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

Lecture - 56
Bioinformatics I

Dear students, welcome to our lecture on bioinformatics. So, in this part of our course applied environmental microbiology, we will see how we can use tools from computer science and mathematics to answer and make sense of the data that we generate using biotechnological and microbiological essays. In one of the previous lectures, I have talked about sequencing techniques all the first generation, the second generation, the third generation and the fourth generation; from the third generation and four generations sequencing techniques, we generate tremendous amount of data and that tremendous amount of data is not something we can analyse using our excel sheet.

And while we are at it, I must also mention even from second generation techniques, we generate enough data that we require sophisticated tools that can help us make sense out of the data. So, it is not just the sheer amount of data that we generate, for example, if you do meta genomics, we might generate anywhere from 60 gigabits of data to 120-200 gigabits which implies millions of base pairs per sample would information would be provided to you.

So, that is lot of data. So, it is not and in addition to it, being big data, we have another other hurdles too, for example, in our sequencing techniques regardless of which technique we use, there will be sequencing errors that will come into picture there were the other chimeras that would be formed and sequenced false leading to false positives and confusion also with the data said, you need to align it to the existing data bases and all this cannot be done manually.

We need algorithms and programs to do that for that reason biology has in has collaborated with computer science and mathematics to come up with bioinformatics and that is what we have talked about briefly today.

(Refer Slide Time: 02:18)

So, bioinformatics is the word suggests is bio informatics. So, we are trying to make get information from the data collected from biological studies now as environmental engineers; what would our environmental scientists environmental students or students of this course applied environmental biology for us typically the most important big data generating technique that we use is sequencing.

So, we can refer to it as meta genomics, if you remember about meta genomics one of the key feature of meta genomics is their is ability to sequence genes to sequence genome and extra chromosomal gene material to a generative material to from microbes it cannot be cultured.

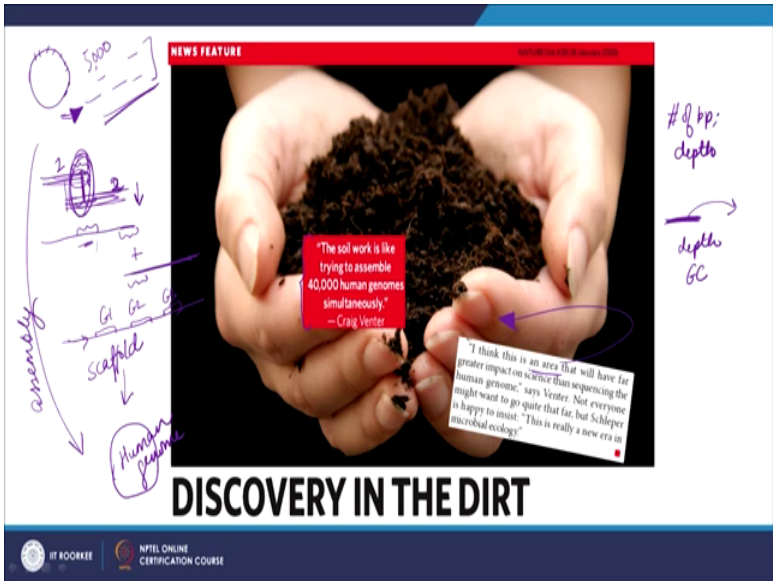
So, without culturing them; so, we skip the culturing method culturing step and we directly sequence or sequence them after amplifying the genetic material. Now, in meta genomics; what will happen is we will get a good snapshot of who is present in our sample and perhaps even what they are doing.

So, we can do functional meta genomics; that will give us small clear information on the functionality of the microbial community, either case meta genomics is very useful for environmental engineers environmental scientists and researchers and meta genomics as name suggests generates tremendous amount of data which we cannot make sense of without bioinformatics, all right. So, let us now look at meta genomics here and briefly, I have talked about this earlier in one of the lectures of this course that when sequencing

techniques were developed we finally, managed to sequence the entire human genome which was a big achievement.

It was not done single-handedly by one research lab, but was a collaboration between many P highs in many researchers. Now as such you should get an idea about how difficult it is for us to sequence.

