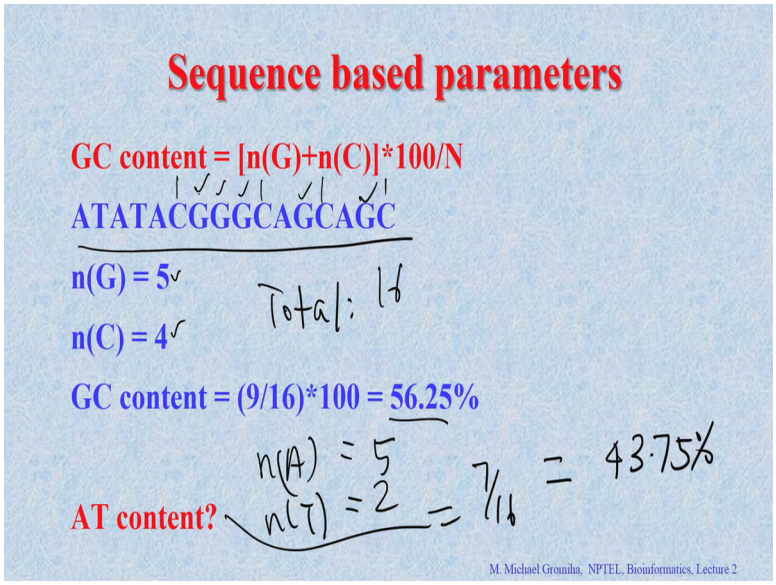
Bioinformatics
Prof. M. Michael Gromiha
Department of Biotechnology
Indian Institute of Technology, Madras

Lecture - 2b
Sequence based Parameters

(Refer Slide Time: 00:17)



I will explain few examples right I will show few examples, now if you have a sequence right for a example this a sequence ATTACGGGCAGGAGC, right so this example. So, what is the AT content or what is the GC content right because this is very important because somehow some (Refer Time: 00:34) some genomes, they are highly reach an AT content and some cases highly reach an GC content. And if you look into this genomes if there are highly AT contents. So, they are highly flexible and they are (Refer Time: 00:46) GC content they seems to be little bit rigid.

So, and also there are places where we have more AT sequences right this have sequences they have high possibility to interact with the a protein side because there are highly flexible. So, they can change the conformation and they can have high probability to interact with the proteins. So, this will give you an option, how where are the flexible sequences or the rigid sequences and the AT content or GC content will give you the information.

So, now this is the sequence. So, what is the G C content? So, for this case G C content means what is the percentage of G and C in the particular sequence for this first we get the number of G right. So, commonly G is here 1, 2, 3, 4, 5 5 Gs number of G equal to 5, then number of C how many number of C Is 1, 2, ,3 4 right number of C equal to 4 now what your total number.

6 7 8 9 10 11 12 13 14 15 16 right 16. So, G C content you can calculate out of 16 how many G and C this equal to 5 plus 4 9. So, 9 by 16 into 100 this equal to see 56.25 percent. How do you get the AT content, how many number of As?

Student: (Refer Time: 02:08).

Five number of Ts.

Student: (Refer Time: 02:13).

2. So, AT content equal to?

Student: 7 by 16.

7 by 16 this equal to.

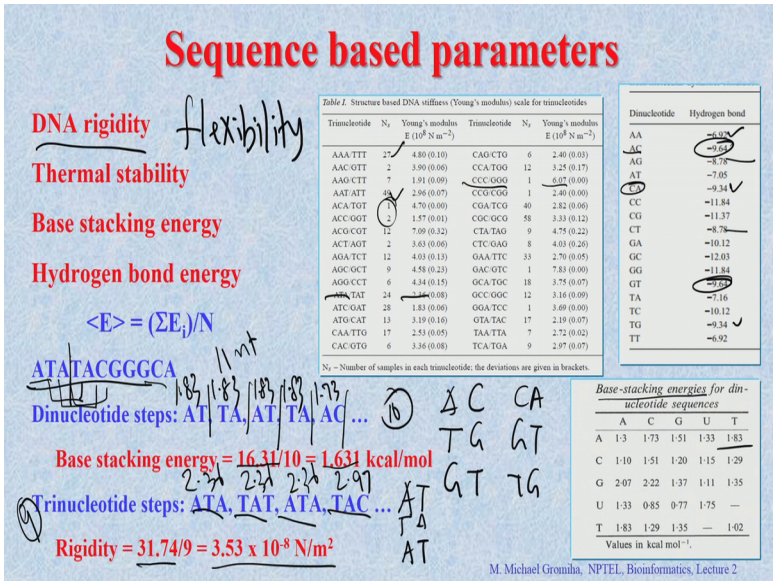
Student: 43.

