Functional Genomics
Professor S Ganesh
Department of Biological Sciences & Bioengineering
Indian Institute of Technology Kanpur
Lecture No 02
The Genomics Era

So welcome back to this course functional genomics and this is the second lecture of week one. So in the previous lecture we looked into how the field of molecular biology is evolved and that led to the emergence of a new era called genomics era right. So that is what being summarize in this slide here.

(Refer Slide Time: 0:42)

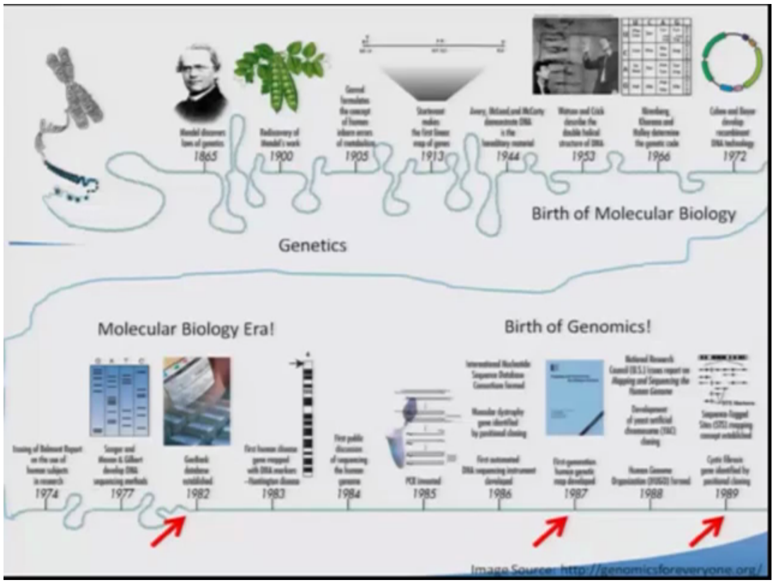


Image Source: http://genomicsforeveryone.org/

You see that we started with the Mendel’s concept to the DNA double helix how Watson and Crick solved the structure of the the DNA and then all the codons and then you have for the first time developed what is called as a recombinant DNA technology in 1972 and that led to what is called as DNA sequencing write by Maxam and Gilbert in 1977, there you have now a tool by which you are able to sequence the DNA.

So the moment this technique was developed and shared with all the scientists and you will find that there are so many labs now are able to make the recombinant DNA either from the DNA itself or from the RNA by converting it to cDNA or copy DNA and they started sequencing. So obviously I going to have you know large amount of data coming up because you know that the DNA sequences made up of four bases, the combination of which gives us sequence.

So that led to a new concept called gene bank, meaning it is a bank which stores not currency but the sequences. In 1982 the American establishment went to form a data base called gene bank data base where you, if you sequence a DNA you are able to deposit the sequence and anybody who is trying to understand the function of the gene or DNA can also access this data base, it is a free open data base and that is the first one but soon there was another subject of as called EMBL from Europe and and Japan also developed very similar kind of a sequence data base.

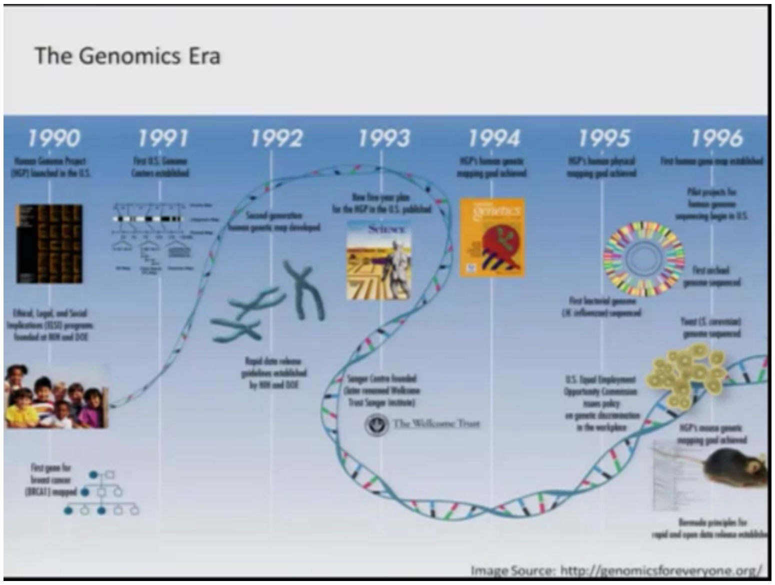
And this revolutions led to understanding of the genome better and then the first gene or for a human disorder you know the gene was mapped using purely based on sequence that are present in the chromosome, these are repeat sequences they are used 1983, a gene was mapped and then then soon they understood the power of understanding the genome that would really help us to understand how the diseases are caused and how the you know the different races of the human population h different than give us some unique advantage.

