

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

Lecture - 58
Bioinformatics III

Hello student, welcome to the next lecture on bioinformatics. In the previous lecture we learnt about, how we are utilising the latest techniques in biology, latest techniques in other fields like; electronic to gather as much data as we can about our environment on microbial communities and other kind of microbes.

Now, the data that we generate can be up to the order of 60 gigabytes and more, and this is too much data for us to use conventional techniques such as the plane old excel sheet. So, in order to make sense out of our immense data that we are generating now, using next generation sequencing techniques and other techniques, we need to bring in the expertise from different fields and together we called this synergy between different fields as bioinformatics.

(Refer Slide Time: 01:09)

What is Bioinformatics?

Genetic seq.
Metabolomics
metaproteomics

Biological + IT, DA → sense

bash
python

Conceptualizing biology in terms of molecules and then applying "informatics" techniques from math, computer science, and statistics to understand and organize the information associated with these molecules on a large scale

IT ROORKEE | NPTEL ONLINE CERTIFICATION COURSE

So, let us explore bioinformatics. As it as a name very; obviously, suggest bioinformatics consists of two different words bio and informatics. Here bio hints at the biological origin of our sample. And the informatics here constitutes of information technology and data analytics, which together help us make sense out of our data. Now when we come to

the bio part here we can have genetic sequences, or maybe we can go for metabolomics or meta proteomics or any of the other advanced techniques that help us gathered tremendous amount of data for our biological sample.

Now as I mentioned that this amount without; when we do this technique so kind of information we receive are simple text file is usually now with this text file we are not really getting information about, what is going on in environmental microbial communities for example, a long text file that is very hard to open on a simple laptop that has just 80 gcd repeated all over again in different combinations and permutations does not make sense to an average environmental engineer and scientist.

So, in order to make sense out of it in order to be able to answer questions that we have asked in our research, we should be able to make; we should be able to use the technique given by informatics. So, let us and this together this exercise together is called as bioinformatics. Now, when will look at bioinformatics; let us look at the definition of issue definition of the bioinformatics it says here conceptualizing biology in terms of molecules and then applying “informatics” techniques from maths, computer science, and statistics to understand and organised information associated with these molecules on a large scale is called bioinformatics.

So, let us breakdown this really long definition into small different phrases and try to understand, what it saying. The first the very first phrase we opt to notice is conceptualizing biology. So, we have biological data that we have generated using chemical reaction, using electronics or using optical measurements of just measuring the ph changes as sequencing proceed. We have gathered some data, now we need to understand; what phenomena; what environmental phenomena, biological phenomena is being represented by this data. This part is conceptualizing biology this is going back to our original research question.

And we want to understand it in terms of molecules on this could be a genetic material and it could be your it could be your metabolites could your proteins. So, it depends whether you are doing meta proteomics, transcriptomics or genomics or whether we are going for metabolomics. So, once we had this data on our say molecules, we can apply informatics techniques from maths, computer science and statistics.

So, let us say I have lot of data that have generated from metabolomics or metagenomics. Now the next step comes to do quality control qc to make sure that the data that has more error or the data that I cannot be sure of is removed. Now for this we required certain algorithm said we have for mathematics. For example, we have we have algorithms that can actually predict, what are the next nucleotide should be in a sequence; and if it is not that nucleotides, which it as predicted there is the chances of it being a chimera which is basically false constructed sequence is higher and in that case the sequence can be removed.

So, in this we use the tools from mathematics that they have the mathematician who have development amazing algorithms to predict the genetic sequences, and how do we do that; we do it by using computer high performance computing more often than not and where is where we required some programming skills typically. So, bioinformaticians are pretty good at basic bash they are pretty good at python and just generally any it s, happened that if you are good at one programming languages others are very easy to catch up so, just general programming.

So, once you know once you can feed the program in your computer your computer can applied algorithms, predict basic sequences are more likely to chimera which are more likely to be true representative of the sequence, that is present in the environment you can eliminates the chimeras, then you can keep the good ones you can do other kind of quality control tests too for example, this repetitiveness we can remove that remove that also.

So, now, I have cleaned my data, I have even used my; programs to align the data with existing database and I sort of I given it annotations, what means what is meant by annotation here is that the sequence now has been cleaned it has been given a name. So, I can call it let us say this is cell 48 sequence. So, cell 48 is specific to cellulose degradation. So, I can say that this sequence represents the ability to do cellulose degradation; particularly the particular step rate limiting step of cellulose degradation.

So, once I have done this data this analysis once I have done the annotation, the next step to answer my research question is to do sound statistical analysis for it. Now when you talk of statistical analysis this is basically testing the hypothesis we want to find out, what is the probability; that the whatever claim we are making is not a fluke or whatever

observation that we have had is not a fluke is not by chance it is, but it is real reliable and producible phenomena, that we can observe over and over again.