43.

Student: 7 5 percent.

7 5 percent; so you can do that we are easily you can subtract content from this, will be get this right, ok.

(Refer Slide Time: 02:27)



Now, also there are we can derive various sequence based parameters. So, if you see the nucleotides. So, they have various features right we can calculate various features based on this sequence. So, I explain few aspects right. So, one is DNA rigidity right or DNA flexibility.

Right so totally how many bases?

Student: 4

If a DNA 4 right; so if you take the dinucleotides there will be.

Student: There will be 16.

16 dinucleotides, but if you consider a pairing then 16 you can divided you can reduce to 10 right 16 divided is to 10. For example, if you see a a a a is this one if it is C A, C A is similar to right if you A C right what is the complementarity?

Student: (Refer Time: 03:30) TG.

T G if you take the reverse.

Student: G T.

G t right; so if you see the A C and a G T, they both are same right likewise if you take another one for example, if you take the AC I would take the C A right. C A means this equal to G T. So, T G right this is 9.34 and T G is 9.34 right if you see the how many unique pairs 12. So, A G and C T right this will be same and the 7.05, this is a AT right again in the same you will get the same one. If you would the TA and AT right if you take AT, what is the complementarity?

Student: TA.

T A if you reverse.

Student: AT.

A T here are the same right. So, you will get the 10 wise you will get right. So, this is for the di if you go the tri totally how many possibilities?

Student: 6.

64 possibilities and due to this complementarity, you will you can reduce this 32. This way if this is the dinucleotides you can this pairs right, but here trinucleotides you have the out of 64 you have 32 cases right here. So, now, first we got the flexibility, how you can either use the dinucleotides or the trinucleotides what is the advantages of using dinucleotides and trinucleotides? Dinucleotides have less information because we have the less number of possibilities, trinucleotides have more information because d a a d a a is we can get more information right. So, the trinucleotides can provide better information than dinucleotides right. So, we provide trinucleotides.

Then what will happen if tetra nucleotides more information, but they what are the effect what is the problem issue the tetra nucleotide like amino acids right if you have to how many amino acids.

Student: 20.

20 if you go the pair 400 right 20 into 20 if you go the tri.

Student: 8000.

8000, but the address it is same if you want to derive the features we use the same data set for example, 1000 or 10,000 sequences. If you use the di you will single amino acid case you will get more number of data. If you go the di then you will your data distributed. If you were the tri then it is very (Refer Time: 06:00) right some case you will get zeros for many cases right.

Likewise the nucleic acids; if you take the di you get could number of data, but tri provides more information and go with the tetra you will get a better information, but the number of data are less. So, the values you obtain for each tetra nucleotide is statistically not significant, because here even trinucleotides if you see some cases you get very less number right say some case if you see 1 2 like this.

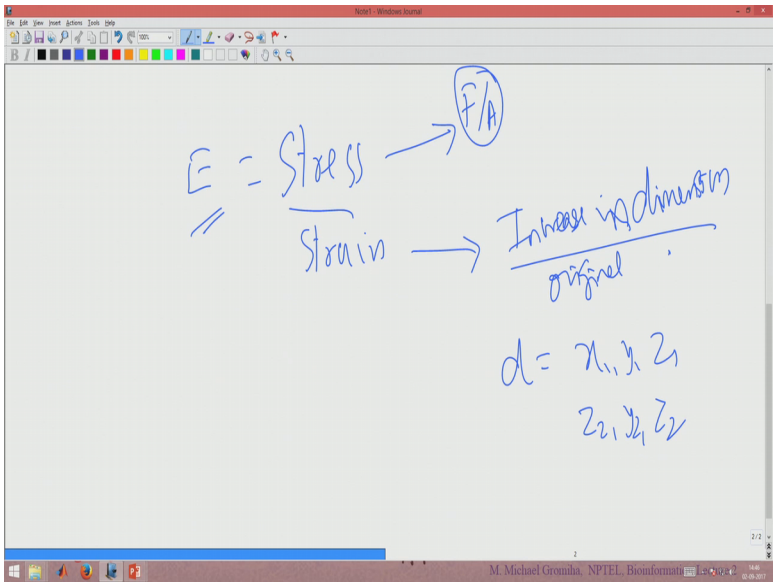
So, if we go the tetra then they data will be less right how to get this one? For the case of DNA flexibility there are various ways to get the flexibility, you can consider the DNA as a rod right. So, put the one base second base and third base right then you can bend you can apply a force, and you can bend it right, then you can see the same force you apply for all the trinucleotides and see how far they can bent based on that you can compute the flexibility or the rigidity of each trinucleotide. If you have the 3 d structures where shall we get the 3 D structures of the DNA?