So then the American government came up with you know a kind of a concept that we should sequence the human genome using the technology that were available and this was still a proposal but that was in 1984, then 85 the PCR came because that made things much easier now we can no really wait for large amount of DNA, we can have small amount, we can make millions of copies and then you know likewise the technology its not only the chemical based approaches but electronics and instrumentation based approaches also evolved then there were you know discoveries or inventions which led to what is called as automated DNA sequencers where you can sequence a large number of samples.

A computer is able to analyse the sequence and give you the sequence straight away that came in. So that really led to what is called as an explosion because then people sorted using this methodology for sequencing the human DNA and 1987 the first generation human genetic map was developed you know a kind of what is called as markers which are landmarks kind of a mile post in the genome and then they came up with that concept. And then 1988 the human genome organisation was formed and U.S. is national research council of the U.S. issued a report that we should map and sequence the human genome and then as a prelude the yeast chromosome was sequenced and that is how they started testing methodology and the challenges that come along with analysing the sequence and how to store them and so on.

So this has led to a new era that you call as genomics, the birth of genomics and 1989 you know the first successful discovery of a gene causing human disease without knowing what is a function of the gene. This is called as positional cloning that led to the discovery of the gene for a disease called Cystic Fibrosis so that is the first one but soon that approach that is you are going to use that in markers these are nothing but repeat sequences and use them to understand, identify, locate gene has become much easier and thousands of human genetic disorders gene have been genes have been identified since 1989, the Cystic Fibrosis gene was identified.

(Refer Slide Time: 5:57)



The Genomics Era

1990
Human Genome Project (HGP) launched in the U.S.
Ethical, Legal, and Social Implications (ELSI) programs founded at NIH and DOE
First gene for breast cancer (BRCA1) mapped
1991
First U.S. Genome Centers established
1992
Second generation human genetic map developed
Rapid data release guidelines established by NIH and DOE
1993
New five-year plan for the HGP in the U.S. published
Science
Sanger Centre founded (later renamed Wellcome Trust Sanger Institute)
The Wellcome Trust
1994
HGP's human genetic mapping goal achieved
genetics
1995
HGP's human physical mapping goal achieved
First bacterial genome (H. influenzae) sequenced
U.S. Equal Employment Opportunity Commission issues policy on genetic discrimination in the workplace
1996
First human gene map established
Pilot projects for human genome sequencing begin in U.S.
First archaeal genome sequenced
Yeast (S. cerevisiae) genome sequenced
HGP's mouse genetic mapping goal achieved
Bermuda principles for rapid and open data release established

Image Source: http://genomicsforeveryone.org/

So going to you know this lecture we are going to run through some of those important mile post so you can see that with every advancement when you talk about human genome sequence and so on. It also comes with certain issues, the issues could be you know when when you sequence the DNA and then when you want to put that sequence available in a in a portal in a data base that is accessible for everyone so that is one of the major decision taken by the human genome organisation that that all the sequence should be deposited in a portal that is accessible to anyone any everyone.

So that also comes with its own challenges like if you have the sequence of different races and if somebody identifies what are the sequence that gives you a risk of developing a disease and so on can this be used against certain races or individual. So there were ethical legal and social implications that were discussed and certain guidelines are being formed by the consortium and then these discoveries you know understanding the sequence and coming up with markers led to as I said many new genes that are implicated in disease process.

One of them being the Breast Cancer gene what is called as BRCA1 was mapped likewise, so again it is a very landmark discovery in terms of Cancer genetics perspective. Then with the such you know advancement we have several landmarks than in 1992 the second generation human genetic map was developed which really tells you where are the markers for every chromosomes. The density has gone up so people are able to use this markers to you know understand and and investigate the chromosome better and then you have finally the human genome programme sequencing programme was established it is a five year plan.

