Applied Environmental Microbiology

Dr. Gargi Singh
Department of Civil Engineering
Indian Institute of Technology, Roorkee

Lecture - 57
Bioinformatics II

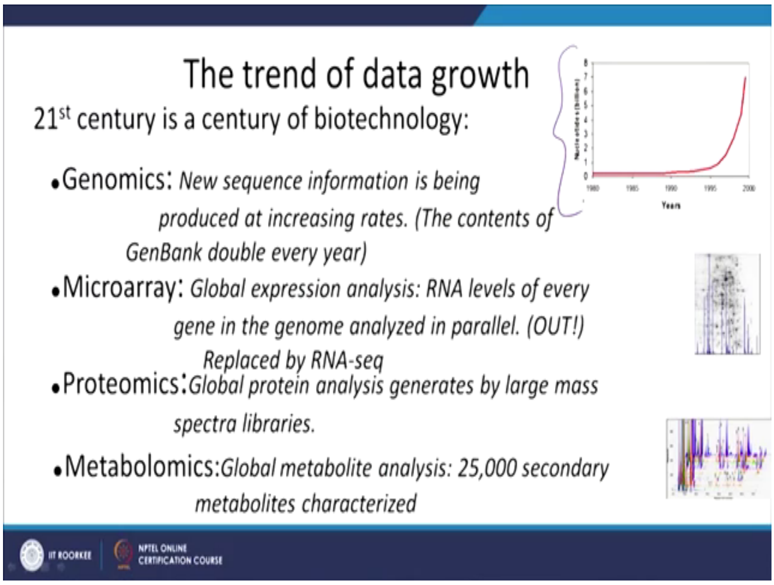
Dear students, welcome back to our course Applied Environmental Microbiology. Today we will continue a conversation on Bioinformatics. So, previously we talked about what bioinformatics is why we need it and we revised? What we have already studied on Meta genomics? And why Meta genomics is important for us as a memento, scientist students and engineers and why we cannot make sense out of Meta genomics? Why it will not help us unless we have strong mathematical and computational tools to make sense out of her data.

Today, we will continue this conversation further try to understand about different kinds of options we have in bioinformatics to make sense out of her data. So, let us get started already.

So, we one thing that we mentioned in previous lecture was that the price of sequencing has dropped steadily following Moore's law. In fact, even faster than Moore's law in last detail definitely and before that and every now and then every few years we are noticing we are getting a new generation of sequencing techniques. For example, after the fourth generation sequencing technique which uses PH instead of optical measurement? The next generation technique are actually directly measuring, the directly measuring the change created by cleaving cleavage of a particular D DNA, but a particular nucleotide by passing it through pores.

So, Nanopore would be the next generation sequencing technique, and not only has the price reduced not only had the techniques improved, but; obviously, the amount of data we generate has increased very fast.

(Refer Slide Time: 01:57)



The trend of data growth

21st century is a century of biotechnology:

- Genomics: *New sequence information is being produced at increasing rates. (The contents of GenBank double every year)*
- Microarray: *Global expression analysis: RNA levels of every gene in the genome analyzed in parallel. (OUT!) Replaced by RNA-seq*
- Proteomics:*Global protein analysis generates by large mass spectra libraries.*
- Metabolomics:*Global metabolite analysis: 25,000 secondary metabolites characterized*

.

So, this is a very important panel here on the right top corner of the slide, that the amount of nucleotide data in billions that we have created in past decades is enormous and it just continues to rise.

So, it is safe enough to say that 21st century is a century of biotechnology and when we say about biotechnology we are not referring nearly to the department of biotechnology the subjects there basically the technology that related to life. Now conventionally biotechnology means we are using techniques to change life to affect life in a way that it suits our purpose. For example, I might modify microorganisms. So, that they make insulin and now I do not have to sacrifice thousands of sheep's to make insulin dose.

Similarly, same is true with vitamins and many other proteins and amino acids that we require for medical purposes and for other civil purposes. So, generally this is what we refer to as when we are talking about biotechnology, in this particular slide we are also talking about all the tools that are either used on living or living beings, biological beings, or use biological agents as tools.

