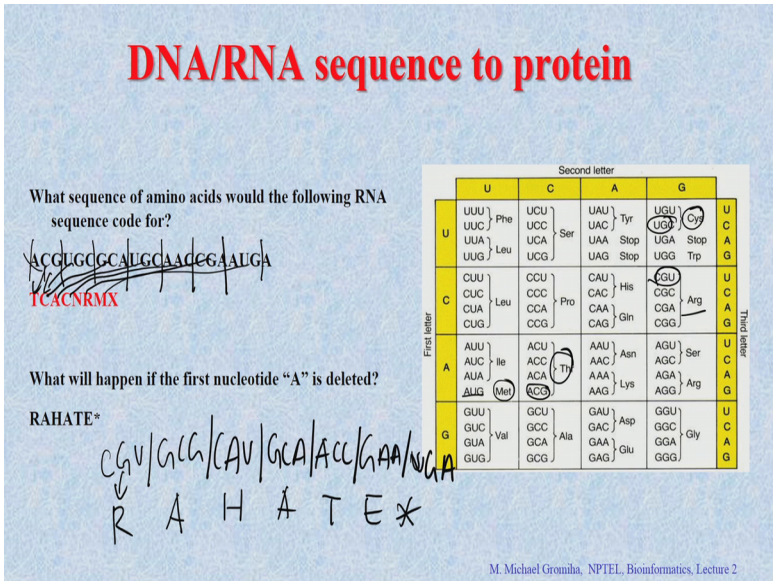
So, if you give the 4 into 4 into 4 64 combinations, but a 20 different amino acids. So, there are several codons which code for same amino acid that is called the degeneracy; not one triple single codon to one right same different codons, they code for the same amino acids. We see the phenylalanine right. So, there are two codons UUU and UUC they code for phenylalanine. If a leucine you have six here UUA, UUG and CUU right CUC, CUA and CUG all this code for leucine.

So, as a one stop codon what is a stop codon you can see here this is a stop codon.

Student: U (Refer Time: 14:11).

Right UGA is a stop codon. Also tryptophan is code with the only one codon right UGG. So, this is also one of the reasons why tryptophan is occurring a very less in the case of the a protein sequences. So, 18 of the 20 amino acids are coded with the more than one codon right and this is called the degeneracy system.

(Refer Slide Time: 14:33)



So, if we have a DNA sequence, can we get the protein sequence if you have DNA sequence or RNA sequence? We had I gave RNA sequence, can we get the protein sequence from this RNA sequence yes what to do? Yeah first take 3 right take this 3 right this 3 3 3 3 steps right fine.

So ACG, what is ACG? A here C here and G here right ACG, so AGC is for the 3 (Refer Time: 14:59) right. So, the ACG code for t, then UGC that is cysteine right UGC right UGC is here this is cysteine. So, it can put cysteine right then GCA is for the (Refer Time: 15:20). And this again UGC bar cysteine and a a c for the as per (Refer Time: 15:24) c g a is for the (Refer Time: 15:25) and AUG right AUG is for the methionine right. So, now, that is (Refer Time: 15:31). So, they could the x.

So, in this sequence what will happen, if you delete this nucleotide a if this not there, you will get the similar sequence are difference sequence.

Student: (Refer Time: 15:43).

Difference sequence right, because now the codons are.

Student: CGU.

CGU.

Student: GCG.

GCG.

Student: CAU.

CAU.

Student: GCA.

GCA.

Student: ACC, ACC.

ACC.

Student: GAA.

GAA.

Student: UGA.

UGA right; so now CGU code for CGU.

Student: (Refer Time: 16:20).

