

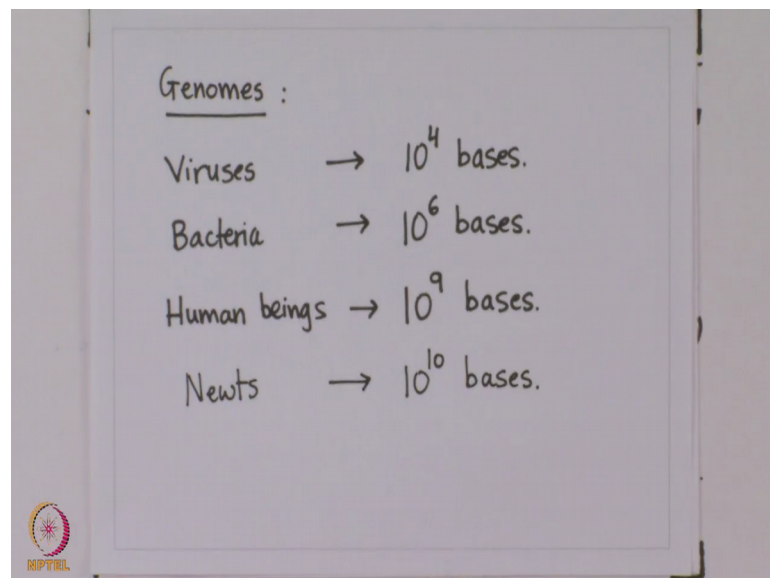
Introduction to Evolutionary Dynamics
Prof. Supreet Saini
Department of Chemical Engineering
Indian Institute of Technology, Bombay

Lecture – 12
Genetic Code & Sequence Spaces

Hi, I am welcome to the next part of the course. We have just concluded our discussion on the three tenets of evolution which are reproduction, selection and mutation. From this we move to the next chapter of the book that we are discussing by (Refer Time: 00:32) and we are going to be talking about sequence spaces and fitness landscapes associated with genotypes or protein structure.

So, we need to define a few things before we start. But before we get into the definitions we want to understand what are the relative sizes of the genomes of different organisms that we see around us. So, if we are to get to a sense of the genomes sizes associated with different organisms let us understand those numbers and their relative magnitudes compare to each other.

(Refer Slide Time: 01:00)



A photograph of a whiteboard with handwritten text. The text is organized into a table-like structure under the heading 'Genomes :'. It lists four organisms and their genome sizes in bases, using powers of 10. The organisms are Viruses, Bacteria, Human beings, and Newts. The genome sizes are 10^4 , 10^6 , 10^9 , and 10^{10} bases respectively. In the bottom left corner of the whiteboard, there is a small circular logo with the text 'NPTEL' below it.

Genomes :	
Viruses	$\rightarrow 10^4$ bases.
Bacteria	$\rightarrow 10^6$ bases.
Human beings	$\rightarrow 10^9$ bases.
Newts	$\rightarrow 10^{10}$ bases.

So, genomes viruses bacteria, human beings; viruses the typical sizes of a viral genome of this very simplest once can be of the order of and we are only going to be interested in very crude order of magnitude estimates of these numbers and we are not interested in

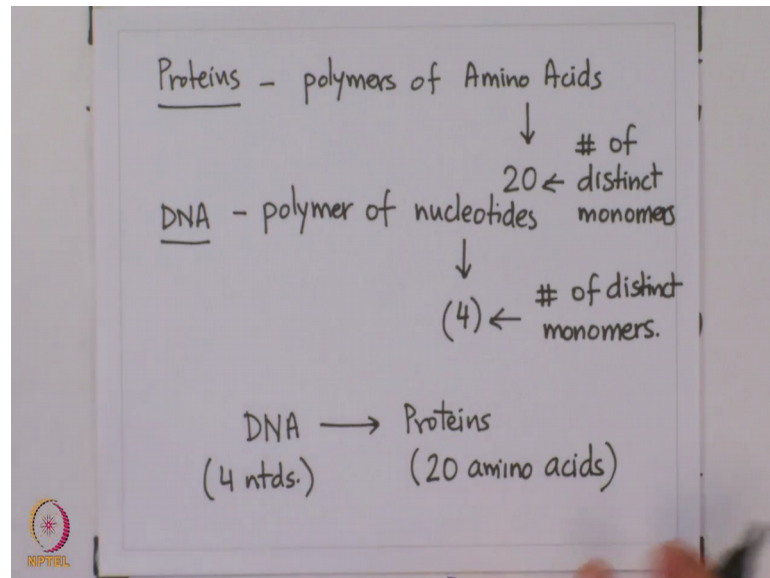
the excruciating detail. For instance human genome is three into 10 power 9 billion bases, but we are for our purposes 10 to the power 9 is a good enough approximation.