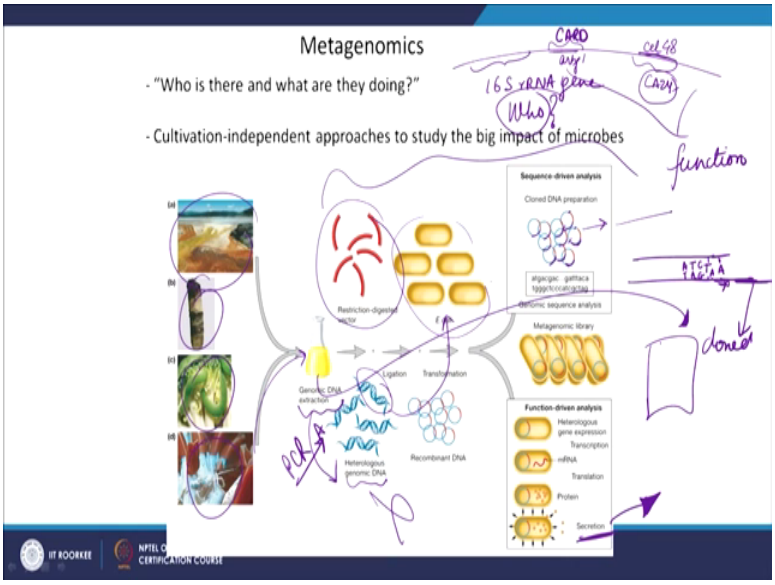
For example biosensors which we did in previous lecture all of them can be collectively called as biotechnology all right sometime. Now let us revise revisit some of the words that we have phrases and terms that we used earlier genomics. So, remember genomics. So, whenever you have or makes you means all. So, whenever we can sequence all the genes in your sample we call it genomics. When you are referring to the all genes and their analysis it is genomics and now here is the thing genomics does not only mean genes, because there are non-genetic area stood and non-genetic regions non coding regions in the chromosome in plasmid. So, including all of that now here is the thing new sequence information is being produced at an increasing rate, the contents of gen bank double every year this is very fast not crazy fast growth.

Next is microarray even in biosensors I briefly covered microarray and, but microarray is not used a lot now it is out of the picture, now it has been replaced the RNA seq. RNA seq will help me sequence RNA directly and I do not need to have a microarray to evade the hybridization there is a good point to mention here that microarray array was typically used for rapid detection of RNA.

Now, if you remember the central dogma you will remember RNA is a better representative of what is really happening right now whereas, DNA also includes the potential for future and the baggage of history. Now in proteomics that is analysis of proteins all proteins global protein analyses generated by large mass is generated by large mass spectral libraries metabolomics we are looking at meta metabolites, global metabolite analysis 25 secondary metabolites have been characterized that is an immense large number.

So, I hope this gives you an idea of how much data we are generating every year. Now yeah now doubles every year you have to understand it grows really fast genomic information. Most of the environmental scientists they nowadays focus on genomics proteomics and metabolomics mostly genomics and proteomics.

(Refer Slide Time: 05:23)



Metagenomics

- "Who is there and what are they doing?"
- Cultivation-independent approaches to study the big impact of microbes

So, Meta genomics I mentioned this in previous lecture. So, let us reuse it again it does me who is present and what they are doing? So, because I have the genetic sequence I can generate I can generate assemblies.

Let us say this region is your 16 S r RNA gene region and this is the sequence that you have created this is a sequence that you have created. So, we can align it to a database and then we can get information of who is present. The other way is also finding of taxonomic identity of the sequences and the assemble, assemble scaffolds or assembled contents or assembled secure genomes.

But, this is the most popular one and we can also look at the other genes that are present for example, if I am interested in antimicrobial resistance I might say align this with the database called card and comprehensive antibiotic resistance database. Then I will know what kind of antibiotic resistant genes are present. If I am interested in cellulose degradation I might look at cazy carbohydrate this is this cazy database involves all enzymes all the genes that are involved with carbohydrate degradation.

And, then I might get genes what genes are present here that are associated with carbohydrate degradation. So, not only do I know who, but I also get information of their function basically what are they doing already. Now the beauty of Meta genomics as we mentioned earlier is that dis cultivation independent approach. So, we do not need to cultivate it we do not wait because ninth remember from last lecture 99 percent of micro communities we suspect members of micro communities are non-cultivable even if they are viable.