Student: NDB or PDB.

NDB or the PDB right nucleic acid database or the protein database bank right. So, if you take all the trinucleotide steps right. Take the whole DNA structures are cut into pieces right do you want you have the overlapping pieces right all the trinucleotides you take right and then if you if it is completely a straight right, you can put one stack on another. So, you know the width of the diameter of this DNA you know, the thickness you know. So, you can see this is some point to (Refer Time: 07:40).

But if you see this structures that it is not straight, it is bent. So, in this case based on this bending angle you can this related with the flexibility right because.

(Refer Slide Time: 07:55)



If you know Young's modulus what is Young's modulus? Stress by strain right equal to stress by strain right what is stress? Force per unit area because this we know this is circular rod cylinder right you know the radius.

So, in this case we know we can calculate the area. Force also we know how much force we give there is 480 pico Newton's right if you make it constant because we use the constant, in this case does not matter right. So, it is (Refer Time: 08:23) 480 piconewtons right. Now, we can calculate this, what is strain? Increase in dimension increase in dimension divided by original dimension right. So, if it is without any change then this is straight. Now, in the 3 D structure you can calculate the distance from the first one and the forth one.

So, you can calculate the distance right if it is the coordinates x 1, y 1, z 1 and x 2, y 2, z 2. So, then calculate distance now from this distance and the original distance you can calculate difference. So, form this difference you can calculate the rigidity. Then this is the direct way to get the Young's modulus. So, based on that for all the different types of trinucleotides, you can calculate the Young's modulus this is the inverse of flexibility, this will give you the rigidity.

So, this is how we get this data. So, we take the set off nucleotide structures valuable the data bank, and cut into trinucleotides and get the dimension, how the change in the distance and we convert this, this values to calculate the strain and stress we know. So,

stress by strain is Young's modulus. So, we know the values. The Young's modulus that is Newton per meter square these are the values.

So, now if you see these values, then it will few nucleotides which are rigid and some of them are flexible for example, ATA its ATA right ATA what is the value for ATA? ATA this is 2.36 this is known to be flexible also if you take the C C or G G G that is rigid right C C C or G G G here, which is the value this is 6.07 now this is rigid.

So, if you compare with the non values the expected values you know, most of the match and the main problem from this table is some cases we have less number of data or now if you have more number of nucleotide structures available. So, we can use all these structures. So, you can rewind this, but some case it is it is 27 or 49, but some cases it is very less say 1 or 2. In this case the data are not statistically significant you have to go for more number of data right. But anyway you can use these as a source for the flow the stiffness of each trinucleotide. This is the one example for the hydrogen bonds the calculate of hydrogen bonds, this is for the flexibility or rigidity and also we have data for the base stacking energies what is the base stacking energy?

Student: (Refer Time: 10:52).

If you have the two bases stack to each other right what is the energy right for it is between this two bases right (Refer Time: 10:59). So, you can calculate this choosing any potential (Refer Time: 11:02) potential or any potentials, you can use the different types of interactions like van der Waals interaction or electro static interactions right and if you put this to bases and then you can calculate the energy between them, and estimate the you can estimate the base stacking energy.

So, here I show the data for the different dinucleotides though here this I use give the data for both the DNA and RNA. So, you can see A C G U T here also A C G U T because U and T are not coming together this why we put dash. So, for each cases A and A or A and C. So, A and G, A and U is 1.33 A and T is 1.83. So, you have the table for the base stacking energies fine. Now, we have different this tables also we have a set off values for different parameters. Because that may be next time next class when I discussed about data base I will show a data base which contains the information regarding different values.

Say hydrogen bond energy and the melting temperature and so on right fine. Now, if you have a sequence, how to calculate the average property of any DNA sequence that is a question. If we take any sequence, I will explained with two examples one is with the rigidity and one is with the base stacking energy.

So, if you take this sequence AT ATA C G G C A right here we need to divide this into overlapping segments here data are known for the trinucleotides right. So, you can classify into overlapping segments right. So, what are the different overlapping segments if we take the trinucleotides? ATA right ATA and from here TAT and from here ATA and here TA C and so on; so, totally how many nucleotides here? 1 2 3 4 5 6 7 8 9 10 11 nucleotides; 11 nucle. So, how many possibilities will have?

Student: 9.

Nine right. So, 9 trinucleotide steps right and then take the values from this table. So, what is the value for ATA? ATA is 2.36 right 2.36 then what is the value for TAT? The same 2.36 ATA.

Student: Same (Refer Time: 13:19).

Same 2.36 and TAC.

Student: 2.97.

2?

Student: 9 7.