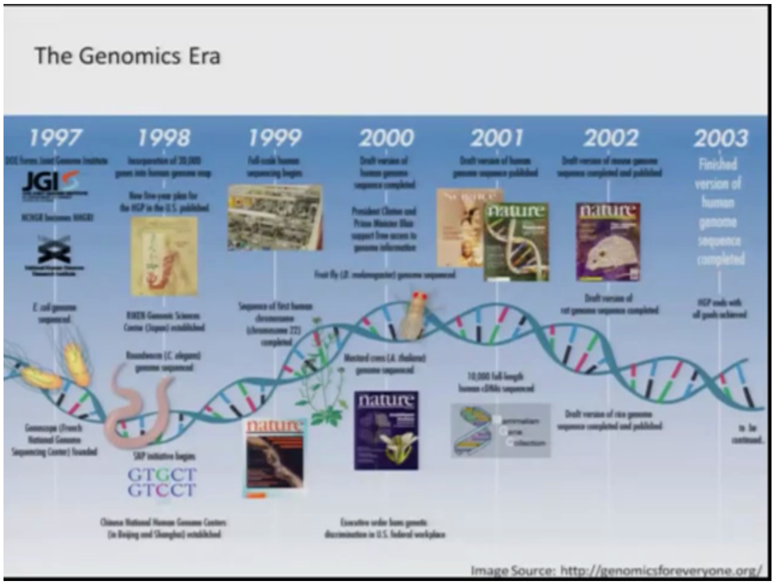
It was established and then and then there are consortium mode because there are about five- six countries came together and they have kind of understood and agreed upon as to which region of the genome they should sequence and they welcome for us to sanger centre you know join the force in you know in speeding up the process sequencing and in 1994 the human genetic mapping you know pretty much the (())(08:33) able to make pieces of the DNA of human genome label them and sort them and able to put all the mile post, so you have pretty much the framework has been done for the human genome sequencing.

So that really has not process of the sequencing complete physical map, the distance between different segments of the genes and others came up in 1995 and then with all this you know advancements the bacterial genomes which is much much smaller the influence of genome was sequenced and then there were other issues that came up along with certain you know when they how sequence the genome and then they also used government came up with the policy how genetic discrimination work place cannot be you know tolerated so again these are ethical issues right.

The other simpler organism like yeast sequence was also completed in 1996 because these are much smaller genome and by then the scientific community also sort of understood that it just sequencing the human genome alone will not suffice. We need to have very good models to understand the genes functions. So one of the models that are closer to the human it very good in terms of genetic tool was mouse and therefore they decided that even the mouse genome should be sequenced and then genetic map was achieved in 1996 and then and then there was a consortium which together discussed and is called as Bermuda principles that really agreed upon that any such large scale sequence h initiative funded by the government be it the U.S. or the Europe or the Japanese or Chinese the data should be in public a domain therefore all can access is called as open data policy that was agreed upon in 1996.

So that really has changed the way people can look at the genome sequence because these centres have understood creating the data that is generating the data is one but analysing them and understanding is going to be much more helpful in task and therefore it should be open therefore many groups who have competence in analysing the data can do is going to really help h the society. So that is the reason why they went for open source h data being deposited.

(Refer Slide Time: 11:13)



The Genomics Era

1997
DOE forms Joint Genome Institute
JGI
NCHGR becomes NHGRI
E. coli genome sequenced
Genoscope (French National Genome Sequencing Center) founded
1998
Incorporation of 30,000 genes into human genome map
New five-year plan for the HGP in the U.S. published
RIKEN Genomic Sciences Center (Japan) established
Roundworm (C. elegans) genome sequenced
SNP initiative begins
GTGCT
GTCCT
Chinese National Human Genome Centers (in Beijing and Shanghai) established
1999
Full-scale human sequencing begins
Sequence of first human chromosome (chromosome 22) completed
nature
2000
Draft version of human genome sequence completed
President Clinton and Prime Minister Blair support free access to genome information
Fruit fly (D. melanogaster) genome sequenced
Mustard cress (A. thaliana) genome sequenced
nature
Executive order bans genetic discrimination in U.S. federal workplace
2001
Draft version of human genome sequence published
Science
nature
10,000 full-length human cDNAs sequenced
Mammalian Gene Collection
2002
Draft version of mouse genome sequence completed and published
nature
Draft version of rat genome sequence completed
Draft version of rice genome sequence completed and published
2003
Finished version of human genome sequence completed
HGP ends with all goals achieved
to be continued.