So, the survival environment would weaken grow them in the lab despite so, many advances in microbiological tools. So, Meta genomics helped us there for example, we can take samples from very different very diverse microbial communities and we can extract their genomic DNA, DNA extraction are very simple process. Nowadays because we have a kit from most kind of samples like you have solid samples you have soil samples, you have faecal matter, you have blood samples, you have a tissue sample, your bacterial sample, you can you can take all kinds of samples and you will have a kit for it and then you can extract DNA.

And after you have extracted DNA what we do is we get heterologous genomic DNA and then this can be broken oh well this is the cloning part by the way. So, what we can do is this we have restriction digested vector. Now, this restriction digested vector will make sure that these DNA, DNA fragments now will be transformed into the competent equalizers.

Now, one thing you want to know because here we are doing cloning typically there's another step here, which is from the we do not we do not like to see usually we do not like to clone the entire sequence what we will do is will amplify region of it. So, we will use PCR and the kind of cloning kit that I am familiar with we leave a sticky end with an a overhang. So, basically it is like at the end if you have here t we have a we have g c t a a t. So, at the end we believe and a overhang there's nothing here to match it a nothing.

So, this a end is a sticky end and this will attach with your vector. So, your vector will attach here and now your vector is attached to it. Now this can easily be cloned into competent cells. So, we have cloned into competent cells, now your competent cells are expressed the protein you are interested the protein might be secreted. So, you can extract this secretion.

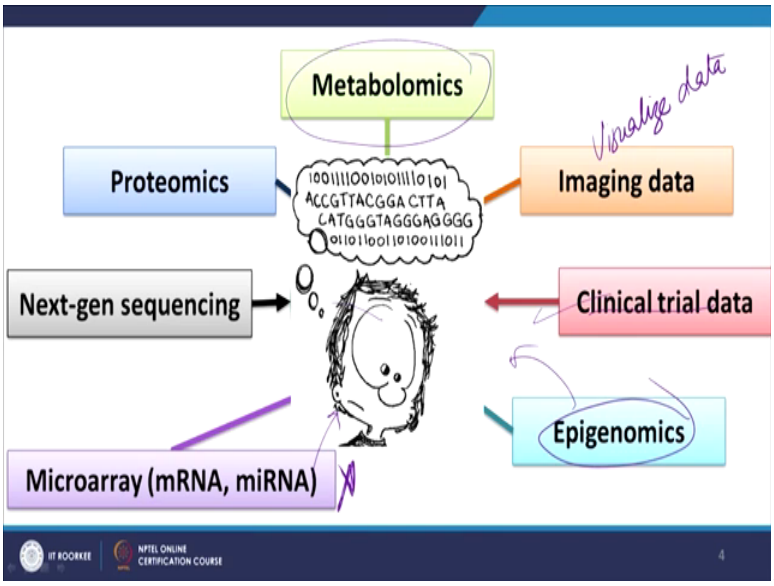
For example, the protein could be some of medicinal values when you want to do it at high quantities. The other possibility is that when you have done the cloning what you can do is remember the red one is your vector and the blue one is here the sequence that you wanted to clone. Now, each of your successfully transformed equalize cell will carry the same thing here like this.

Now, let us say you want sequence this. Now in order to sequence this, what you can do is you can amplify regions of this particular plasmid and when and you can sequence them. Now here you will get very high fidelity in sense that you will get very good results for your sequencing if this is useful for typical sanger based sequencing technique in meta genomics we skip all these steps.

We do not need to do this remember cloning you take nearly one day to set up your cloning, and then you have to leave them overnight, and then you send them for sequencing, and then you get your data and then you look at your data using your tools it takes at least 3 4 5 days depending on what kind of resources are available to you and whether any way.

But in Meta genomics we skip all these steps and we go from genomic DNA extraction directly to quality check quality control and then you can send it to your sequencing agency and then they will do the next steps for you, which do not look like this already.