9 7 right; so add up all the values then finally, you will get 31.74 and divide with the total number of possibility that is 9. So, 31.74 divide by 9, this equal to 3.53 into 10 power 8 Newton per meter square. So, you will get the exact num average rigidity that for this particular trinucleotides particular DNA sequence. Now if you want to have the base stacking energy. So, base stacking energy we have the data for tides are trinucleotides 9 nucleotides right where here we have data for trinucleotides, here we have data for dinucleotides.

So, we need to divide this DNA sequence into overlapping dinucleotides right how many dinucleotides steps you will have; N minus 1 so this equal to 11. So, totally we will get 10 . Now, how to classify this AT?

Student: TA.

Then TA, then ATTAA C, and so on. So, what is the value for AT 1.83 here 1.83 TA 1.83, AT 1.83 TA.

Student: 1.83.

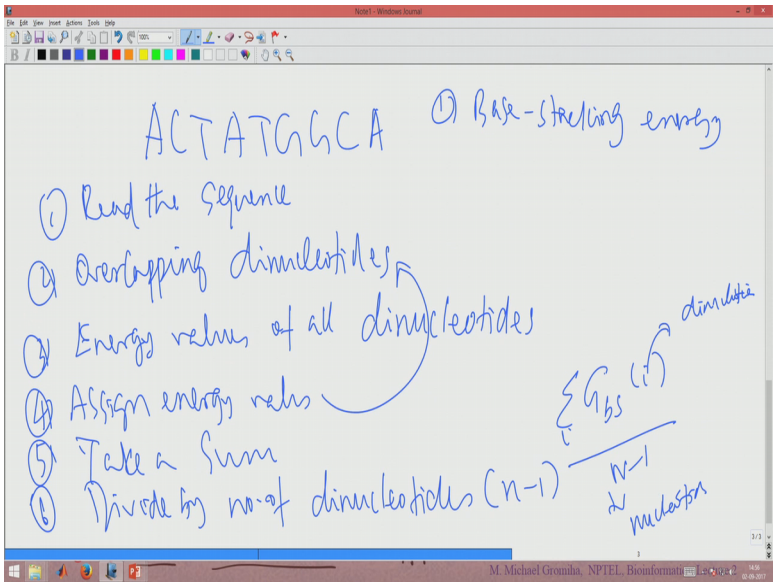
1.83 and AC.

Student: 1.73.

1.73. So, you add up all the values, that this will a that is 16.314 divided by this n minus 1. So, that will be 10 dinucleotides. So, divided by 10 this will give 1.631 kilocalorie per mol. So, now, if you have set of sequences, you can calculate the values and you compare the sequences which sequences are rigid, which sequence are flexible right and what about different types of energies. And from this you can calculate that where is structure base factors are whether this, a region is able to interact with the DNA this particular protein and so on right.

And you can see the DNA interactions or DNA drug interactions, DNA protein interactions and so on. So, these you can calculate right then how to derive this parameters using computational tool right.

(Refer Slide Time: 15:27)



Then we can develop a program to calculate this steps right how to do that? For example, if you have sequence (Refer Time: 15:29) sequence.

Student: (Refer Time: 15:31).

A.

Student: T.

T.

Student: G.

G G.

Student: C A.

C A; so how to calculate the base stacking energy? So, the question is how to get the base stacking energy, what to do first.

Student: (Refer Time: 15:54) sequence.

So, first you write the sequence right you need the sequence.

Student: (Refer Time: 15:59).

So, now the overlapping dinucleotides: so you have the overlapping dinucleotides right and then. So, we need the table right we need the base stacking energy values.

Student: (Refer Time: 16:19).

Energy values for all dinucleotides and then what to do? So, we have to assign the values right. So, the assign the values for each stacking from this a step two: so assign values.

Student: (Refer Time: 16:42).

Right in the step 2; this step right and then.

Student: Take a sum.

Take the sum take a sum and then.

Student: Divide by total number.

Divide by total number of.