Image Source: http://genomicsforeveryone.org/

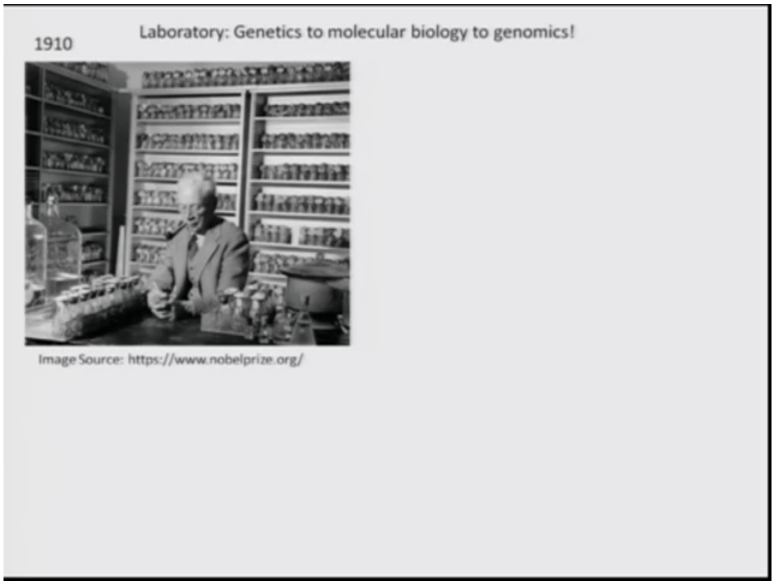
Then came the genomics era 1997 to what you see now, their large number of genome sequence or made available so you have for example many initiatives that came for example equalie genome one of the common bacterium that is present in your gut equalie sequenced and then there are other centres for example which started sequencing other models for example the (()) (11:37) genome sequence was initiated and they also developed other ways for example it is not only the sequence that helps us. We need to understand what is the variations between two individuals or one that are healthy and not healthy.

If really the DNA sequence gives you that attribute then even the variations should be understood so there was another initiative which really helped to understand the single nuclear type polymorphism or the variations that came in and then in 1990 you know that is when really you know the full scale human sequencing began. People started doing it in a rapid pace because technology is developed you have very good automation came in instrumentation came in , they are able to do better and then 99 of course the first human chromosome that is chromosome 22 smaller chromosome sequence was released.

So that led to a series of you know data release with for example Drosophila genome was sequenced, mustered another plant sequence was released and then 2001-02 you see that that human draft version of the human genome sequenced was published and then mouse sequence was published and then also they started in parallel sequencing all the RNAs from different you know tissues. You take tissue from human body convert the RNA into cDNA and sequence therefore you know what are the genes that are expressed in a given tissue that also began called as a mammalian gene collection.

So which led to you know the sequencing of humans cDNAs. So that is how went on and 2003 of course the finished version of the human genome sequence accomplished, pretty much what we have now for the human genome. Whatever reasonable chromosome could be sequenced was completed in 2003. We have them in the portal, so anybody who wish to look into what are the genes present in a given chromosome we can go on look and there will a session after a couple of classes that we would introduce you to various such data bases. So you can also learn how to browse through this data base and then understand how the genes are.

(Refer Slide Time: 14:03)