(Refer Slide Time: 10:19)



Metabolomics

Proteomics

Imaging data

Next-gen sequencing

Clinical trial data

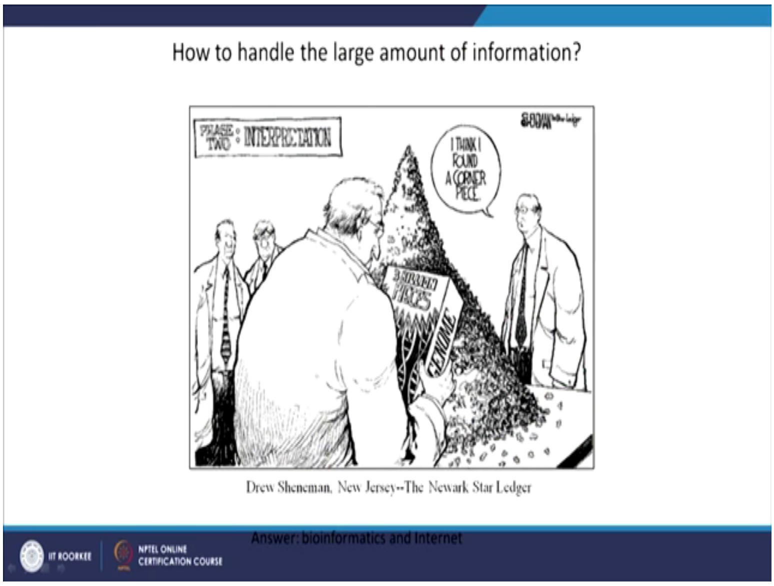
Epigenomics

Microarray (mRNA, miRNA)

Now, the trouble of today's world is that we have so, many omics around us we have metabolomics that is giving us information on what metabolites are present, proteomics that is (Refer Time: 10:28) us about proteins.

Next, generation sequencing and doing a lot of Meta genomic data microarray not popularly used now anymore, but even microarray kind of techniques give you a lot of data epigenomics, we have talked about them. They give you data clinical trial data and now you have and then you need to also visualize your data. So, all this data is very high to visualize as a result we come in fix.

(Refer Slide Time: 10:54)



How to handle the large amount of information?

PHASE TWO : INTERPRETATION
I THINK I FOUND A CORNER PIECE.
3 BILLION PIECES
GENOME
IIT ROORKEE
NPTEL ONLINE CERTIFICATION COURSE

Drew Sheneman, New Jersey--The Newark Star Ledger

Answer: bioinformatics and Internet

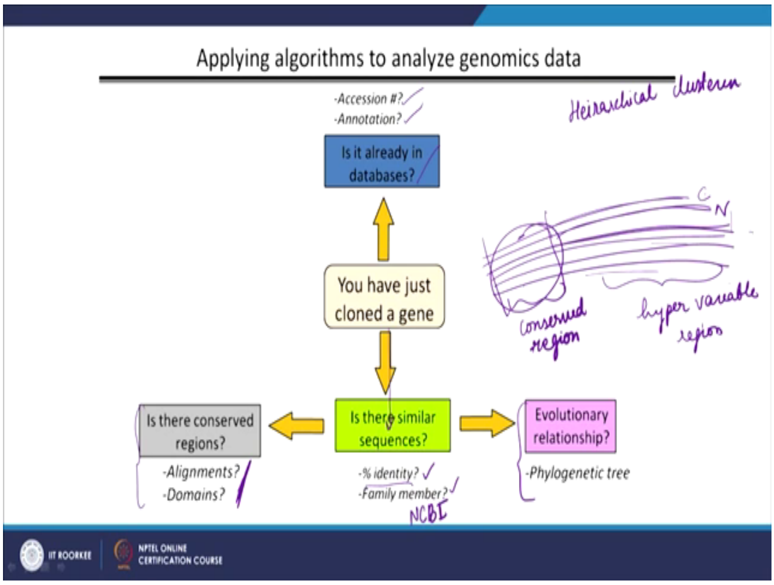
.

So, how do we handle this large amount of data that is a very very pertinent question and this is a very nice cartoon.

Here you have a puzzle that you need to fix with 3 billion pieces. So, you have seen the jigsaw puzzle we have to fix them and it is really impossible to even understand where to get started from and here is a joke when a scientist is saying I think I found a corner piece the answer is bioinformatics and internet. So, if you can use bioinformatics and the computer will decide, which pieces go well together remember what I told you about sequencing and assembly in the previous lecture you create millions of these sequences.