So, this is done by using statistical techniques and tools, and if you are into environmental science, environmental engineering or bio biology bioinformatics I highly recommend you to get in touch with either a good statistician or maybe a good statistics course on NPTEL and understand atleast the basics of statistics, because without this no matter how good your expensive design is, no matter how expensive techniques you have used together good data and even the good informative informatics techniques you have used to from computer science you have used to clean your data and make sense out of it; without a backing from statistics you cannot really have any you cannot make any solid claim.

So, in our environmental science, when we get then we generate data such as genetic sequences, information of metabolites, information on proteins and we apply techniques from mathematics from computer science and even statistics not together this process is referred to as bioinformatics; another word here is large scale. If we already were doing this on small scale for example, if I have my waste water treatment plant scale waste water treatment reactors, and I am working on them so I am basic statistics gathering basic data I am trying to make sense out of it this is not bioinformatics for say, because it requires that this data generation collection is done on a large scale.

So, basically anything that is high through put and you analysing it you are stepping into bioinformatics. Now, another thing I want to mention that bioinformatics also includes apart from sequencing and annotating them bioinformatics also includes modelling. So, let us say I know I have identified the protein its sequence, and now I want to understand what its 3 d confirmation is in that case, what I can do is; I can use my computational techniques here and I can make 3 d models of the protein which may be able to predict to some degree of certainty what the function of the protein is and that I can then test in my lab all right.

(Refer Slide Time: 08:42)

Introduction: What is bioinformatics?

- Tools, algorithms needed to handle large and complex biological information. *within discipline* *multi-d* →
- The NCBI defines bioinformatics as:
"Bioinformatics is the field of science in which biology, computer science, and information technology merge into a single discipline"

IIT ROORKEE NPTEL ONLINE CERTIFICATION COURSE

So, let us again brief as a bioinformatics. So, what is a bioinformatics what kind of tools to be required bioinformatics will give us, the tools and algorithms that we need to handle large and complex biological information this is the definition given by NCBI. NCBI by the way is the largest bank of nucleotides not the nucleotide per say.

But for the largest bank of nucleotides sequences in entire world and it is open the gold standard in environmental microbiology and other and other kinds of microbiology bioinformatics is the field of science in, which biology computer science and it merge into a single discipline, which really bring me a up to very important point that I would like to make here often in our colleges often in our studies; whether we are in undergraduate classes of whether we are doing post graduation our expertise are divided into very narrow, very finely define narrow disciplines for example, this is environmental engineering this is transportation engineering this is chemical engineering right; this is physics, this is chemistry problems, this is biology problem even within biology.

We have divided into different groups to general biology this is microbiology this is molecular biology and so, on and so, forth; however, one thing we need to recognize that nature itself does not follow these narrow definition of biology of disciplines, what does follower is it is own nature. So, the disciplines are manmade and now what science is progressing towards is removing the boundaries and going from within discipline research. So, this is defined as I within discipline.

So, you go in depth within a discipline to trans discipline research, where we actually take experts of different disciplines they come together and they tried to solve a problem this is more like what bioinformatics; we have computer science, we have mathematics, we have biology, and this is and then we have multi disciplinary search and then we go to interdisciplinary research. So, interdisciplinary research we are actually not only our people expert in their own disciplines, but their idea of expertise is fluid and it moves through multiple discipline for example, we might have a biologist whose real expertise lies in how to write programs to make sense of biological data.

So, for example, if you are a computer engineer and a computer engineering student and you are a part of this course this is the an opportunity for you to use your understanding of environmental microbiology and apply it in your computer engineering to develop new tools and do research that would make advancements there and we also moving towards trans disciplinary research, there we do not define a problem as an environmental microbiology problem or as a computer engineering problem is just a problem and we use whatever tools are available without any constrain of what discipline it is.

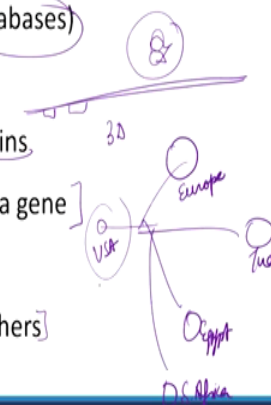
So, to me bioinformatics is really the cutting edge of moving from within discipline research to trans disciplinary search and you would not be surprised to know this that I hope, that bioinformatics is a lot of research is happening across the globe and there is lot of demand for it, because most of the bioinformatics processes require lot of high performance computing and any kind of algorithms can reduce the computational cost would be major demand anywhere right.

Now let us move on to; why do we use bio informatics or what do we used it for we used for storing and retrieving biological information from databases.