(Refer Slide Time: 04:11)



And make sense out of environmental samples using this code by Craig Venter the soil work is like trying to assemble forty thousand human genomes simultaneously. So, and if you have not already done; so, go ahead and do it look up human genome sequencing online, find out about the scientists, who were involved in it and what kind of supercomputing tools were required; how much time did it take and what you will find out that it required lot of computational facility.

Because human genome is a eukaryotic genome it has thought of Introns, Exons and other genetic materials around not present and prokaryotes like bacteria which usually tend to interest us more from environmental perspective. So, when you are gene sequencing human genome not only we need to have sequencing part done.

But then we need to get rid of the errors or suspected errors we also need to have an idea of the quality of the sequences you are generating we need to get rid of chimeras and after we have done all the qa qc work next step is to assemble them. So, let us look at

briefly about what assembly is. So, when we sequence and this is shotgun sequencing you have let us say a bacterial genomes it is circular and because we are doing shotgun sequencing, what we do is we split it into small small pieces and we get small small sequences fragment. Now one singular genome can be split into let us say 5,000; 5,000 pieces 5,000 fragments. Now each of the fragments can will be sequenced by your sequencer and once you have the information, the small sequence a small length that the small fragment of your genetic material may not be as informative as the entire genetic material.

Because see many things are happening in our genetic material whether it is plasmid or chromosomal genetic material the sequence of genes is important the location of genes is important also; one if sometimes when we do shotgun sequencing at not sometimes many many a times; when we do shotgun sequencing, we do not get the entire gene in one sequence they get split up. So, we need to assemble.

So, that we remove all confusion we get click gain clarity and we know the sequence of genes in their genetic material. So, for that; it is very important to stitch back these fragments. So, now, you have generated, you have generated this data and it is important to stitch them back. So, stretching them back would look like this let us say, I have a sequence here and here it has a particular sequence that I find pretty unique and then I find another sequence that are similar or exactly same sequence on an in the right orientation.

So, what I can do is I can assume that because of this, similarity, there is a good probability that they are actually in continuation with each other, but there was a break here for one bacteria or one genome and for another there was a break here during short gun process and thus, we have this split. So, what and what we can do is we can stitch them together and now we have there is a longer fragment.

So, what we can do is we can take our sequences and assemble them and assemble these longer fragments further; for example, they have unique signature here there might be another unique signature here and if we find another sequence or another content that has same sequence here.

So, we can stitch them together and this way we keep growing longer and longer sequences until we reach scaffold level. So, we have scaffolds really long up to thousand

base pair and more ah. So, these are more meaningful because they typically would have genes sequences; you will be able to know gene one appears before gene 2 appears before gene 3 and it is in the right according to the orientation or in other; in other in another sequence.

So, and then the scaffolds can be further assemble together to create recreate the entire genome. So, this is what we call about genome sequencing the genome the bare sequencing is the first step of recreating the entire genome and from re getting the information from the entire genome. So, all right I have a fourth generation third generation sequencing machine and it gives me this data.

But I still need to go through all this step and these steps are called assembly this is assembly process basically stitching the smaller sequences together. So, we have a bigger picture a broader picture and I really want to go for whole genome sequence. So, we get the sequence for whole genome, we can assemble it and we can stitch them together and now we can say that said this is bacteria one bacteria one has been sequenced or if this was a human gene human cell. So, we can say the human genome has been sequenced, right.

So, this was this is computationally very challenging. So, we have algorithms which decide similarity we stitch together which check the error rate because every time you are stitching together sequences we are losing our certainty certainty about the sequence sequences. So, we are adding to error. So, for that we have algorithms which decide whether it is worth adding stitching two sequences together or not another thing, I want to say that here in this example that I gave you notice that I have only stitched one sequence with the other sequence typically in meta genomics; what you will have is you will create multiple copies of same fragment.

So, let us say; this is fragment 1 from 1 bacteria; this fragment 2, 2 from second bacteria they have some similarity, but they are not be just one sequence or one fragment that would be like this because sequencing sequencing techniques often involve amplification step will have multiple copies of first bacteria's fragment 1.