So, these are your millions of sequences or 3 billion pieces. So, you have created 3 billion sequences you do not know how they fit together? How they fit together how can they be stitched together to great to create greater length of sequences, do not have to worry about finding the corner piece what you can do is you can use bioinformatics tools and the internet page tools and what they will do is they will stitch it for you and you if you sequencing data is good enough you might even get the whole genome sequence.

(Refer Slide Time: 12:10)



Applying algorithms to analyze genomics data

-Accession #?

-Annotation?

Is it already in databases?

You have just cloned a gene

Is there similar sequences?

-% identity?

-Family member?

Is there conserved regions?

-Alignments?

-Domains?

Evolutionary relationship?

-Phylogenetic tree

Now, how does this work let us say you have just cloned the gene using the typical way, then you can what you can do is you can check out if it is already present in the database then what said what is it is accession number? What is it annotated after that you can look are there similar sequences in the environment that other people have reported? What is the percent identity?

What is the family member and one of the lectures on bioinformatics we will be going to NCBI and I will show you how to find out person identity and find out about family members. There is a thing if a sequence that I have cloned is very similar to another sequence that is already known. So, I know the percent identity, then a percent similarity with a known sequence then I can get an idea what kind of microbe what kind of geniuses this is so, this is very important.

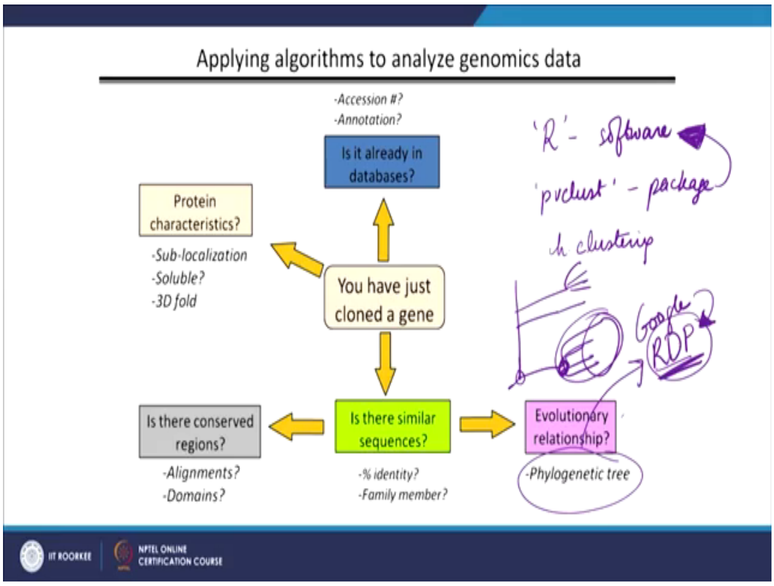
And the next step is I want to find out is are there conserved regions in it if you remember from 16 yard on a analysis one of the beauty of 16 S RRNA is that it has conserved zones.

So, let us say this is the hyper variable region or just a variable region. So, what we can do is we can align all 16 S here, because the conserved region should be same in most of them or all of them very similar. So, this is very similar. So, this stays serve us for aligning sequences with each other and with database and then we can see this one belongs to clostridia this one belongs to negativity cuties this one belongs to something else. So, this kind of differences we can then analyse.

But we need to have something to staple them together and say this is my landmark form which I am comparing. So, that that is why alignment is important domain analysis is important and those you want to understand evolutionary relationship, all right you know you know different kinds of a bacteria were detected in your sample what is the evolutionary relationship how close are they in relation to each other? So, to do that kind of analysis you will you can make trees.

So, for that you can do cluster dendrograms it is called the one that is used most commonly it is called hierarchical clustering.

(Refer Slide Time: 14:17)



Applying algorithms to analyze genomics data

-Accession #?
-Annotation?

Is it already in databases?

Protein characteristics?

-Sub-localization
-Soluble?
-3D fold

You have just cloned a gene

Is there conserved regions?

-Alignments?
-Domains?

Is there similar sequences?