(Refer Slide Time: 12:11)

Why do we use Bioinformatics? *NCBI → Genetic info*

- Store/retrieve biological information (*databases*)
- Retrieve/compare gene sequences
- Predict function of unknown *genes/proteins*
- Search for previously known functions of a gene
- Compare data with other researchers
- Compile/distribute data for other researchers



III ROORKEE NPTEL ONLINE CERTIFICATION COURSE

For example, I have generated lot of database a lot of information and I want to make sure, that it is uploaded on a database it is made public, but manually uploading a sequence very painful so, because I have billions of sequences. So, using bioinformatics, I can write commands that will uploaded for me that will give metadata to users who might be interested in the data.

And also let us say there is a database or there is some information available on mg rast or another platform and I want to download it and I want to make sense out of it, then it can also retrieve biological information for me the other way we used bioinformatics in the same sense of retrieving is let somebody in Austria, Hungary or in Austria sequenced a pure culture and found out the importance of a particular gene and that sequences available on our database.

We uploaded data on let us say NCBI and they wrote information about it like, what is the speciality of this gene. So, I can come match I can match the gene that I have sequenced in India with those databases, and then if there is the similarity I will get an I will get information already this gene that has sequence in India matches the gene that was sequences in Austria.

So, very much is really what there is a very high chance that the function of that gene similar to the function of the gene, that I have here and in this way I actually retrieve biological information I can compare my sequence I can get an idea of how unique or

how similar my sequence is and how unique or similar it is function might be. So, the other option that we have is that; if you are not if you if you do not want to compare with gene, but you have a protein sequence what you can do you can translate a protein sequence and you will have at least the sequence of the protein amino acid a follows amino acid be and so, on and so, forth.

Now, in this case we can do three dimensional modelling using bioinformatics to find out, what would be the secondary structure of your amino acid sequence and what will be your tertiary structure of your amino acid sequence, what will be it is three d confirmation and then it can match it and predict what kind of properties this protein will have all right. Next what we can use biometrics for researching for previously known functions of a gene. So, for example, I have a gene and I know it is doing some job that it is doing at source my purpose, but I also want to know what other people have searched before.

So, basically this is a very good literature review for me, I can match this gene with databases or I can translate this gene into protein in silicon, which means I do it on a software do it on a computer I do not have to actually clone it into a bacteria using a vector and in make the bacteria expressive. So, once I have (Refer Time: 14:52) the translated it, I can match with the existing database and find out, what is the experience? What is the; what are the observations of other researchers that is it I can compare data with other researchers.

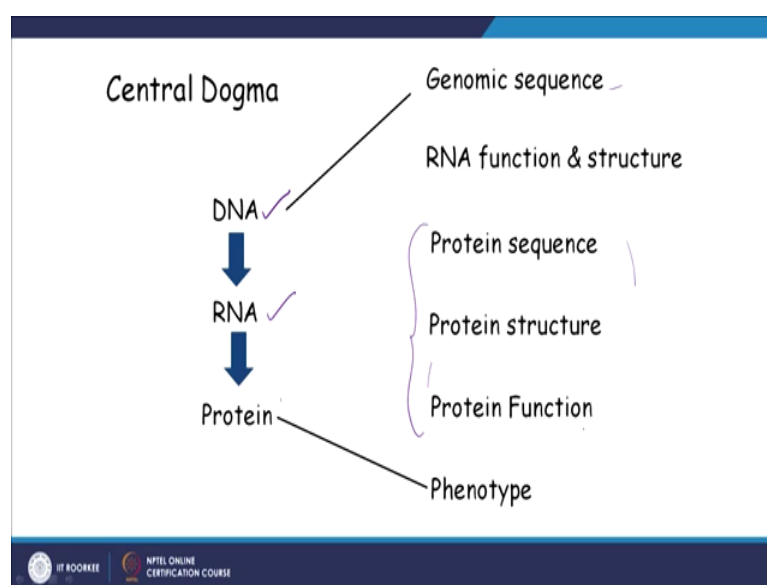
And then I can compile and distribute data for the researchers and this is very fantastic because recently with Meta genomics we have generated so, much data that. So, let us say let me start here again. So, let us say we have a big part we have a big world here. Let us say there were some study done in USA, and the about soil microbiology and they observe some kind of microbial community some kind of interaction and all is well and good.

And let us says some other study big study, that happened in Europe and they observe same soil microbiology did different communities different characteristics and stuff and another study in India and another study in Egypt and then let us say another study in South Africa. So, all these studies have collected data from different parts of the world

the one way possibility; that the bioinformatics gives us is that we can collect data if the data has been uploaded on databases.

We can collect this data from different researchers and then do a kind of Meta analysis to understand, what are the global patterns? And there have been some really good papers the people have published to see the global patterns of antimicrobial resistance which is a major environmental challenge. So, there is comparing and distributing data for the researcher.