So, multiple copies and the more copies we have that agree with each other I did yes all these copies are very similar. So, we can say that they are copies of the same fragment they are same size they have same similar sequences and if all of them if two of them

have same match. So, sick fragment one and fragment two both have same match no matter how many sequences we have that we can could label as fragment one and fragment 2, then we can with confidence stitch them together. So, this is the importance of the word depth.

So, number. So, when you have meta and that brings to me me to another point that when we are doing meta genomics or sequencing meta genomics basically you need to know number of base pairs that you will generate this will give you an idea of the width how much data are you going to create; obviously, the more is better next you want to know about the depth which is basically one single fragment let us say in my sample I have one bacteria that is unique and there's only one of its kind.

And I split it into fragments in the short gun process. So, this one single fragment how many times will it be sequenced or how many copies of this sequa of the sequence will I get this is very important the depth and depth is very important because when we are assembling the more the depth we are the most sure we can be about our assembly process note the depth will depend on many thing depends on the platform that you are using you know some platforms give you good depth.

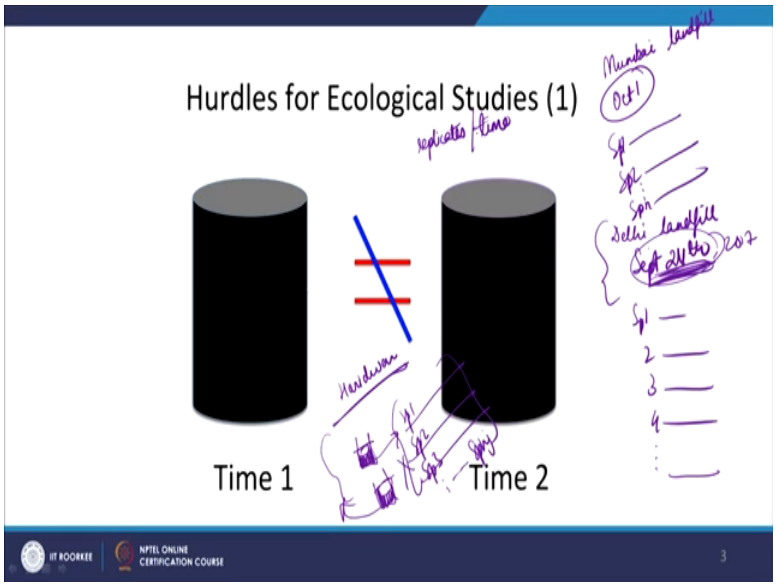
Some do not it depends on the diversity for a lot of microbes that the depth would not be as much or if you have pure culture then definitely you have very good depth the other thing it depends upon this GC content because as GC content changes the annealing this is this annealing step that in the sequencing process gets affected. So, there multiple factors that will determine depth, but depth is very important all right. Now let us come back to the code by Craig Venter here, he is mentioning the soil work is like trying to assemble forty thousand human genomes simultaneously.

He is trying to explain to you this was very difficult getting a human genome assemble together stitched together, we had to use high performance computing in USA perhaps the best in the country there and one of the best in the world definitely. Now, imagine 40 times 40,000 times over that is the challenge for soil micro biome and I hope this also gives you an appreciation of the kind of complex microbial communities we are dealing with here.

So, one human cell no way near as complex no way near as diverse as soil microbial community and let us continue listening to enter, he says I think this is an area and he is

talking about meta genomics and sequencing and a including bioinformatics that will have far greater impact on science and sequencing the human genome which area sequencing the soil micro biome say is mentor not everyone might want to go quite that far, but Schleper happy to insist this is really a new era in microbial ecology already.

(Refer Slide Time: 13:11)



Now, let us look at the hurdles for ecological studies and how meta genomics is a promising feature and why you would perhaps want to use it and why that will push you towards bioinformatics. So, in ecological studies and I hope you have learned this that microbial communities are not static; for example, if I have a wastewater treatment plant, I do its characterization and it will do well in my wastewater treatment plant I have these protozoa have these bacteria I have these fungi I have these bacteria fugues this is how the microbial community looks like.