-% identity?
-Family member?

Evolutionary relationship?

-Phylogenetic tree

'R' - software
'pvclust' - package
h clustering
Google
RDP

IIT ROORKEE | NPTEL ONLINE CERTIFICATION COURSE

I encourage you to install a software called ‘R’ it is a statistical software and I encourage you to download a package and install it called pv clust in this software. So, once you get the r software you all you need to do is you need to download the package pv clust. And once you have this package you can very easily do hierarchical clustering of your samples and they will get they will be aligned together in basis of their similarity.

Now, if you are interested in phylogenetic tree then you have other options with you like for example, you can use you can use the ribosomal database project this is an online.

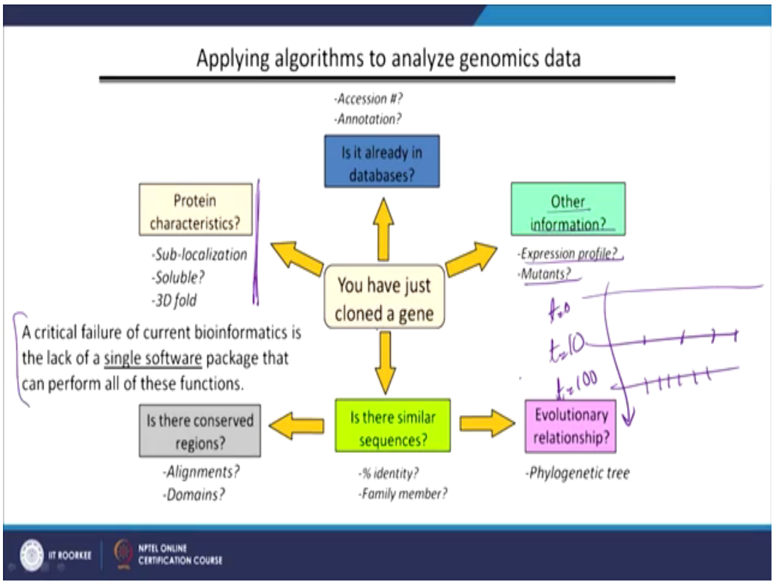
So, if you can just Google RDP you will get information on how to create phylogenetic tree. So, once your phylogenetic tree it would look like this typically it will have dendrograms and then, whenever you see different microbes coming from similar stem we can say this is the common ancestor of all these, and then if this is the initial stem and this is the common ancestor of all these. So, this is evolutionary relationship.

And, then we have protein characterizations what you can do is you have the gene let us say you have your cell allow the competent cloned in which you have cloned that competent bacteria to express the protein, you can actually you can actually analyse the protein or if you have not expressed the protein with the genetic sequence, you can do modelling to find out what how will this what will be the confirmation of this protein what will be it is putative activities and you can get a very good information on protein characteristics.

Next is you can also get other kind of information you can find out expression profile you can also look at mutants. For example, there was a study in 20 reported in 20 14 20 13 I think 20 14 or 20 13 in nature, where what they did was they fed monkeys that had tuberculosis anti tuberculosis drugs and some of the monkeys had tuberculosis I think some of them did not and then, they noticed over time how did the tuberculosis actually all them have tuberculosis here how did you micro bacteria that causes tuberculosis to these monkeys.

How did that change over time? So, this kind of analysis can also be done using Meta genomics or using genomic sequencing techniques.

(Refer Slide Time: 16:46)



Applying algorithms to analyze genomics data

-Accession #?
-Annotation?

Is it already in databases?

Protein characteristics?

-Sub-localization
-Soluble?
-3D fold

You have just cloned a gene

Other information?

-Expression profile?
-Mutants?

A critical failure of current bioinformatics is the lack of a single software package that can perform all of these functions.

Is there conserved regions?

-Alignments?
-Domains?

Is there similar sequences?

-% identity?
-Family member?

Evolutionary relationship?