(Refer Slide Time: 16:23)

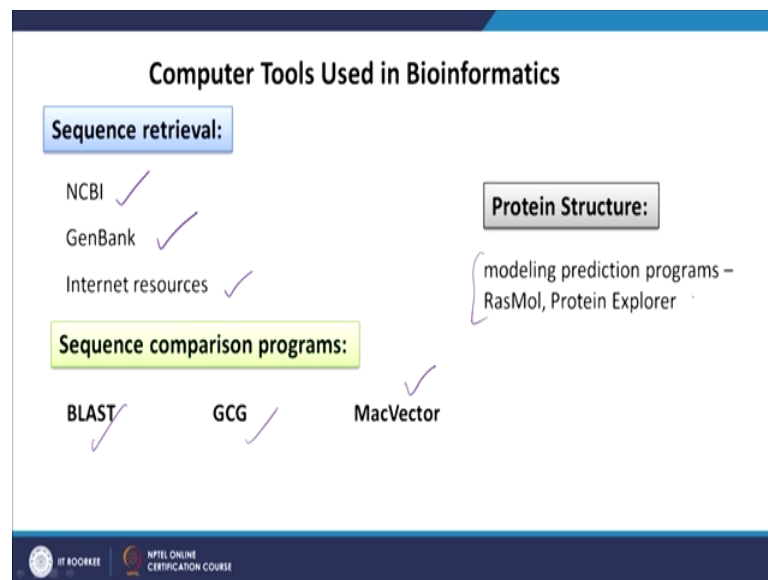


Now let us come back to central dogma. So, our central dogma if you remember is I have DNA, which stores data we have RNA which is messenger basically in (Refer Time: 16:31) is replaced by (Refer Time: 16:32) and then RNA is expressed into protein. Now at each of these steps we can do different kinds of analysis, if you are sequencing genome DNA genomics sequential for example, meta genomics.

So, it is a Sanger sequencing if you do for RNA you can find out the RNA function you can find out RNA structure, if you do at protein level you can do a lot of you can do protein sequence, we can find out what the sequence is amino acid I think you can do 3D modelling and try to understand, what is structure would be again the 3D modelling of protein sequence to understand it is primary structure secondary structure tertiary structure quaternary structure.

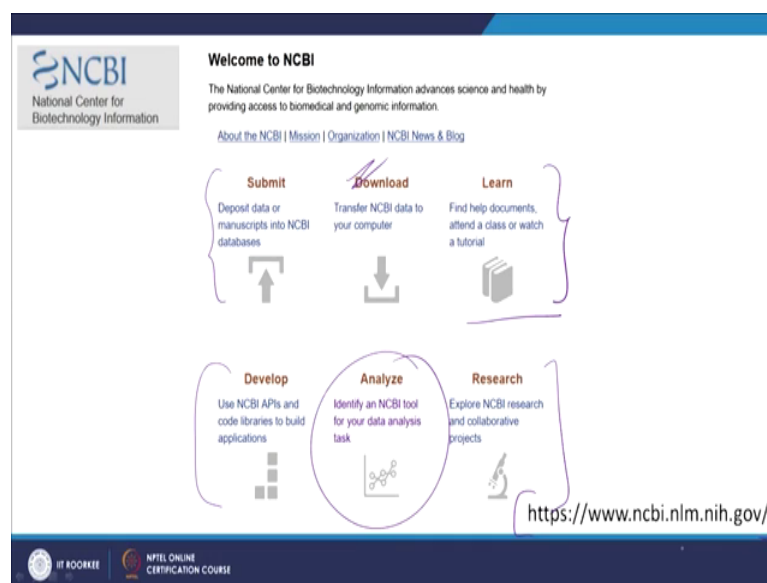
Now, all these and predict it is function all this is part of your bioinformatics by the way modelling and bioinformatics and then you can also do phenotyping from your data. So, at all levels of this central dogma you can get the; you can get data and we can do we need bioinformatics to do different kinds of analysis.

(Refer Slide Time: 17:27)



Now, again as a division of what you already studied the basic tools used in bioinformatics most commonly for sequence retrieval to find out the sequence matches NCBI, GenBank and internet other internet resources, that I will be talking about in the next lectures for sequence comparisons most commonly used is BLAST, GCG or MacVector for protein we have some RasMol, Protein Explorer and many more.

(Refer Slide Time: 17:50)



If you log into NCBI pay this is how it will look like to you and there will be icon welcome to NCBI page and now look at the 6 different options that, they have on their home you can deposit data or manuscript into NCBI database. So, let us I did a study in India and now I have collected from sequences I mean I have sequence some samples and I have sequences and I can deposit those sequences.