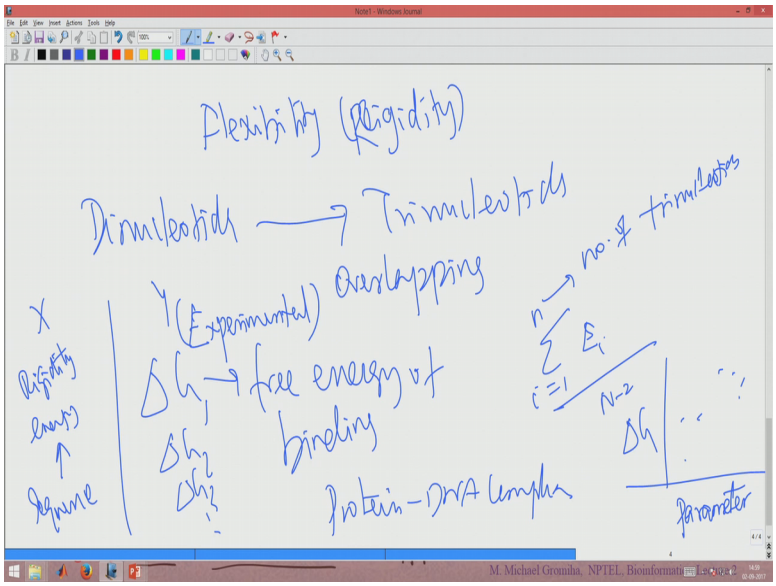
Student: Dinucleotides

Dinucleotides. So, what you call how many numbers of dinucleotides?

Student: n minus 1.

N minus 1, right. So, here this will give sigma, i they will take base stacking energy right. So, you can put G b s right of i. So, I is i is (Refer Time: 17:24) number of dinucleotides right. So, i is since for the dinucleotides. So, add up then here divided with this n minus one right. So, n is the total number of nucleotides.

(Refer Slide Time: 17:42)



So, I will get the values. Likewise, if you have the flexibility what to do similar way what to get the flexibility? The same way, right.

R rigidity; so in this case what to do? The only at the same procedure you have to follow only is the difference is.

Student: (Refer Time: 18:05).

The initial of di you have to move the tri right. So, you get the sequence and make the sequence into overlapping trinucleotide sequence say you have to give the overlapping trinucleotides right and then you have the table and assign the values and add the values right sigma E i right i equal to one to n, n is the number of trinucleotides and they divide with the n minus 2 because the trinucleotides say to n minus 2.

So, you can do the same way and then you can calculate rigidity. So, depending upon the data whether you have the data for the each nucleotide or they dinucleotides or they trinucleotides right you can assign the values and you have divide you have the overlapping dinucleotide nucleotides right, then you can you calculate rigidity.

So, why we need to derive this parameters what is the importance of this; if you to see the rigidity values, or the rigidity values have some information to do with the binding affinity. So, if you have the proteins right and the interactive the DNA. So, right if you

have the proteins and the DNA. So, you know the binding affinity. So, there are various databases available, which will give you the binding affinity of protein DNA complexes.

Sometimes we have the same proteins interact with different DNAs in this case the affinity is different. So, take the DNA sequence because protein is same. So, difference mainly comes from the DNA sequence right. So, in this case you have to look for the contributions from the DNA. So, you get the DNA sequence and then see the various parameters for example, rigidity or flexibility right and they how this flexibility or rigidity right varies with respect to the DNA sequence and we have the delta G values right there is a free energy of binding right and compare this delta G with the non sequences non e values.

So, for example, here delta G 1 delta G 2 delta G 3 for different complexes right protein DNA complexes. So, this is y. So, you have the x different parameters for example, rigidity right. So, you have the rigidity or any types of energies right you can compute from the sequence, this you can calculate from sequence this is the delta G, we know experimental this is experimental. So, we have the x values and we have y values because set off data. So, you can plot right this is the delta G, this is the any parameter right. So, you can use any parameter or any property right, say rigidity or the flexibility or the binding energy or hydrogen bonding energy.

So, then you can plot when you make a plot connecting the each parameter in delta G and you can see that which parameter fits very well with the banding free energy. So, this will helpful to understand; what are the various factors, which are important to understand the binding affinity of the protein DNA complexes right you can see that.

So, and the other hand you can do with the potential also, depending upon the different types of complexes. If the complex varies only the DNA side and mainly the DNA part is important if there are different protein complexes with the different proteins and different DNA, then you can use the parameters derived from the protein side as well as DNA side. DNA side the ways parameters right may be in the next class I will discuss about database which contains more than 50 parameters right.

So, we can use different parameters and see which parameter relates well with the binding free energy are with any combinational parameters we can combined different parameters which can be used to understand the binding affinity of protein DNA

complexes if you see some correlation right between the rigidity or the flexibility the delta G what can be infer.

Student: That is particular parameter (Refer Time: 22:13) properties.