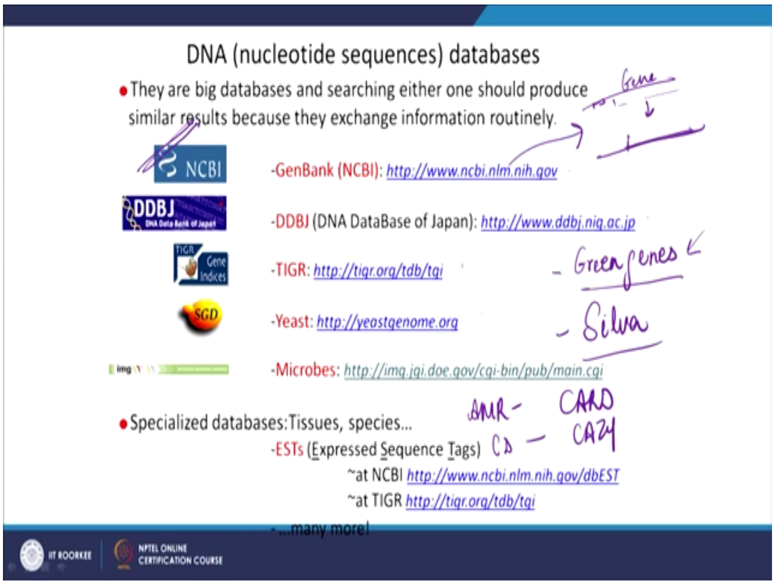
-Phylogenetic tree

So, basically you will have at time t equal to 0 you will have a particular sequence of micro bacteria at time t equal to 10 days you will notice that there are some changes here, some single neutral nucleotide polymorphism some shifts or deletes or additions and then time t equal to hundred days you will notice more more shift more changes.

And, then this way you can get an idea of the how the what is the evolutionary relationship you can also get an idea of how are the sequences changing with time , but so, all these tools are available for us all in online you can look them up if you know that term what you are trying to do, if you are clear about and you can look them up and solve, but remember that a critical failure for current by informatics tools is that they do not we do not have a single software package that can do all of this.

So, there is no standalone single software package that you can install in your computer and get do all these analyses. So, there is something that there is room for improvement. So, if you are interested in coding for interested in computers; I highly encourage you to check out what is happening in biometrics, what are their limitations currently and how we are trying to improve them because there is lot of scope of research of creativity and even entrepreneurship in by informatics.

(Refer Slide Time: 17:59)



DNA (nucleotide sequences) databases

- They are big databases and searching either one should produce similar results because they exchange information routinely.

-GenBank (NCBI): http://www.ncbi.nlm.nih.gov

-DDBJ (DNA DataBase of Japan): http://www.ddbj.nig.ac.jp

-TIGR: http://tigr.org/tdb/tgi

-Yeast: http://yeastgenome.org

-Microbes: http://img.jgi.doe.gov/cgi-bin/pub/main.cgi

- Specialized databases:Tissues, species...

-ESTs (Expressed Sequence Tags)

~at NCBI http://www.ncbi.nlm.nih.gov/dbEST

~at TIGR http://tigr.org/tdb/tgi

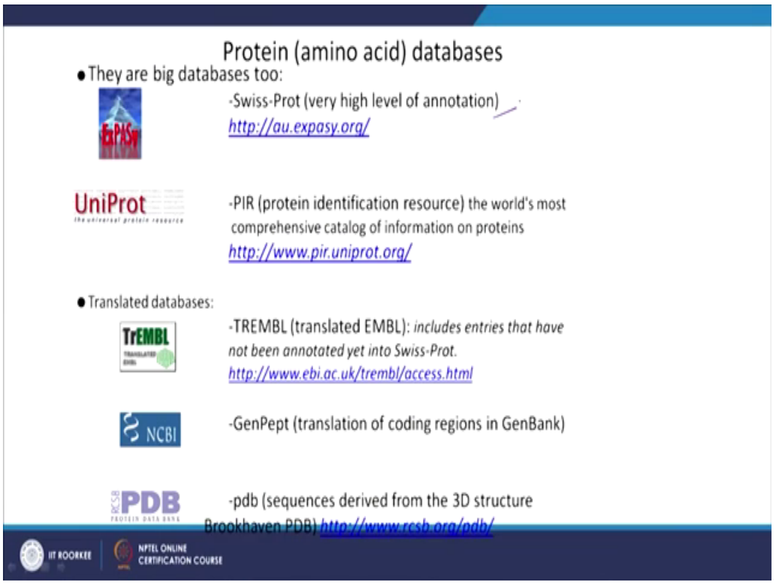
-...many more!