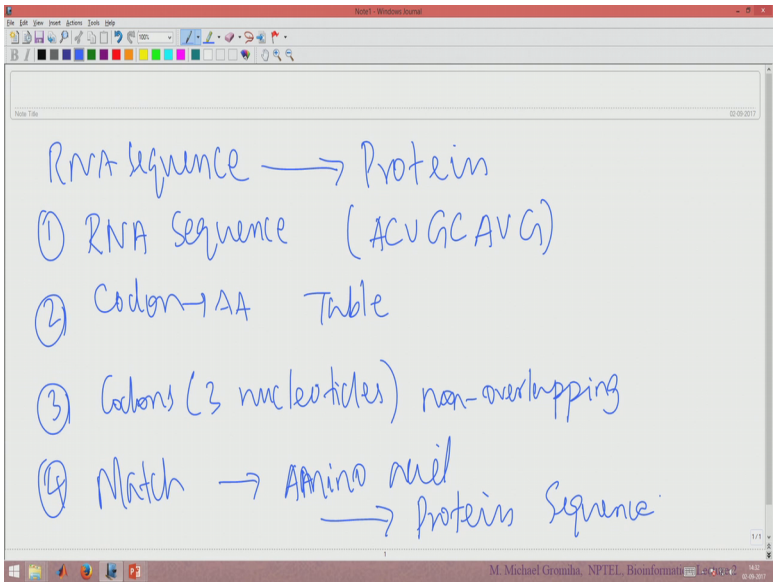
Arginine right; this arginine.

Student: (Refer Time: 16:21).

This is (Refer Time: 16:22) this is (Refer Time: 16:23) GCa (Refer Time: 16:24) TGA A G A A E and UGA UGA sub codon, they will put a sub codon here. So, if you remove one single nucleotide. So, if there is this totally (Refer Time: 16:43) right. So, if you this way if you have the DNA sequence, I will look for the protein sequence we should know exactly where we start the codon in otherwise you will get the different ah protein sequence.

So, there are various resources, you can also use programming you can write you one code write you get translate the RNA sequence and protein sequence how to write a program right? So, first order you what the information required inputs you can need this input sequence right you need the RNA sequence here.

(Refer Slide Time: 17:13)



So, you need to get the protein right. So, how to write the program to get the protein sequence in RNA sequence what are the input necessary?

Student: (Refer Time: 17:30).

First you RNA sequence right you only first we need the RNA sequence for example.

Student: ACU.

ACU.

Student: GC.

GC.

Student: AU.

AU.

Student: G.

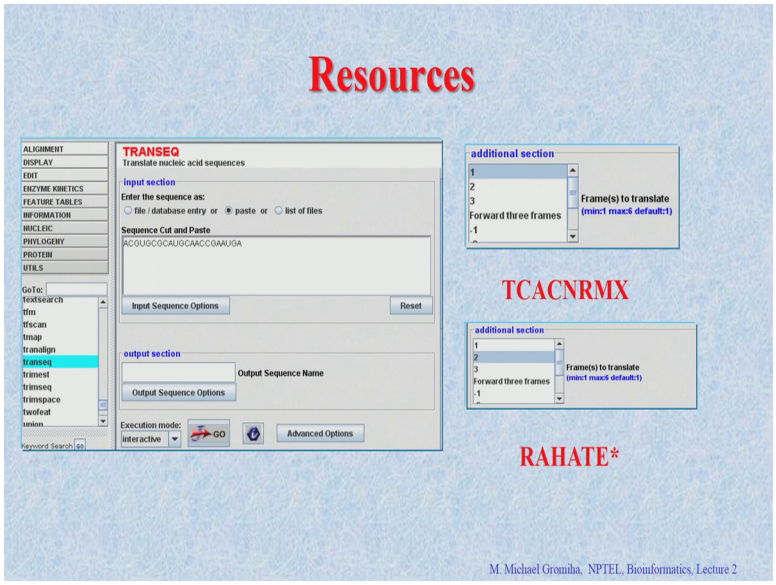
G. So, you have the RNA sequence (Refer Time: 17:49) other information require the code right. So, you get the mapping right you have the table right the codon to amino acid table right. So, then did will get the information regarding each codon codes for

which amino acid then what you have to do? Is not first change the RNA into (Refer Time: 18:11) right you have to codons right; so 3 nucleotides.

Say cut into 3 pieces right not the overlapping right (Refer Time: 18:24) to stop when I start from the next one. So, non overlapping right there is very important. Then you have to match right with the codons with the table right then you have to map match the sequence. And then finally you obtain the amino acid.