So, viruses, some of the simpler viral genomes are of the order of 10 to the power 4 bases. Bacteria we have been talking about E coli and many of the things that we are going to do later on in the course are going to be based on our understanding of the work cause of microbiology which is E coli and E coli has a genome of about 4.5 into 10 power 6 bases and again because we are interested in only order of magnitude numbers this is 10 to the power 6 bases, in human beings is 10 to the power 9 bases.

But examples can be counterintuitive, these three examples here seem to suggest that as complexity of an organism increases its phenotypic complexity, its physiological complexity increases, the genome size increases to. So, it gives you that impression, but in general that is not true. There are examples with organisms which are much smaller in size which have not evolved to be of a particular size, but have much larger genomes. And one of the examples that is listed in the book is for newts which have a genome of 10 to the power 10 bases.

And those of you who are who are (Refer Time: 03:12) fans would immediately associate newts with a gusyfink (Refer Time: 03:17) who is obsessed with studying newts. So, complexity of an organism its size of an organism does not always correlate with its genome size, it is a very rough rule at the very best. So, we have this numbers and of course, the protein numbers also very widely between these species as we are talking about their genomes.

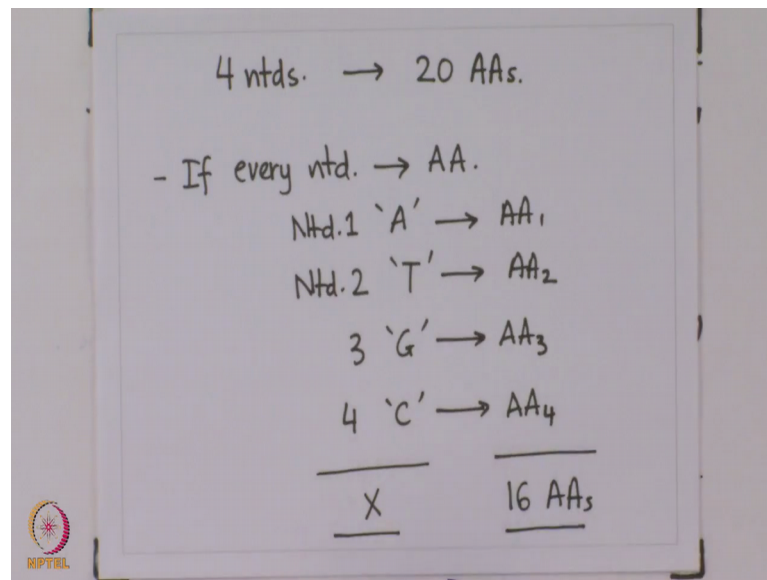
(Refer Slide Time: 03:46)



Proteins which are polymers of a minima acids and amino acids are 20 natural amino acids that comprise these proteins and of course, DNA is a polymer of nucleotides and the number of monomers is 4. So, think of these as the number of distinct monomers this is the number of distinct monomers when it comes to the DNA polymer.

And for a large time this was a puzzle as to how does information which is stored in a polymer with 4 monomers get eventually translated into a polymer which is comprised of 20 monomers. So, we have information going DNA encodes the information for what proteins are going to be made inside an organism cell and this is a polymer of 4 nucleotides and this is a polymer of 20 amino acids. So, how does this happen? This was a big puzzle for a very long time.