And yeah I know how to how to behave because I know the community structure I know the functionality, but hopefully by this time of the in this course you know that microbial communities are not static; they are constantly evolving. So, they are constantly undergoing succession for many reasons a one is the random error that crops and then the random deletion of a population or addition of a population.

Next is that influx and efflux or different kinds of samples and in ocular that changed my rural community and yet the most important is that the environment of the sample changes over time for example, we have talked about lakes if you remember. So, at the

bottom of the lake no no photo trophy definitely because a lake does not in that level lake does not receive sunlight and most likely if it is a deep lake it is anoxic or anaerobic and in this case aerobic microbes cannot thrive.

So, they would we expect aerobic heterotrophs where would we expect auto photographs at the top of the lake now as oxygen gets depleted from the bottom because there are oxygen consuming the reactions happening and which is basically oxidation of food oxidation of electron donors and when that happens oxygen gets depleted and we start getting rid of a redox gradient at the lake depending on the stratification of the lake we will get a redox gradient.

Now, that redox gradient also does not remain static it to changes it to changes, but time as oxygen keeps on getting depleted or if re addition is well enough and it starts entering into the lake and then there are seasonal patterns. So, the take home message here is that ecology is not static biology is not static definitely environmental microbiology is not static and a very clear cut example is that your for 15 years your wastewater treatment plant work perfectly no problems microbes are happy you are happy with the microbes and the work that they are doing and then one day it is lodge refuses to float why because microbial community has undergone a succession. So, this is a major hurt this is a great point for appreciating environmental studies and yet it is a major hurdle; why because let us say I took a sample from Delhi landfill on September 24th already.

Let us say I took this sample from Delhi landfill and then I characterized some microbial community, I said species one, these presence species 2, these present 3, this present four this present so on and so forth and I did all kind of analysis alpha diversity I compared and then I this was in September 24th, 2017 and then I went to let us go to Mumbai, I went to Mumbai and I collected sample from Mumbai landfill on October the first and then; obviously, I got information species one, so much species 2, so much species n so much.

Right now, the question is can you compare Delhi landfill with Mumbai landfill; for example, let us say on to publish a paper comparing micro communities from different landfill different demography different region different geography different climate and then I do this I have sequences for microbial community in Delhi landfill I have sequences for microbial community in Mumbai landfill.

And when I compare the question is how much similarity are you expecting how many differences are you expecting are the differences only because this is deadly in that is Mumbai or other differences according to the age according to the time this was September 24h that was October 1, right. So, maybe there was a rainfall event or maybe geography climate.

So, there are many variables that in are involved and which in changed the composition of my community which make microbial communities undergo succession the other aspect of this time one is not equal to time two is for example, you go to hardwar a sample the wastewater treatment brand and you are like already hardwar waste water remove plant has this much problem has this much emerging contaminants and has this much of do this much of dod. So, if you get the information for the water form from this hardwar wastewater remove plant.

Now, you go again next day and not only that not only this basic information you also do microbial community analyses you know species 1, species 2, species 3 so on and so forth, right. Now, if you go next again to hardware wastewater remove plant and correct another sample, then it is very likely that the microbial community will not resemble this will not resemble this. So, this is a very important point that the time of the sampling makes a big difference in the microbial community I am hoping to record.

So, this is a major hurdle for ecological studies and one simple way out of this is to do replicates or to sample everything at the same time and to have identical conditions which is very challenging when we are trying to do in field related work the next hurdles for ecology studies is 99 percent of bacteria are non cultivable.

(Refer Slide Time: 19:00)

So, basically the kind of bacteria that you can culture in the lab like tuberculosis and you can culture so many other serum staphylococcus aurous all these can be cultured. So, we know a lot about them thankfully, but 99 percent of bacteria uncultured we have absolutely no idea who they are or what they are and I if nothing is decide I hope definitely serves the purpose of humbling us into realizing that very little less known.