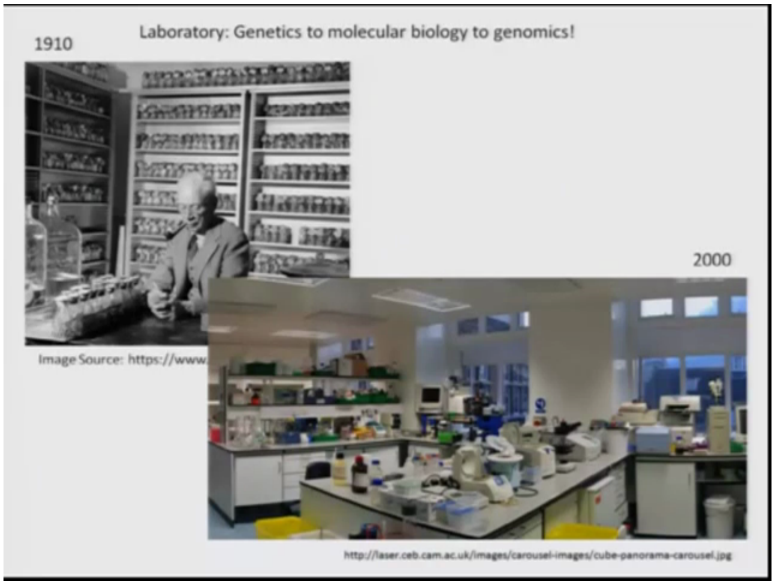
1910

Laboratory: Genetics to molecular biology to genomics!

Image Source: https://www.nobelprize.org/

So as I told it is the sequence information that that we have now so many different organisms being sequenced and so on. There is an enormous task which did not come only from the lab genetics or bio-chemistry based approaches. The way the science was done the genetics or molecular biology was done has changed from a lab centric to more industry like approach. What I am showing here is photograph from Morgans lab where one of his co-workers studying Drosophila genetics. All you see that at the back rack which is the bottles are kept or different strange of Drosophila being you know maintained there. That is how it began the genetics began.

(Refer Slide Time: 15:06)



Laboratory: Genetics to molecular biology to genomics!

1910

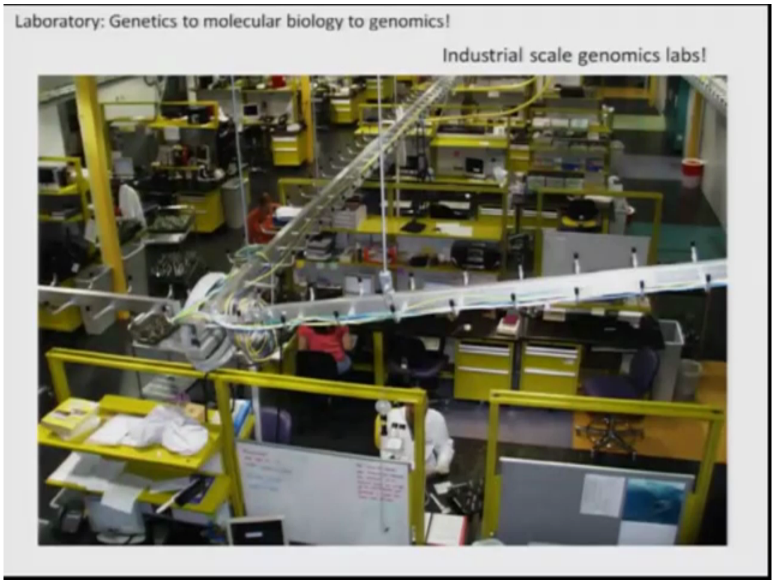
Image Source: https://www

2000

http://laser.ceb.cam.ac.uk/images/carousel-images/cube-panorama-carousel.jpg

But in 2000 or so when molecular biology was really at its prime, this is how the labs looked like. So you have lab benches, we have certain equipment but still it is a manual work. So you have to take the solutions put them in a tube allow the reaction to get (())(15:13) and load it on a gel analyse and so on. Still it required the manpower to conduct experiments.

(Refer Slide Time: 15:34)



Laboratory: Genetics to molecular biology to genomics!

Industrial scale genomics labs!

But things have changed when with advent of what is called as genomics it was only possible because of the automation and a large number of bio-medical equipment that were developed in parallel to carry out many of these experiments without human intervention and to analyse the data, the what is called the you know (())(15:45) screens all became possible you know in the later 2000 onwards and it is become like you know if you look into large genomics lab it does not look like the lab that was shown in the previous photograph.

It is more like a kind of an industry and this is called as industrial scale because you scale up everything there will be hundreds of people working, hundreds of equipments you know thousands of samples being sequenced at the same time and and it led to you know data that is being generated just like anything.