Now their 2, 4 benefits of depositing sequence to GenBank or to NCBI to GenBank because, now it is most generous will ask you have you deposited your sequence somewhere and the reason is, because he want more easy availability of sequences and he want more transparency in our research. So, the sequences if you deposit in NCBI you can say yes I deposited in NCBI or in GenBank; this is my accession number is my annotation and, then you can publish it in a manuscript and if you have published to manuscript you can deposit manuscript. So, if anybody is matches if their sequence to your sequence in this case my sequence then they will know ok; this sequence was found in this kind of environment and this was the study there is a manuscript of the study that was working on it.

So, there is an option for submitting and then you can also download this is very important. So, what NCBI does is that it should send a request to NCBI through internet it will take your request ok. Now you are now you are in waiting and they are many other users across the world has submitted their requested one by one by sequence NCBI

processes there high performance computing centres will compute the a query will solve it and give them the data and there is the waiting time, but if you have a good computer in your house in your lab in your room, and then you could you can do you can download the NCBI database and using very simple Linux based a bash commands or whatever platform your using windows Linux windows very simple commands you can use to actually to align data to do all kind of analysis that NCBI can help you do. So, if you download NCBI database you can do lot of analysis and this is a beauty of NCBI that it is open it is very transparent.

Next is, if you have any questions you want to learn more and I highly encourage you explore, this part if you are at all intrigued by bioinformatics at all intrigued by the latest advantages and advances in environmental microbiology is to export the learn portion it can help you find documents attend classes watch tutorials. For example, you have a genetic sequence and you want to align it and find out what bacteria it is, and or if you do not know exact bacteria, then what kind of the bacteria it might be and where else has it been found.

For example, I remember when I was working on oil spill in Gulf of Mexico for my mom masters for my master's thesis I got I got some sequences by Sanger sequencing and not only was interested in knowing, what kind of about; what microbes are present, but I also wanted to know were these microbes are found in other oil reservoirs or areas where the oil spills or not because the they were the I can alright there is these microbes might be associated to oil spills reservoirs.

So, if you want to do this kind of analysis and you are not sure how to go about it then you can click to the learn option and learn more about it, and then we will be discussing this in next lecture and then what you can do with you can actually use the in this develop you can use NCBI api and build your own applications. So, let us say you want to do some kind of analysis that is not available here you have a cool idea of what kind of analysis could be have look for the researchers you can jump into this and code it.

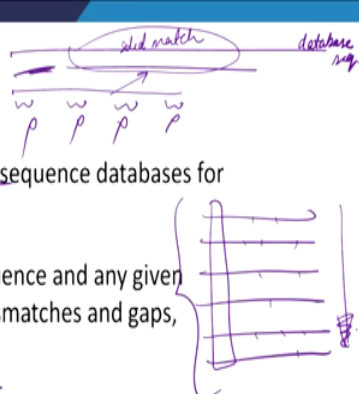
And then this is very important let say you are not very clear about, what kind of analysis to you click analyse it will guide you step by step and, then if you want to collaborate with NCBI they see their bioinformatics company and you are a bioinformaticians, then you can collaborate with NCBI explore research and collaborative project here and this is

the webpage of NCBI and the homework for this week actually involves you going to NCBI page aligning the sequence that I will give you annotating it finding the person maths; similarity and then translating the sequence and then running the alignment again.

So, once you will be using blast n or blast x depending on the sequence I give you in the other case you will be using blast p for protein, and then you have to compare how the data is different in your case ok. So, please bookmark this slide make sure that you visit NCBI familiarized yourself with it and then you can do good in your homework.

(Refer Slide Time: 22:23)

Similarity Searching:



A tool for searching gene or protein sequence databases for related genes of interest

Alignments between the query sequence and any given database sequence, allowing for mismatches and gaps, indicate their degree of similarity

The structure, function, and evolution of a gene may be determined by such comparisons

IT ROOKEE | NPTEL ONLINE CERTIFICATION COURSE

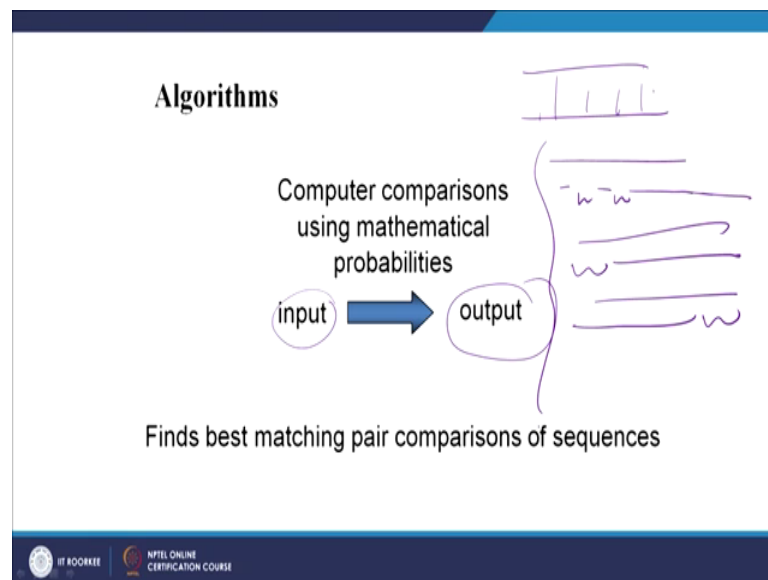
Homework all righty let us look at similarity searching um. So, the reason, why we use NCBI most often is for similarity searching and we want to find out what is our sequence most similar to and we can do it for gene this sequence we can do it for protein sequence and um, then we align we align the alignment between the query sequence and any given database sequence align for mismatches in gap indicate their degree of similarity. So, this is a beauty of the NCBI it allows is it allows mismatches and gaps.