So, now, let us take a recap and find out; why 99 of microbes are unconceivable; obviously, the simple reason is because the environmental conditions are very different here . So, basically this is the environmental conditions in the lab are different; for example, let us say there is a scientist a very renowned German scientist who goes to hot water springs and lakes and vents thermal vents to collect samples. So, that he can find out the bacteria.

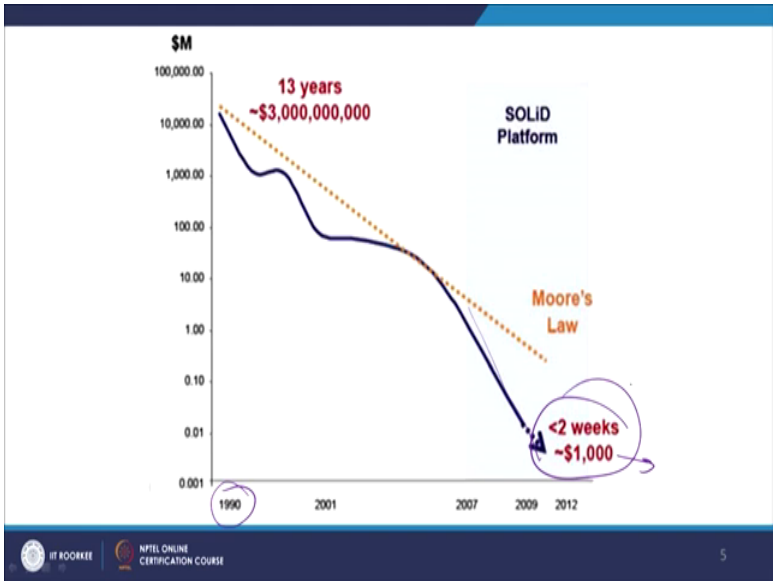
Now, let us say he gives me his hot water sample in which there are Archaea and there are extreme extremophiles growing and I take them to my lab and I try to grow them on the broth now here is the thing my broth is at 25 degree Celsius, the bacteria they were living at 98 85 definitely above 70 degree Celsius, how will they survive in the play there; it is not possible.

So, the environmental conditions when they change, they can because the environmental conditions in petri dishes are very different from environmental condition where these microbes actually grows that is why 99 percent are uncultured cultivable the other reason

could be that they like to exist in microbial communities, this is another reason because microbes they prefer to exist as communities, they do not like to grow individually in the lab in only few of them can there is only few of them are comfortable and then there are many other reasons that we have not completely understood yet so, because we cannot compare two time intervals of ecological samples..

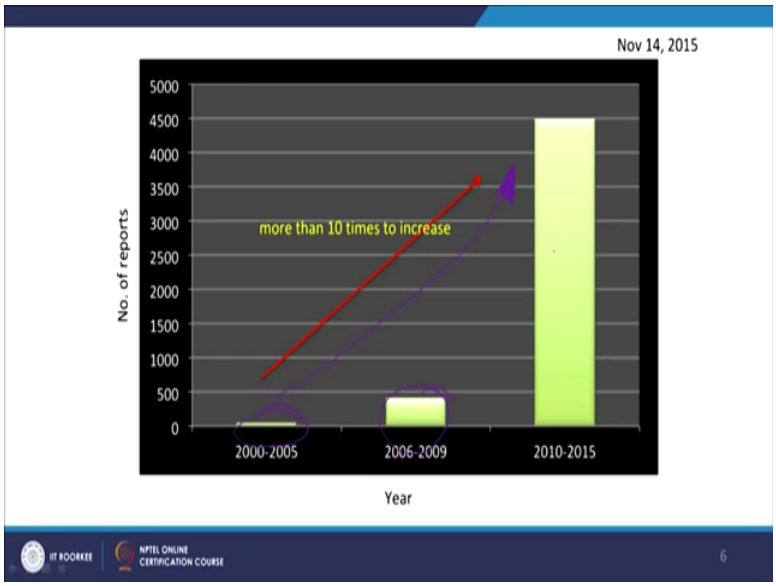
So, we need to do replicates and we need to rule out all other factors we need to collect a lot of sample generate lot of data and because 99 percent of bacteria are uncultivated; we need to find non uncultured independent sequencing techniques and; obviously, meta genomics shines as the most popular culture independent sequencing technique now let us look at the cost of sequencing.