(Refer Slide Time: 16:22)

http://www.zeiss.com/microscopy/

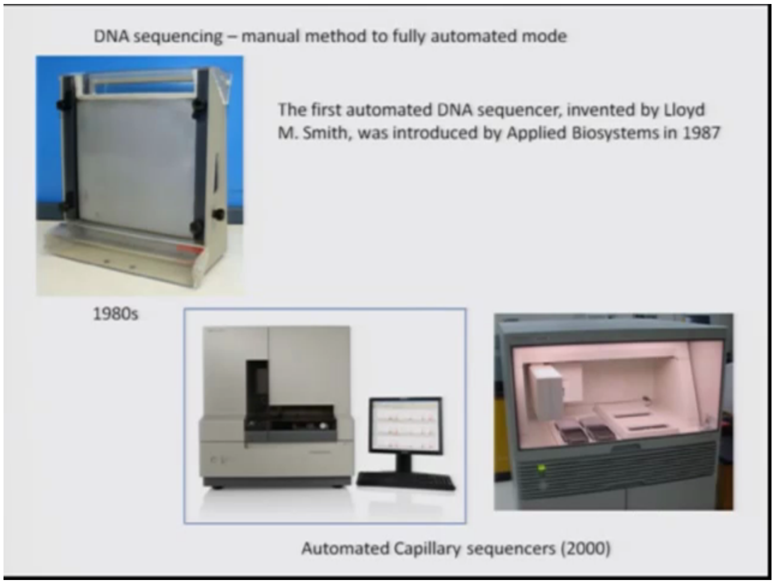
https://www.beckmancoulter.com

So if you look into the equipments there is a you know revolution if it is not just evolution. You know most of you must have used microscope in the college or schools that is something that is shown on the left side is simple microscope that was used during Morgan period but that is no longer you cannot use that for when you are analysing large number of cells so now you have what is shown on the right side is a work station what you call as you can pretty much put the sample and programme at the microscope would look at the cells score them pattern and give you the data. So that is without anyone sitting there that is possible because you know the technology is developed and that kind of an approach is required if you want to look into the dynamics of large number of samples.

One person cannot really do all these things so only missions can do so that technology really played a major role in changing the way biology done be it looking at the cell even in liquid handling for example, the one shown on the left side is some of you must have used the (()) (17:30) that in colleges you must have used for taking the reagent to what you called as calibrated little hand held (())(17:38) to know you have work stations that is shown on the right side. Is the machine which you give a programme as to how what volume of sample should be added in which tube and it does in more time and with precision without any error so that is how things have changed.

So now we can pretty much set the system where in you have for example cells put in tubes that are grown there and you can access system to extract the RNA convert that into cDNA and sequence it, analyse and give a data. So everything can be now automated, the platform would do and at the end of the day you are going to have the sequence which you have to validate analyse and so on. So that is how the system has changed and that led to you know more challenges we will talk about a little later.

(Refer Slide Time: 18:46)



DNA sequencing – manual method to fully automated mode

The first automated DNA sequencer, invented by Lloyd M. Smith, was introduced by Applied Biosystems in 1987

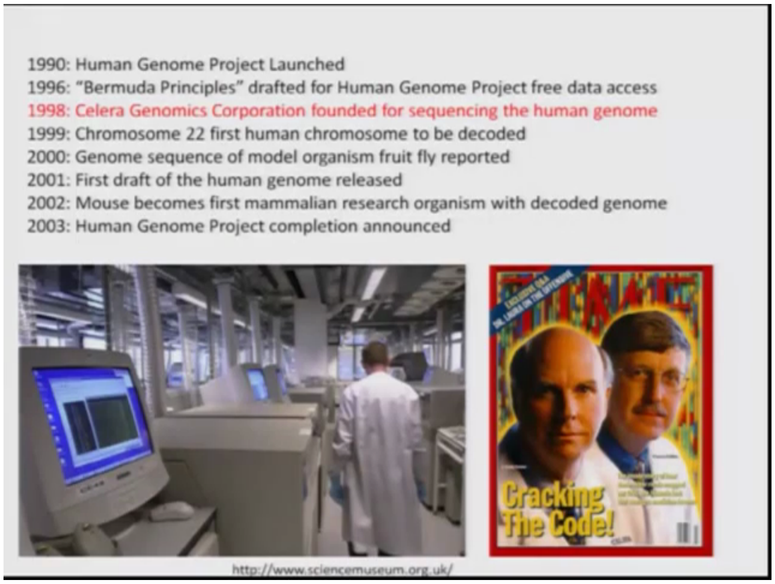
1980s

Automated Capillary sequencers (2000)