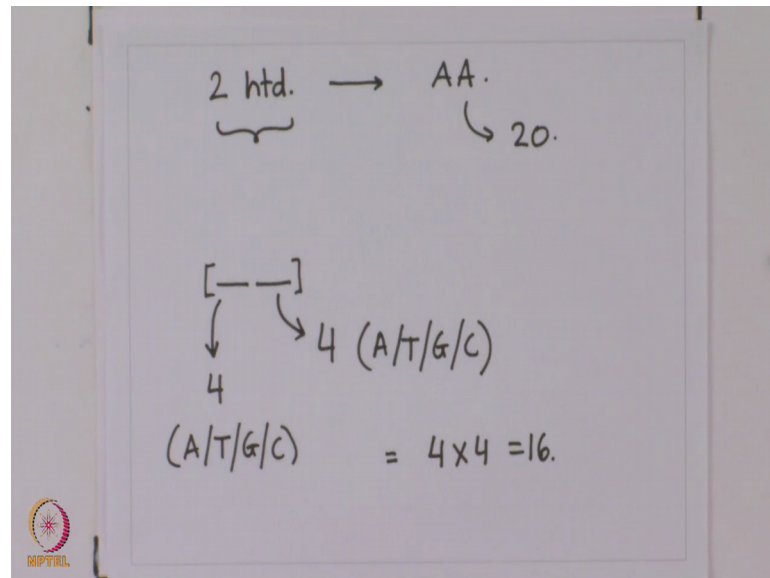
(Refer Slide Time: 05:30)



Let us look at this in some detail we have 4 nucleotides and 20 amino acids AA is what stand for amino acids. Now if every nucleotide, if every nucleotide corresponded to an amino acid we would only go up to 4 amino acids because nucleotide 1 let say A would correspond to amino acid 1, nucleotide 2 say T would correspond to amino acid 2, nucleotide 3 G would correspond to amino acid 3 and nucleotide 4 C would correspond to amino acid 4.

We have run out of nucleotides because DNA is comprised of a polymer which is only 4 distinct type of monomers, but we still have, so there is nothing left here, but we still have to assign nucleotides to 16 distinct amino acids. So, this scheme does not work. So, what could work? Let us built the complexity associated with this association and now let us imagine a scenario where every doublet two nucleotides correspond to one particular amino acid. Does that solve our problem? Let us see.

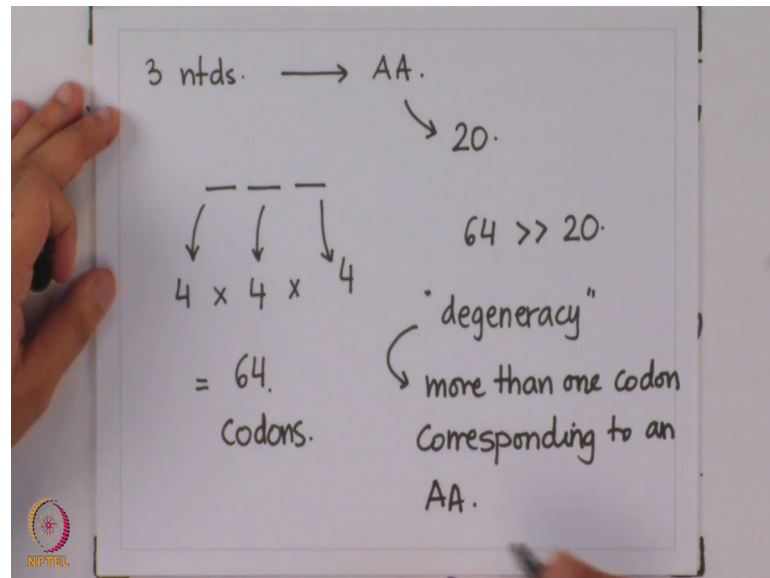
(Refer Slide Time: 06:58)



So, now, we have two nucleotides a pair correspond to an amino acid these are 20 number, we want to see how many distinct nucleotide pairs can we make using two nucleotides at one time. So, we have to fill two places and when we do that how many options to we have to fill the first place of the pair which is 4 because we could fill the first place by any of the 4 monomers and when comes to the second place again we have the same 4 options. Same 4 options of the nucleotides to be allotted to the second place and that would give me the pair that I am talking about.

The total number of nucleotide pairs that I can generate is 4 into 4 which is 16, but 16 is still less than 20. So, even this scheme of two nucleotides in a sequence corresponding to an amino acid does not really work. So, we need to go another step and let see what 3 nucleotides corresponding to an amino acid do.