(Refer Slide Time: 21:51)



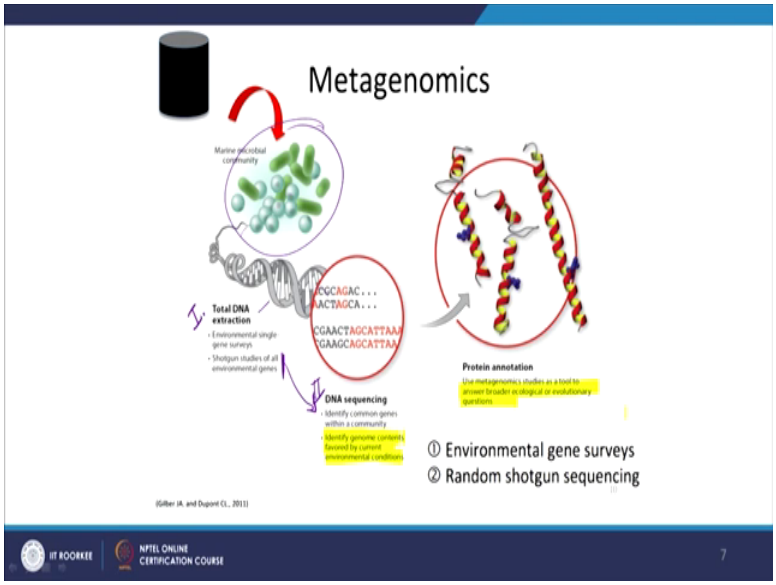
In 1990 sequencing was very very expensive and following Moore's law and even better the cost has reduced a lot and it is even less..

(Refer Slide Time: 22:05)



Now already now, as the price is reducing as a need is being more and more appreciated; obviously, the trend for number of publications in leading databases is like this. So, in 2000; 2005; they were very little papers on soil meta genomics in 2006-2009, we had considerable number in 2010-2014 we just shot up. So, there is a lot of scientific interest and microbial communities interest.

(Refer Slide Time: 22:30)



Now, let us look at meta genomics again briefly as a recap. So, what do you do in metra genomics you take your sample this is your let us say marine microbial communities it

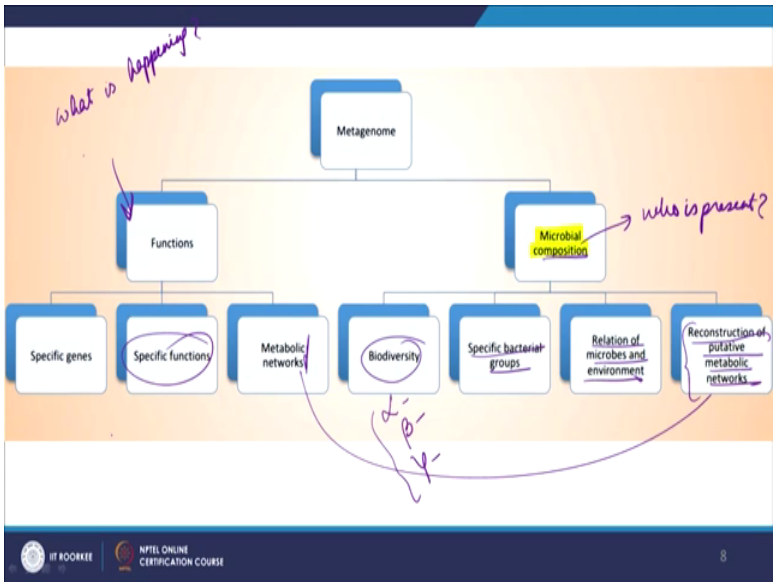
grows in ocean and if they have some sequence of their genetic material. So, the first step is you extract DNA. So, this step one total DNA extraction and and after you have done the extraction you might go for shotgun studies for all environmental genes and after doing shotgun you go for the second step which is DNA sequencing. So, in this case you do the DNA sequencing when you get the data you identify the common genes within a community.