So like ways when you talk about sequencing which is one the major contribution in the genomics field that another sequence of the DNA or RNA being generated. So the one shown on the left top is an electrophoretic setup that developed you know that is how some of us have done DNA sequencing in 1980. So these are the manual gels we have to run it and then and then at the end of the day you have to expose the gel to extra film and then you will have bands then you read the sequence. In a day probably 100- 50 bases you can do and this has changed with the advent of what is called as automated sequencers in early 2000, so you have machines which can take up to 380 samples.

Do all the sequence on its own, give you the sequence in soft copy be it in server or pen drive whatever it is. And you can analyse them so without any human intervention. So things have changed so the (())(19:36) companies really helped in developing such technology. One of them was Lloyd Smith who introduced the automated the first automated DNA sequencers and that was you know done by a company called Applied Bio-systems will really changed the way the human genome was sequenced and that led to the scientific community to take up many different organisms for sequencing. Now even these approaches are out dated we will discuss about the contemporary approach people used for sequencing the DNA.

(Refer Slide Time: 20:17)



1990: Human Genome Project Launched
1996: "Bermuda Principles" drafted for Human Genome Project free data access
1998: Celera Genomics Corporation founded for sequencing the human genome
1999: Chromosome 22 first human chromosome to be decoded
2000: Genome sequence of model organism fruit fly reported
2001: First draft of the human genome released
2002: Mouse becomes first mammalian research organism with decoded genome
2003: Human Genome Project completion announced

http://www.sciencemuseum.org.uk/

In this period you know generally the molecular biology genetics, this was pretty much done by research labs which are more interested in understanding discovery based science, understanding how the genome functions and so on. It was not a commercial entity it was like you know a pure science but you know the implications of understanding the genome into the health has even become a business because you know everyone wants to live better, live longer, disease free so all of us go to doctors and they prescribe medicines. The medicines come from how much you

(())(20:53) industry which are not non-profit organisation. They are for profit, they do R and D to come up with the drugs.

So the genome and its variations and its implications the health is also changed the way industries look into this field called genomics because now they have interest and the molecular biology has gone into industry that is when we came up with industry called biotech industry which uses the genome knowledge to develop new therapy or therapy protocol or interventions strategy or drugs. So that is what something shown in here in this slide, 1990 the human genome project was launched 96 the Bermuda principles that is drafted for human genome project saying that the data generated is free for everyone. Since it is free for everyone there are also people who are interested in understanding the data and using it for you know as a commercial point of view in terms of therapy or drug and so on.

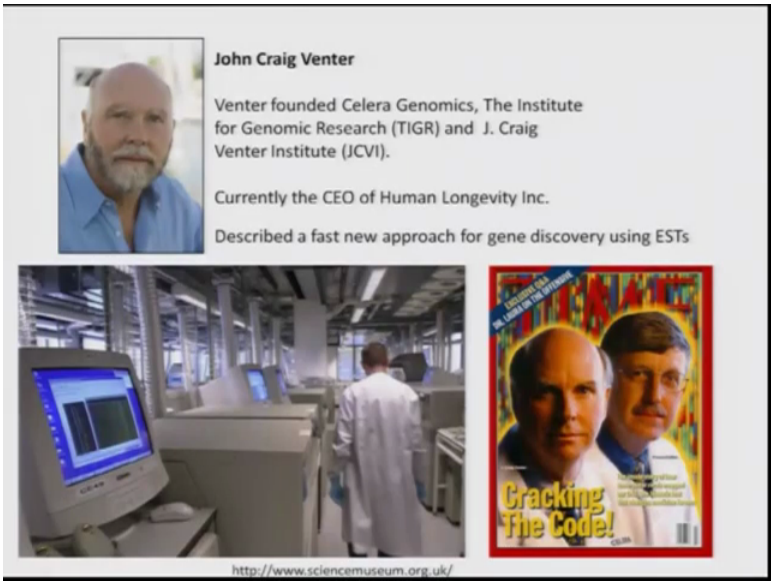
In 98 you know one of the famous researcher you know launched a commercial company called Celera Genomics and they said that we will finish the human genome faster than the 6 nation consortium and they went with altogether different approaches for sequencing the DNA and they have used heavy usage of computers we will talk about little later and finally they gone into many other disciplines which could be of commercial applications. 99 the chromosome 22 first even chromosome was you know was decoded model systems you know genomes have come 2001 first draft of the human genome was released and 2003 human genome project was completed and this was announced.