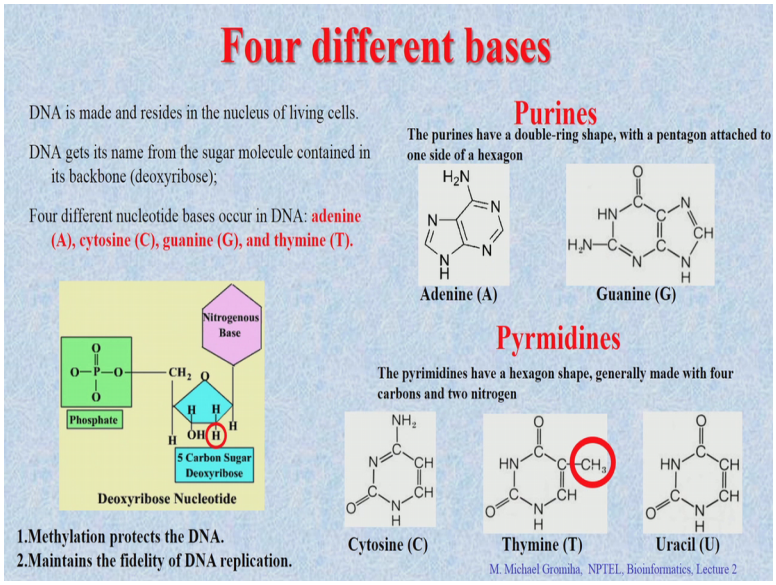
Correct.

Student: Now can be predict the property predict that property (Refer Time: 22: 21).

So, if you for example if rigidity or flexibility is important; so you can see the conformation change of this DNA right because of the flexibility, that mix more contribution that is very important to understand the binding affinity. So, if you change the DNA sequence if you mutate then you can see how the affinity is last right. Then you can see how for this affinity is related of the (Refer Time: 22:43) deceases right for example, if you take the 353 right 353 is mainly between the protein and DNA interactions.

Several mutations which finally, essentially lead to decease right for example, take a answer. So, if you see change this sequences right you can see how for the binding affinity changes right. So, then you relate with the different functions. Likewise you try to see various parameters right and see which parameters are important and which combinational parameters are important to understand the binding affinity of these protein DNA complexes.

(Refer Slide Time: 23:26)



As well as you can do the same (Refer Time: 23:15) put the RNA complexes, if we have the data for the RNA sequences. Till now what we discussed from the beginning just we will summarize, first we discussed about the basics of DNA right. So, DNA contents what are 3 different groups in the DNA phosphate.

Student: (Refer Time: 23:31).

Sugar and the?

Student: Base.

Base right. So, what is the difference between a DNA and RNA?

Student: Two prime (Refer Time: 23:38).

Two prime whether it is h o h and then?

Student: (Refer Time: 23:44).

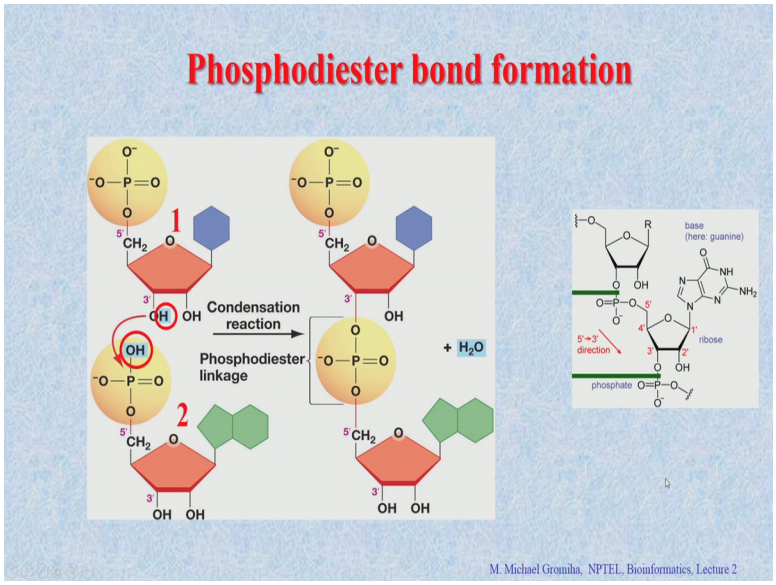
The base right the uracil or?

Student: Thymine.

Thymine. So, we have two different types; so classified into purine pyrmidine. So, purine has how many rings two rings and pyrmidines has.

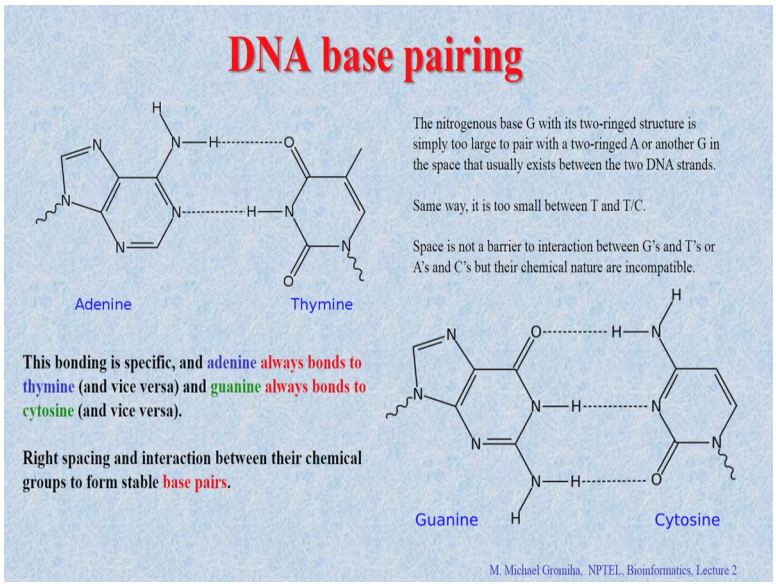
Student: 1.