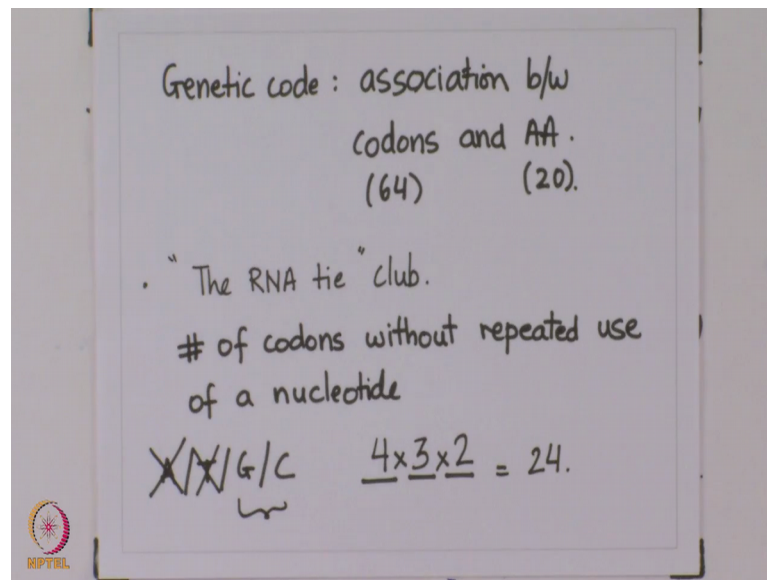
(Refer Slide Time: 08:17)



16 is close to 20. So, this should hope we do it. So, now, we have 3 nucleotides corresponding to an amino acid and this to fill the first place in this we have 4 options, the second place we have 4 options, this place we have 4 options. So, number of triplets distinct triplets that we can generate from DNA sequences where each triplet is 3 nucleotide bases is 4 cross 4 cross 4 which is equal to 64 and the number of amino acids that have to be allotted these codons is 20.

Now 64 is obviously, much larger than 20 so that means, there is what is called degeneracy which is, so these are called 64 codons we have more than one codon corresponding to an amino acid and this of course, forms the bases of what is called the genetic code which is the association between codons and amino acids which are 64 in number and these are 20 in number.

(Refer Slide Time: 09:36)



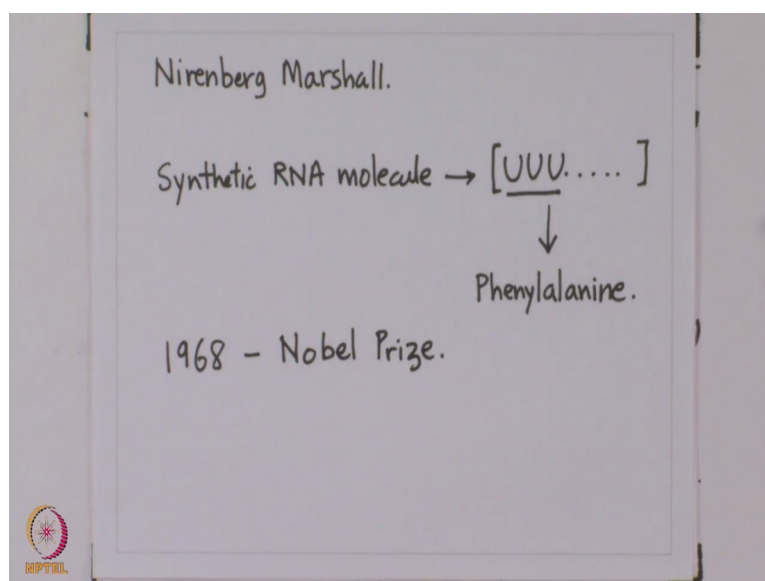
And for a long time this remained a puzzle. How is this correspondence between 64 codons and 20 amino acids made in 11th cell? Watson and Crick, particularly Crick after the DNA structure paper in 1953 was working on this problem and they formed it is a very well known club which is called the RNA tie club. This was formed in the 50s and try to decipher the relationship between 64 and 20. How does a number 64 which is much bigger than 20 correspond to 20 amino acids? It biology would have posed a very neat solution if there were only 20 codons each one of those codons corresponding to a distinct amino acids, but the fact that there were 64 and 64 of these somehow corresponded to 20 posed a problem that these people belonging to the RNA tie club, film and Crick Watson were members of this club, and each one of them the RNA tie club have had 20 members, each one corresponded to an amino acid and their discussions with in this club were here towards figuring out the code going from 64 codons to 20 amino acids.

And the popular some of the approaches that they try was codons with repeated nucleotides are not going to be allowed only codons with distinct nucleotides are going to be allowed. So, let see how many triplets or codons can be make which have distinct nucleotides only. So, repeat of a nucleotide in a codon is not allowed now. So, we are interested in number of codons that we can make without repeated use of a nucleotide. So, our monomers are A T G and C and we are interested in forming a codon, but

repeated use this time is not allowed which was one of the approaches that people in RNA tie club thought would lead them to the answer towards the genetic code.