And finally, if you do it till the end till the end of this sequence then finally, we get the protein sequence. Say it get a RNA sequence read the RNA sequence and get the table right codon (Refer Time: 19:05) table and you have to cut the RNA a sequence into the pieces of 3 a nucleotides, that for each codon right (Refer Time: 19:12) non overlapping ones right then you map with the codon see the amino acid in the table then you get the amino acid right then finally, do till the end and look at the protein sequence.

(Refer Slide Time: 19:30)

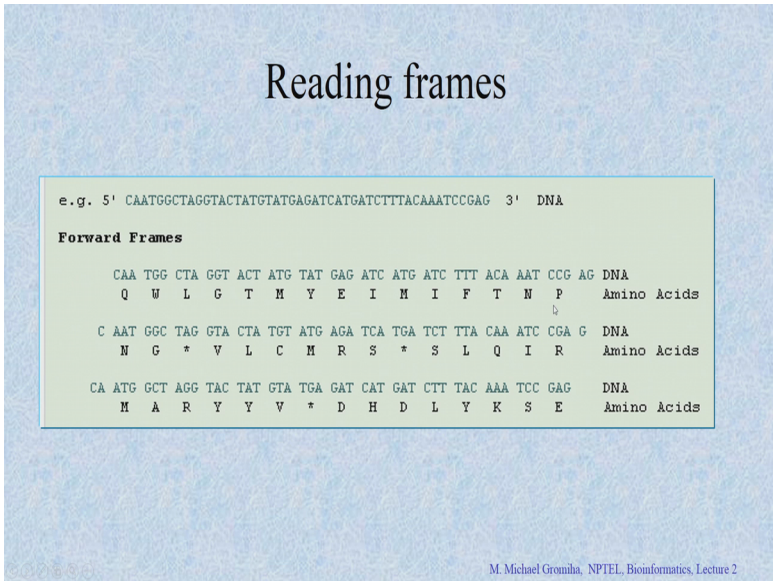


So, nevertheless in the several resources are available in the literature. So, again you can use the emboss software to translate the RNA sequence to protein sequence right. So, here you can use the transeq, this is the module available in the emboss software right. So, here you put the sequence ACG UGC GCA UGC ACC GAU GA right. So, now, if you go with one, it will ask for the different frames right. So, there are 6 different frames right how many frames.

Student: 6.

Six frames right. So, what are 3 6 frames? 3 forward frames and 3 I will just explain the next slide right. So, you get the moment you take this sequence and if you ask to convert the protein sequence right you can see this into this same sequence. I showed the example here the same sequence I give right ACG.

(Refer Slide Time: 20:24)



So, I get the same sequence TCA CNR MX. So, if you cut this one right the first one then you get the same sequence. So, for each sequence we have six reading frames then how do you get the 6 reading frames?

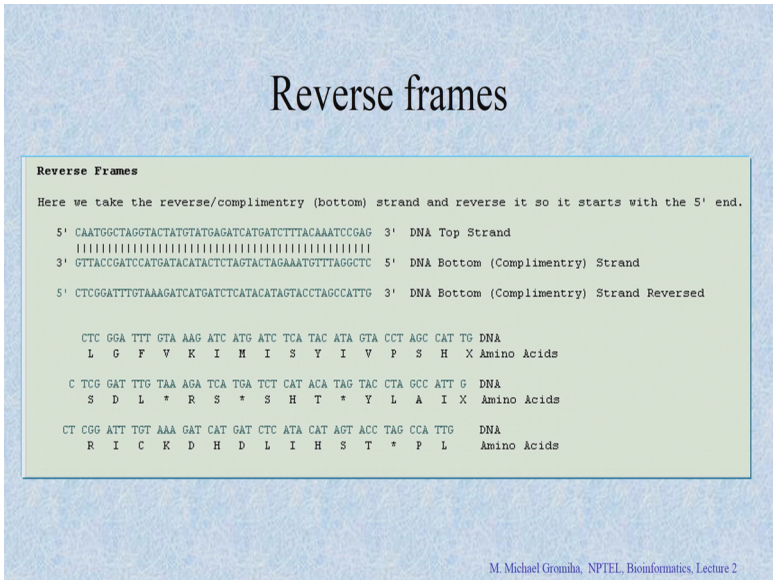
Three in the forward direction, 3 in the reverse direction how do you get the 3 in the forward direction? For example, this is the 5 prime 3 prime this a DNA sequence right. So, first you continue (Refer Time: 20:39) 3 nucleotides. So, get you call it CAA TGG CTA and so on at the DNA. Now, you go with this when I sits the CAA codes for Q, and TGG code for W and this code for l. So, it get the amino acid sequence right this is one and the second one it take this one out c then you make in the another 3. So, AAT GGC TAG and so on. Now, I check this table for the amino acid. So, here you can see the amino acid.