(Refer Slide Time: 23:57)



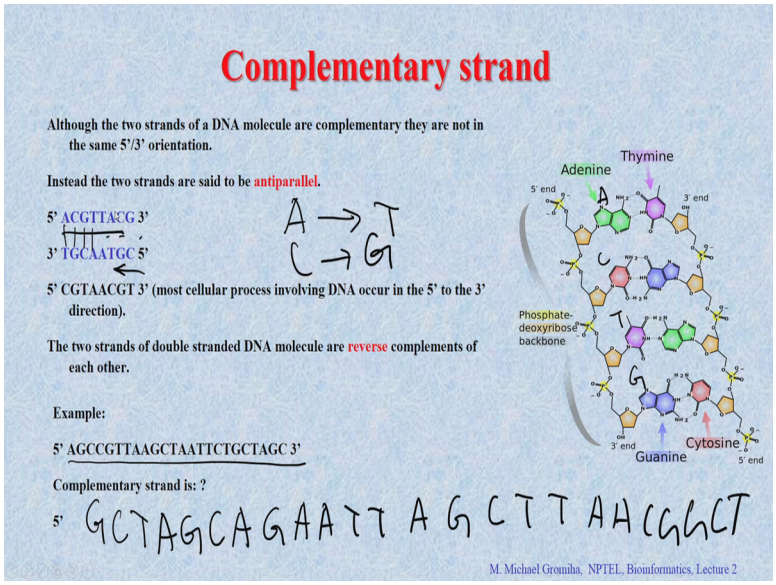
One (Refer Time: 23:55). So, this is thymine and uracil these difference between the DNA and RNA and then form the diester phosphodiester bond they have to estered bonds so this we have they have the sequence.

(Refer Slide Time: 24:04)



Then why are these pairing is between A and T and G and C, aspace constraints as well as a plus the chemical groups right these are two different aspects why these pairing is important.

(Refer Slide Time: 24:16)

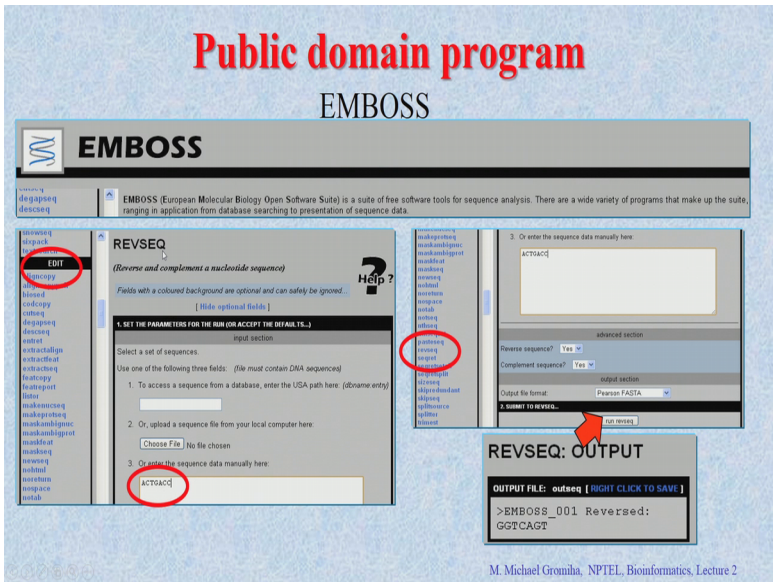


So, if you have a single strand how to get the complementary strand.

Student: (Refer Time: 24:19).