For example I noticed that I put very high ph in some place and still there are microbes growing, but I put high ph. So, they do not grow, but they are growing and I want to find out what makes them grow at such high ph. So, what I do is I can sequence the DNA find out for the similarities what are the similarities between all microbes across the globe that grew at high ph once I get the similarity I can be more sure that already this might be something new some new microbe and I need to pay attention and after you have got the sequences; what you do is you common you identify common genes within microbial community.

So, once you get common genes it makes analysis very very easy and then you identify genome contains favoured by a current environmental condition; this is a very very important sentence next is protein annotation now you have sequences you are pretty sure where they are coming from you know how to deal with it next step is very difficult squat put in annotation again you use meta genomic studies as a tool to answer broader ecological.

So, you use whatever information you have whatever tools you have in your hand to answer broader in questions on ecology and if here it says evolutionary, but just environmental microbiology already.

(Refer Slide Time: 24:14)



So, let us look at meta genomics; what manage enemies look like meta genomics we can do 4 functions. So, basically kind of functional meta genomics or we can do microbial composition which is basically telling me who is present and function is what is happening .

So, if you are doing micro biome composition as say which are which are the most typical as say nowadays you want to find out biodiversity. So, you want to look at alpha diversity and look at beta diversity potential even gamma diversity these are not the gamma bacteria alpha beta and delta epsilon these are very simple sign and all that, but this is very simple. So, what you can do is you can look at biodiversity; how are the sequences in this sample different from the other sample.

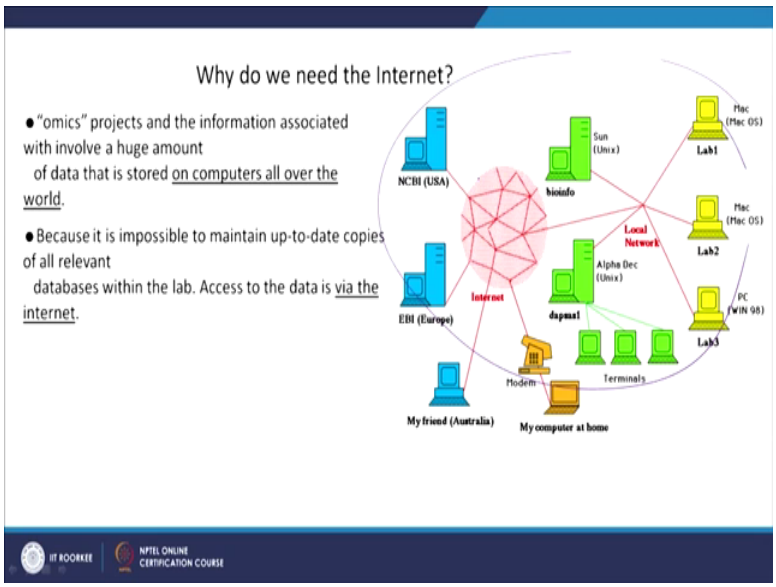
So, for example, samples from Mumbai landfill different from Delhi landfill; this analysis can be done here when we are doing better dive biodiversity we can also do n sample diversity is all microbes in Mumbai landfill more diverse than the ones in the Delhi landfill or is it the other way around. So, all this will come under biodiversity we can also looks for specific bacterial groups already I am doing solicited addition tell me what kind of cost you here present I am looking at only cross trade here and I can look for them.

Next we can find a relation of microbes in environment. So, whenever I take sample from an environment the microbes that are thriving in that environment are the ones that

are very well suited for that environment. So, what I can do is because I know the environment because I picked the sample from the I know its temperature I know its humidity I know its conditions I can pair up in this kind of condition these microbes grow well this information is very very helpful for commercial purposes and also for research.

Next what I can do is using my microwave composition I can look at reconstruction of putated metabolic networks now if I am looking at function with function I can understand the metabolic networks pretty much similar like this I can look at specific functions, for example, if I am doing cell deterioration, I might look at the function let us talk about degrading starches, I will talk about deviating cellulose and I can also target specific genes already this is perhaps going to be the last second last slide here the question is already we need to do meta genomics.