So, the first position we could have any one of the 4 nucleotides. So, we have 4 options. The second position now because we have used one of the 4 up it could be A, it could be any of the other three, only leaves us with three options leaving out the one that we have already used in the first position that is 3, and now for the last position in this codon we are left with only two choices. So, we have to 4 cross 3 cross 2 which is equal to 20 4 still not quit equal to 20. This number still exceeds 20 and of course, it was understood that one of the codons or more than one of the codons also has to correspond to the stop signal which tells the translation machine read that this is the point where translation should end and the polypeptide synthesis, polypeptide chain synthesis should stop. But even if we are count for one codon to be a stop signal that still leads us only 21 codons, but we have 24 distinct codons.

(Refer Slide Time: 13:30)



So, this was a big mystery and the and the answer was proposed by somebody called Nirenberg Marshall and the story goes that there was a biochemical conference in 1961 in Moscow and Marshall what he had done was he had done a synthetic experiment where he had synthesized synthetic RNA molecule and this synthetic RNA molecule was special that it only comprised of use. This was just a poly U, a poly U RNA molecule and when this was added in a test tube with along with all the amino acid sequences. So, he

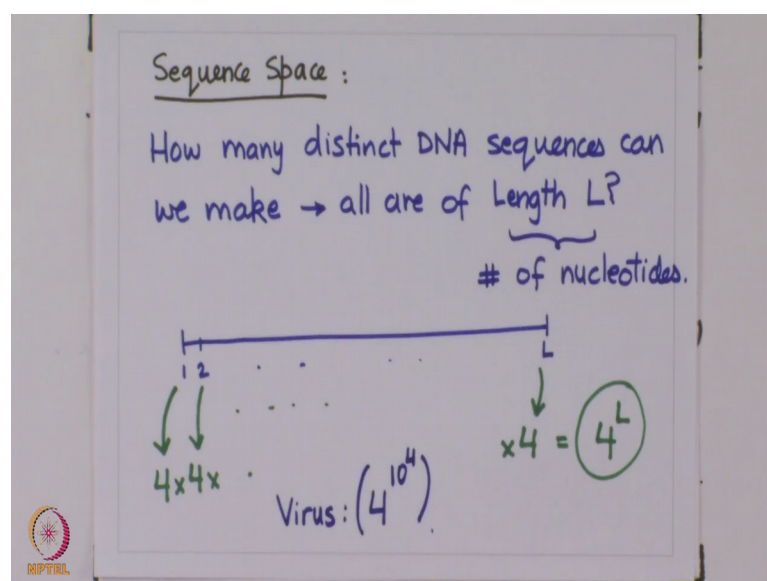
took this RNA molecule and added it to a test tube with ribosomes and with amino acids all 20 of them, the only polypeptide chain that he obtained on adding this was phenylalanine. This gives Marshall the clue that U U U corresponds to the amino acid phenylalanine and this was the beginning of discovery which led to us finding about the genetic code.

And when Marshall presented this result of his in 1961 in our conference in Moscow, Crick became aware of the top that he had given and he sort of changed the schedule associated with the conference that had Marshall present his results again. The following day and this time he is talk was attended by more than 1000 people, and this triggered the race among people in the world this was a 1961 because this was just one of the nucleotide, there were many other codons whose correspondence was yet to be figured out and then this just triggered a race among people around the world trying to figure out this association between codons and amino acids. And this was in 1961 and eventually for this work Marshall in 1961 received the noble prize along with Hargovind Khurana.

So, let us some (Refer Time: 15:54) for how the genetic code was came to be discovered. Let us move on and let us try and understand what is referred to as the sequence space which is associated with the genetic or a protein sequence.

Let us talk a genetic sequence first.