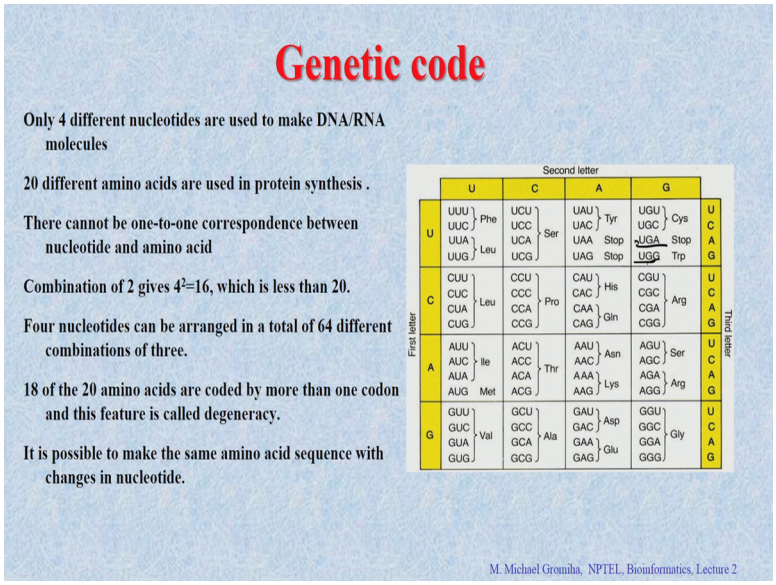
So, you have to get the reverse inverse complementary, you have to get the complement and a reverse it and then you will get the reverse complementarity right.

(Refer Slide Time: 24:28)



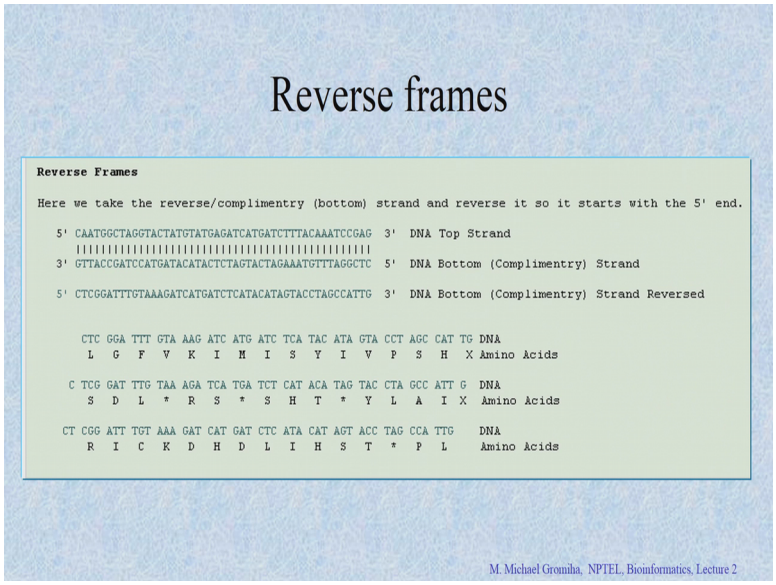
So, there are several programs available in the literature right you can the emboss is one of the suite.

(Refer Slide Time: 24:37)



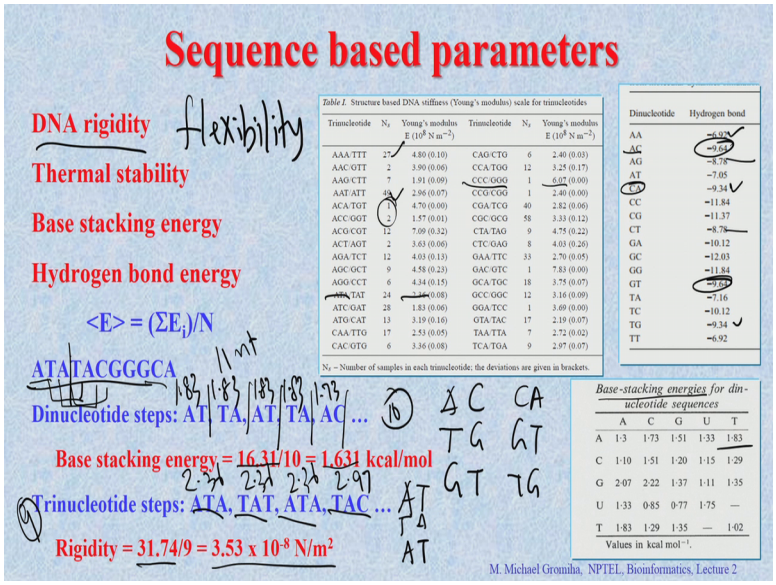
So, we can use it for this one, right. Now we have the transcription and the translation. So, you get the DNA sequence or protein sequence, right.

(Refer Slide Time: 24:41)



And then you have the sixth different frames 3 forward frames and 3 different frames.

(Refer Slide Time: 24:47)



And then we try to see some of derive some of this structures sequence based parameters. If they available then we can get it and then you can use it for the applications. Try to understand the functions, fine.