(Refer Slide Time: 26:26)



We need to do sequencing because these are the conditions that we face as environmental engineers we cannot culture 99 percent of bacteria. So, there is no point cutting to culture anything; it is let us do sequencing.

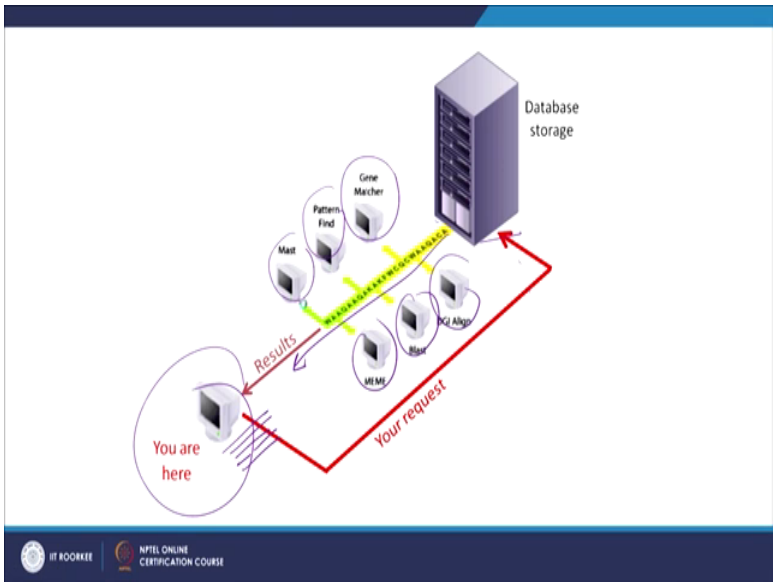
Now, when we do sequencing we need internet why do we need internet well because more often than not our own PCs are not sufficient they are not built for and they are not sufficient to do bio informatics analysis. So, we need to have our institute computer that you can use or in case if you have a good computer at home you can definitely use that.

And once you have used that computer you upload the data to NCBI and CBI verifies, it might send it to GENBANK and then you get published and stuff why is this important look, here the omics project like meta genomics, you know like proteomics metabolomics all these are my projects and we have talked about comics before it means all these omics projects and the information associated with them involve a huge amount of data that is stored on computers all over the world.

So, if you are connected to internet here and we are doing meta genomic analysis our data will be shared by other computers. So, it will not be very taxing on our own personal computer it is impossible to maintain up to date copies of all relevant databases within the lab access to the data is via the internet it is not possible that we have up to date copies of everything every database present in the world within the lab. So, people should be able to access it online for example, recently my colleague from Virginia tech corps informed me that they need access to the Linux computer that I was using when I was doing my PhD.

Why because well we sit at home and we do the work. So, for that reason we do need to connect to internet.

(Refer Slide Time: 28:14)



And then this is the last slide let us say you are here connected to internet this is database storage and then what happens you send a request for example, let us say you have a file in one of the lectures, we will be doing this and you send please align this file with the

database tell me, what it matches best with and what this will do is it will go for gene matcher it will go for alignment for blast for M E M E for mass for parent finder and after it has done all that analysis will get back to you with information .

(Refer Slide Time: 28:41)

Already, there last slide for today the current by informatics market this is in USA by the way is worth 300 million per year prediction is it will rise up to 2 billion per year in 5-6 years, there are at least fifty bioinformatics companies at time of making this presentation in USA. So, there is lot of scope even in India, we have talked about it scope in public health scope in general health in well being.

Also, lot of people who are not able to exercise, but they need to know how their body is changing with it and then they generate a lot of data, then they need to make sense out of this data and to make sense out of the data that they are generating using variable sensors because of their health issues that data analysis also is turned by bioinformatics.

Alright, students this is all for today. In next class, we will go more in depth about bioinformatics try trying to understand what tools are available to us easily and how we can use them best.

Thank you.