(Refer Slide Time: 16:12)

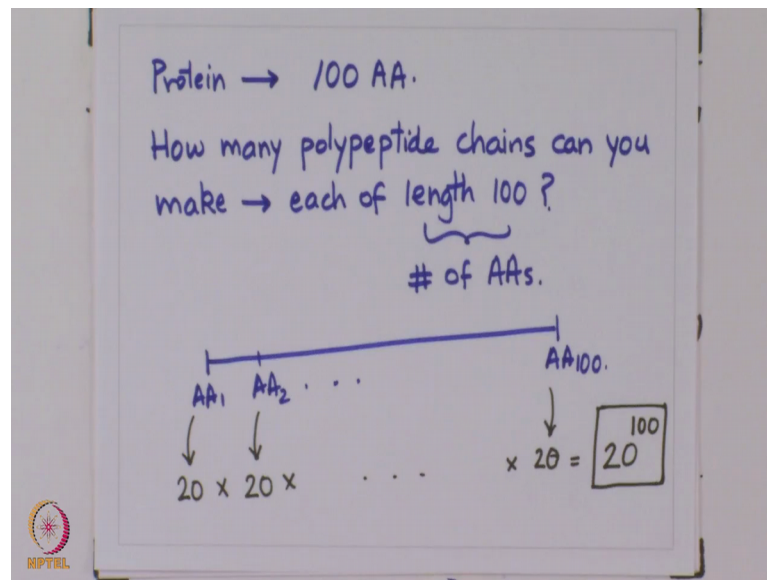


So, we are interested in finding out what is the sequence space. Let us try and answer this question how many distinct, how many distinct DNA sequences can we make such that all are of length L , that is the question that we want to answer. So, you are asked to synthesis a linear DNA strand of length L and when we say length L , this is length is typically just represented as number of nucleotides. So, length is not the conventional length as we think of it, but it is actually just the number of nucleotides associated with the DNA sequence. So, what we are interested in is that how many distinct sequences can we make and these are linear pieces of DNA and each length of DNA is comprised of L nucleotides.

So, I suggest that you pause the video for a minute or, so think about this problem and try and come up with an answer to this problem. So, how do we go about thinking this? We are required to make this linear piece of DNA which is nucleotide 1 2 going all the way up to L and we are interested in how many such sequences can we make. And it is really trivial counting problem when I am allotting nucleotide to the first position how many options do I have in terms of allotting a nucleotides to position number one, position number one can be filled with any of the 4 nucleotides. So, I have 4 options in terms of which nucleotide can go in position 1 and where it comes to position 2, I have the same 4 options again because repeat of nucleotides is allowed here and so on and so forth going all the way up to L , when it comes to allotting a nucleotide for the L th position again I have 4 options.

So, the total number of distinct DNA sequences that I can make which are of length L is 4 into 4 into 4 going all the way doing this L times. So, this is equal to 4 to the power L . So, if a organisms genome is of length L there are 4 to the power L sequences which are also of the same length made from the 4 nucleotides and have the same length has that organism. How big is this number? Remember we started this lecture by getting at sense of the typical sizes of genomes. So, even for a virus with the genome of size 10 to the power 4, this is 4 into 10 to the power 4 sequences that many distinct DNA sequences can be made.

(Refer Slide Time: 19:32)

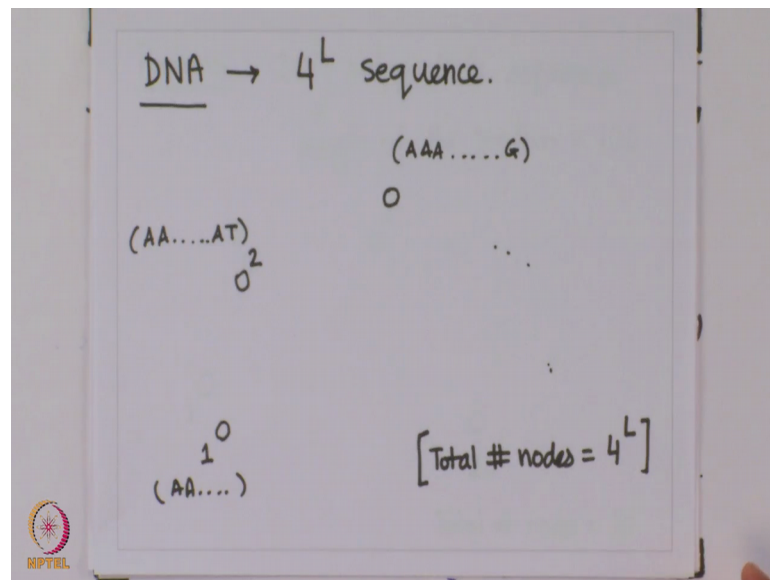


Let us think in terms of proteins, let us imagine that a protein is comprised of 100 amino acids and the question is how many polypeptide chains can you make such that each is of length 100 and length again here is number of amino acids. So, again this is similar approach you have you have a linear chain this is amino acid 1, amino acid 2 going all the way up to amino acid 100. And how do you count the number of such polypeptide chains that you can make.