And the third one if you take this to C and A right, and then it start with the C A and ATG GCT AGG TAC and so on. So, now, looked into the table each corner right you will get the |M A R T and so on what will happen if you more again?

Student: Again the same (Refer Time: 21:26).

The same will repeat right say if it (Refer Time: 21:28) TGG. So, it will start from here right you do not have this amino acid, but you start from same (Refer Time: 21:34) same sequence. So, in this case you will get the three different (Refer Time: 21:39) likewise in the reverse frame. So, how do I get the reverse frame?

(Refer Slide Time: 21:42)

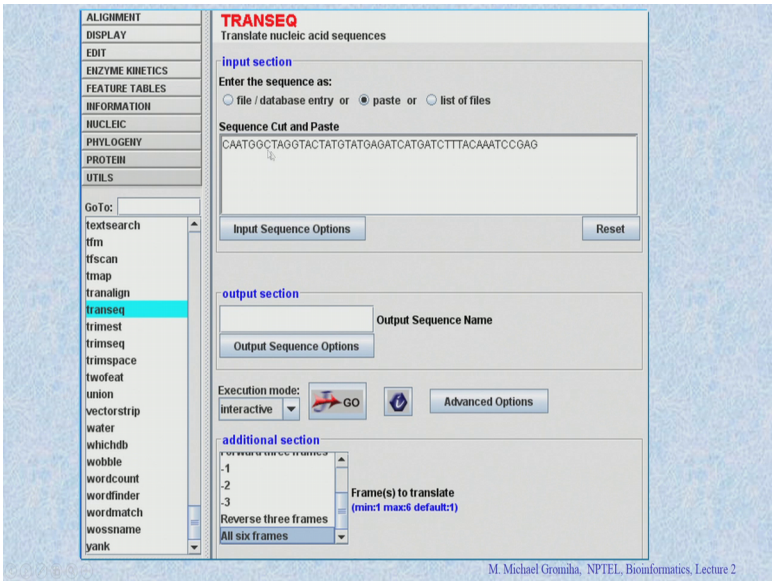


Student: (Refer Time: 21:46).

(Refer Time: 21:46) complement right. So, let me have this 5 prime 3 prime now if you take the complementary right then if you take the complementary then it will get the 3 prime 5 prime direction right and you reverse it right then you get the same strand they give from the 5 prime 3 prime. Then from this 5 3 three prime sequence, then you make the codons right here CTC GGA TTT and so on well this is the this will code for L, GGA code for G and TTT code for W F right. Now, you can get the amino acids.

Likewise just you take one out then you will get the second frame and the third frame. So, that total will have sixth difference 3 forward frames and 3 reverse frames. So, you can use this transeq right to get the sixth difference, here if you give the input sequence.

(Refer Slide Time: 22:29)



And then here you have options whether this is the forward reading frames or the reverse reading frames right are everything. Say if you go with the all six frames right you go here a click on all six frames.

(Refer Slide Time: 22:45)

So, you will get the sequence right. So, this is the forward frame 1, this is 2, this is 3 and the reverse 1, reverse 2 and reverse 3.

So, then if you go into the same software emboss. So, if you see there are various tools available in emboss. So, everyone have the different values and they have gives different

algorithms to get the values. So, we can use any of these different module available in the emboss right to get the characteristic features of this (Refer Time: 23:15) DNA sequences or RNA sequences.