So, for example, let us say I have a sequence here this is your database sequence database sequence and if I allow if I do not allow for gap. Let us say it starting here and my actual sequence is starting here, if I do not allow for gap then I might have lot of mismatches does not match these are region that do not match, but if I allow for gap and I shift it if I shift the same thing here then there is a very solid match.

So, this kind of gaps and skips and misses or permitted in NCBI to account for the possibility that your clone or will you fragmented your DNA from DNA to is different from where the other people in who are submitted to database fragmented the DNA from all right. The structure function and evolution of a gene may be determined by such comparisons. So, if you know if you can read the manuscripts of people who have submitted their genes, if you can read about it and know about it you can get an idea of function you can get an idea of structure and also get an idea of evolution how is the gene changing, because number to understand the evolution, we need to know how the genetic sequence is changing.

So, if you have genetic sequences you can align them on conserved region, this we talked about in last class. So, let us say these are your genetic sequences from different times with different regions and there is a conserved domain. You align it on the conserved domain and then you see how are changes happening here, and then this way you will know, what is the rate of evolution of this particular gene under given circumstances all right algorithms.

(Refer Slide Time: 24:38)

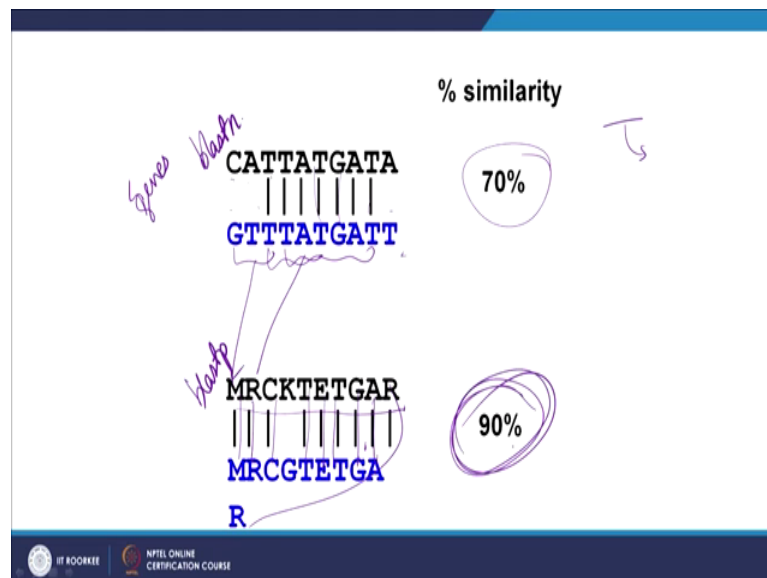


So, competitive comparisons using mathematical probabilities you have some input of data and I will give you some output and the way the mathematical algorithms work is there they find the best matching pair when comparing two sequences. So, basically there

are two sequences that they compare they compare it by allowing all kind of mismatches all kind of gaps.

So, maybe if this is; how it will look like maybe to look like this or maybe to look like this we are giving up here we giving up in between so, tries for all permutation combination and then tells you what is the best match scenario right.

(Refer Slide Time: 25:18)