For the first position you could have any of the 20 amino acids allotted to the first position. So, you have 20 options here, you could have you have the same 20 options because repeat of amino acids is allowed going all the way up to the 100th amino acid which is also 20. So, the total number of polypeptide chains that you can make such that each chain is of length 100 is equal to 20 cross 20 a 100 times which is 20 to the power a 100, right. These are very very large numbers.

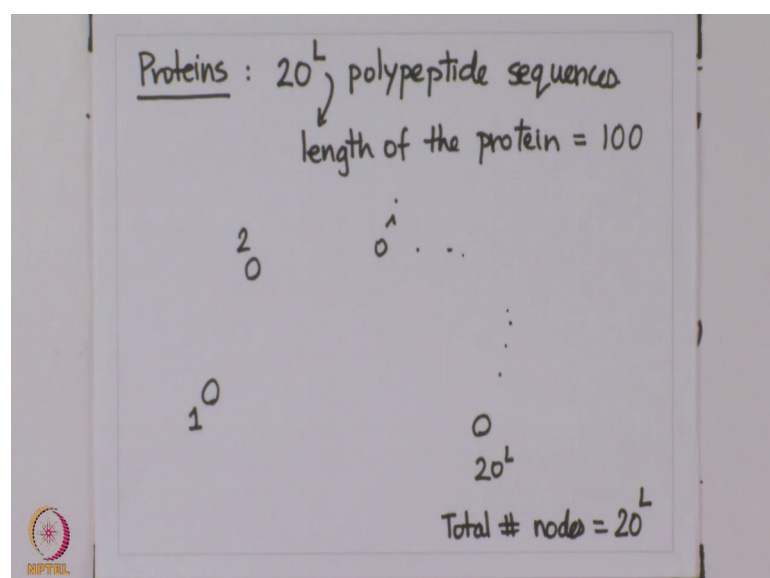
So, the number of sequences possible which are equivalent to the length of the simplest organism that we know is 4 to the power 10 to the power 4 and the number of polypeptide chains for even a smallest protein which is of length 100 is astronomical this is a very very large number.

(Refer Slide Time: 21:49)



Let us now think in terms of DNA and try and arrange them in a particular way. So, when we think in the context of DNA we have 4 to the power L possible sequences, 4 to the power L sequences which correspond to a DNA fragment of length L. Let us represent each one of these sequences by a node. So, this is sequence 1 this may be is all A's, this is sequence 2 may be this is all A's, but the last nucleotide is AT, this is sequence 3 may be this is all A's, but the last one is G and so on and so forth. So, each node here represents the one particular sequence, 1 of 4th to the power L particular sequences. So, this total number of nodes here on this plane will be 4 to the power L.

(Refer Slide Time: 22:55)



And similarly in the context of proteins we have 20 to the power L polypeptide sequences where L now is the length of the protein. Again represent a node let us call this node 1 this represents the first polypeptide sequence of length 100, if L was equal to 100 as was the case in the example that we talked about. This represents the second polypeptide chain upon all the 20 to the power L , polypeptide chains of length 100 and so on and so forth going all the way up to 20 to the power, going to the node which corresponds to 20 to the power are L . So, the total number of nodes here 20 to the power L , L in this case is the length of the protein that we are interested in.

So, what we have done here is defined that given as DNA sequence of length L or a protein sequence of length L , how to enlist and enumerate all possible sequences of DNA and peptides which correspond to the same length as we are interested in. If we are interested in DNA sequence of length L then the total number of DNA sequences which can be created which are of same length, but any sequence the choice of nucleotides can be different, but the sequence size has to be the same the answer is 4 to the power L distinct sequences.

In terms of proteins we have, if we have a protein of length 100 and we are interested in how many polypeptide chains can we make by any by free use of any peptide, but constraining the size of the amino acid chain polypeptide chain to a 100 that answer is 20 to the power a 100 or 20 to the power L more generally where L is the number of the nucleotide; L is the number of the amino acids which comprise the protein. And these possible number of solutions to this answer 4 to the power L or 20 to the power L constitute what is called the sequence space associated with DNA or protein sequences.

So, in the next lecture we link sequence spaces with what are called fitness landscapes and understand what can they tell us about mutations and populations moving about them.

Thank you.