Some of the information that solve the tools that are very popularly used for DNA data DNA databases, that are popularly used by scientists researchers students across the world, to compare the data where to align it to annotate it or Genbank NCBI then DNA database of Japan GDBG, TIGR, yeast for yeast genome and then we have a microbes and then there are others that I would like to say there is the green genes and then there is silver, I encourage you to go through them to the advantage with green genes is it is highly accurate.

What means what it means highly accurate word is that the quality of the sequences is checked over and over again? So, if you get a good match you can be very sure that it is that match and there isn't very less transfers of error. For specialized databases we have ESTs which is for expressed sequence tags and many many more for antimicrobial resistance, we have card as I mentioned for cellulose degradation.

We have cazy and so and so forth all you need to do is Google and you will find out the database that will fit your requirement. For protein the analysis also we have a lot of databases.

(Refer Slide Time: 19:14)



Protein (amino acid) databases

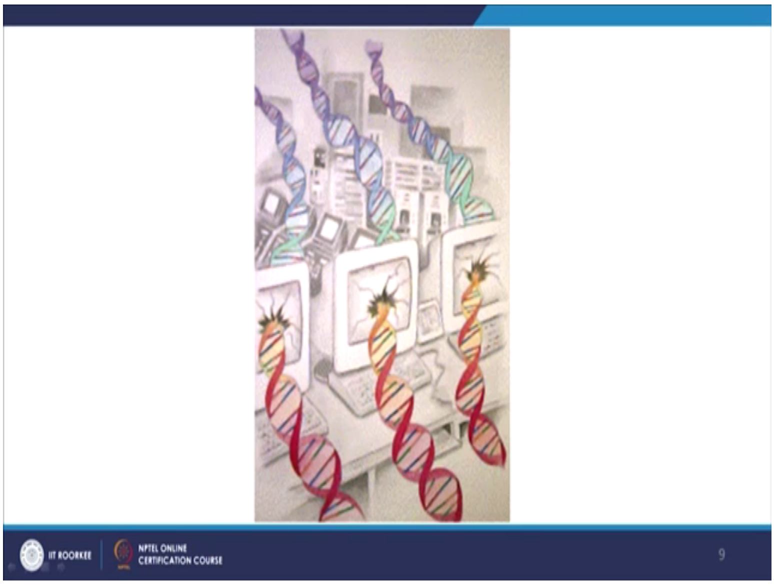
- They are big databases too:
  - -Swiss-Prot (very high level of annotation) http://au.expasy.org/
  - -PIR (protein identification resource) the world's most comprehensive catalog of information on proteins http://www.pir.uniprot.org/
- Translated databases:
  - -TREMBL (translated EMBL): *includes entries that have not been annotated yet into Swiss-Prot.* http://www.ebi.ac.uk/trembl/access.html
  - -GenPept (translation of coding regions in GenBank)
  - -pdb (sequences derived from the 3D structure Brookhaven PDB) http://www.rcsb.org/pdb/

We have Swiss Prot, which happens to have very high level of annotation we have PIR protein identification resort by uni prot it is the most comprehensive database in the world on proteins then here tremble, which is a which means embl by the way and it has it is it has it is like the prevision of Swiss prot.

So, basically the sequences that have not been included in this data we are present here, but again we have NCBI now this is very interesting what NCBI does if you look here NCBI is on top when it comes to DNA databases. So, it has one very nice technique where it will convert it will translate your nucleotide your genes genetic data into proteins. So, if you know where the open reading frame is you know where the start codon is you can start translating it very easily by just looking into the table. So, what you have here is ATGC you will have here as here proteins now.

So, the protein version of your Genbank is called as gen wrapped.

(Refer Slide Time: 20:08)



.

And then we have protein databank again very very important already. So, dear students in summary your Meta genomics will help you take your sequencing data, you run it through the computer. So, that it does not break your computer, but make sense to you and you can get meaningful data to answer environmental challenges to answer the questions that you have been asking.

So, that is all for today in the next lecture we will talk more about bioinformatics.

Thank you.