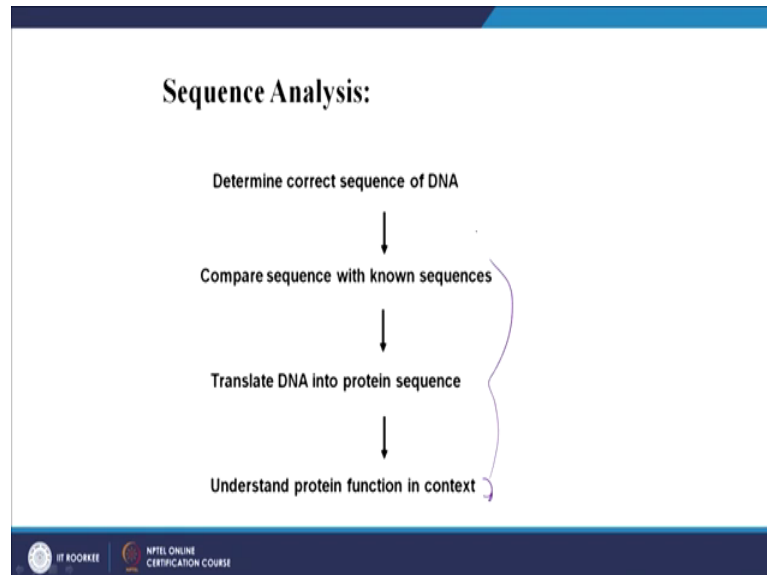


So, let us look at these two cases. So, here we have two sequences and you want to find out the match, if you look at here we notice that there are 70 percent match align for this gaps. Now, if you can translate; now this is genes. So, I hope you can recognise that C stands for cytosine, A stand for (Refer Time: 25:39) T stands for thymine and so, on same thing and T matches with TATGAT. So, 7 out of 10 matching so, we have 70 percent matches, now if you translate this into protein. So, you read it like this and you translate it into protein then this is what you have.

So, here you are doing blast n or blast x and here you are doing blast p. So, you are comparing the proteins, and then you notice that the match percentage is much higher; why would this be. So, because if you remember from the amino acids different code on skin code for same amino acid. So, to account for this degeneracy we prefer that you translated into protein and then match you get a better; similarity and also better representation of what is perhaps happening, because the degeneracy in the genetic

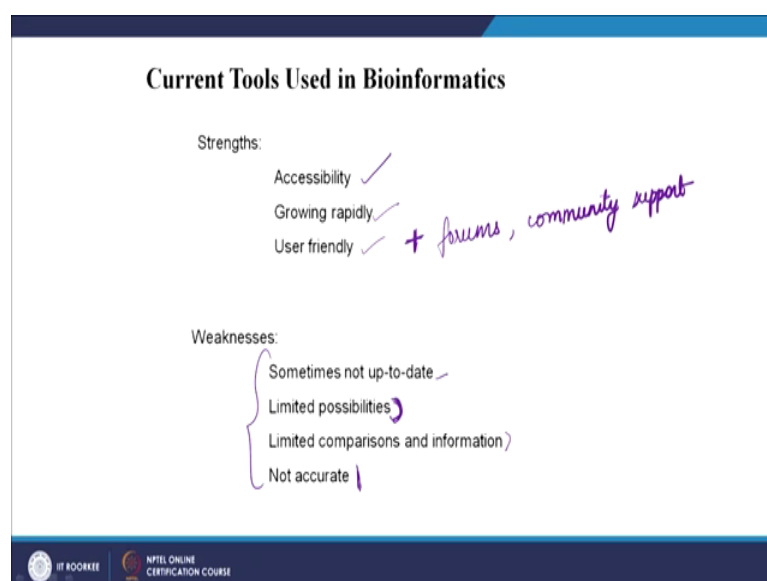
sequence will not impact of the functioning of self it will not impact the functioning of protein, because protein is right protein has the same.

(Refer Slide Time: 26:51)



So, blast p is very useful and now let us look at sequence analysis; first you determine the correct sequence of DNA, then you compare the sequence with non sequential you translate DNA into protein sequence and then you understand protein function and context of all the data you are gathered here all right.

(Refer Slide Time: 27:02)

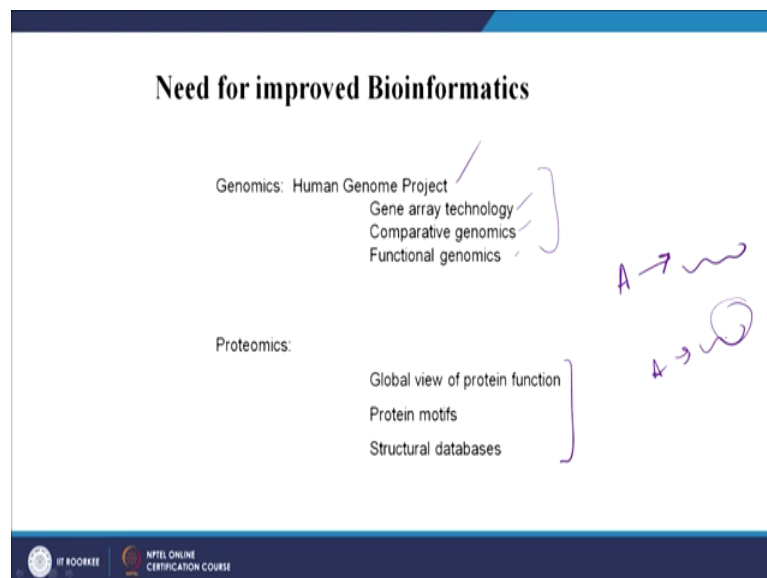


The current tools used in bioinformatics have certain strength a very accessible so, most of the tools of freely available online. In fact, if it is proprietary tends to not have a big market. So, people are making their tools the database available freely to everybody they growing very rapidly, there very user friendly and I must add that they have very good forums where you can ask for help you can post questions and then they are very good community support here most of them have very good community support, what are the weaknesses of current tools.

Now sometimes not up to date, because some of them can be really obsolete it is limited possibilities and one of the limitations. We talked about last time was there is not a single tool that can do all kind of comparison, that someone might be interested and then we have limited comparisons and information and many of them are not are not accurate they have statistical errors we have issues there.

So, this is I like to think of this is room for improvement bioinformatics as it is right. Now is not certain stone it is growing very rapidly as it says here and does for evolving very fast when it comes to bioinformatics. So, this is there the room for improvement is and this is where many people are working to improve our tools.

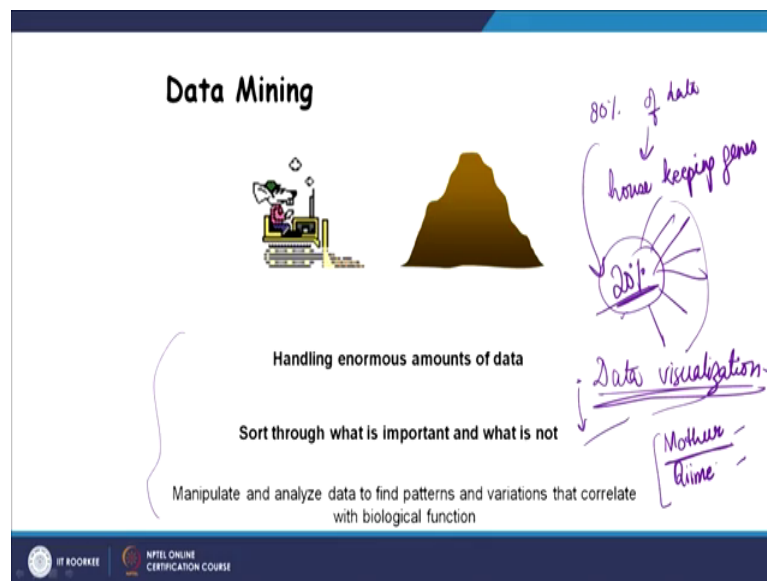
(Refer Slide Time: 28:22)



So, need for improve tools for human genome project gene array technology comparative genomics and functional genomics for this we need specialised sophisticated tools for proteomics, we need to understand global view of protein functions for example, you

know demine with protein a and in one particular cell it does a particular function, and in then the cell can survive, but put it in another cell and then it is not acting in isolation it is acting in presence of many other inhibitors and many other molecules many other proteins many other RNA. So, thinks it is behaviour in different source will be different. So, which very important for us to understand the global view of protein function not just say that it measures to this protein found in the bacteria?

(Refer Slide Time: 29:06)



So, it must be doing the exactly same thing say and similarly for protein motifs and for structural databases. So, basically bioinformatics allows us to do data mining it allows us to handle announce amount of data it makes it approachable and acceptable it helps us to understand, what is important; what is not because here is another think in meta genomics these are rough estimated eighty percent of your data.

So, 80 percent of your base pairs would be housekeeping genes. So, 80 percent of genes that you will detect from meta genomics will be housekeeping genes which are found in every bacteria. So, you would not really see any much difference you would not really see the; you would not really see that the trend that you are looking for. So, only in the meaning for data for you would be 20 percent. So, bioinformatics will help you find out what you 20 percent is, where it is and you can do all kind of statistical tools.

You can use them you can do all kind of analysis to answer your questions to test your hypothesis it also allows us to manipulate and analyse data to find pattern, then

variations that correlated with biological function, there is another added part of this bioinformatics that I would like to mention briefly here is data visualisation most of the tools bioinformatics tools.

Now, there they come with very good inbuilt data visualisation technique, and if you do not have data visualisation techniques and the tools then they will recommend you, why do not you use package a software b for visualising your data and this is very important because we have so, much data, but we cannot really make sense of it unless we can see it as it is every visual creations. So, besides all this we also have the part portion where we using bioinformatics to visualise the data.

And in the next one of the next lectures I will be talking about two, I will be comparing two self-software platforms, when is mothur and another is qiime. So, it is written as q i i m e pronounced as chime and I will discussing; how each of their advantages and what their disadvantages and one of the advantage for one of them is that they have very good data visualisation. So, dear friends this is all for today and in the next lectures, we will be actually looking into these tools and I will be briefly telling you briefly leading you on how to use them.

Thank you.