It was not only the human genome project that consortium but even the cellular genome makes by then within a span of 4-5 years. They have completed the human genome sequence as well so which the other consortium took several years to complete. So these are somehow the you know the famous you know captions one in 2003 you can see the Time magazine has shown the two people one is Francis Colin on the right side with specs, he is the one who was championing the human genome consortium project and on the left side is Craig venter, he is the one who founded the Celera Genomics and completed the genomics sequencing part you know much much quicker than the other did.

The reason being they have used very different approach and they have used what is called as you know computational tool to analyse the sequence which really help them to finish faster.

This also the time when you know things have changed what you see now the left side is you know set up which is like industry like, you know thousands of people working and cutting down the cost by using a large number of equipment and computational tools that really made you know a huge difference.

(Refer Slide Time: 24:06)



John Craig Venter

Venter founded Celera Genomics, The Institute for Genomic Research (TIGR) and J. Craig Venter Institute (JCVI).

Currently the CEO of Human Longevity Inc.

Described a fast new approach for gene discovery using ESTs

Cracking The Code!

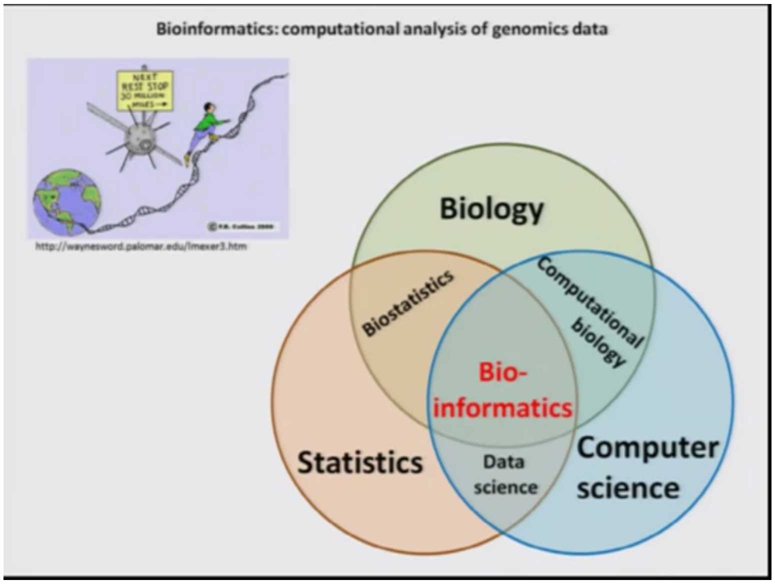
http://www.sciencemuseum.org.uk/

And this gentleman Craig Venter who founded the Celera genomics not only did that but he also established a new institute called The Institute for Genomic Research to understand how genome variations can contribute to health and disease and then the Craig Venter institute and so on. So he was also the one who made a synthetic living organism by you know putting together DNA sequence of different organisms and made a new you know species of course is a lab centric it has never come to the free atmosphere.

So now he is heading this you know company called Human longevity where he is trying to make the living much better without much disease and so on that is his ultimate aim but he is known for also this the speedy discovery of gene and the human genome sequence was started, the aim was to sequence the genome and people have used various approach to find where the genes are. So it is very laborious and challenging task but what he went and did was to do different kind of approach called EST which stands for expression sequence tags. The approach is that you now make cDNA libraries from various tissues for example human body and then sequence only a part of the cDNA therefore you have about hundred bases randomly here and there for a region that is transcribed.

Now you go back and look at the genome sequence and see which region has identical sequence that would tell you where are the genes are. So that is an approach that you know developed by Craig Venter and really helped in understanding where the genes are located otherwise it was extremely difficult because as we know there is a one percent of the genome is represented in transcript so it is going to be herculean task to find where the genes are.

(Refer Slide Time: 26:45)



Bioinformatics: computational analysis of genomics data

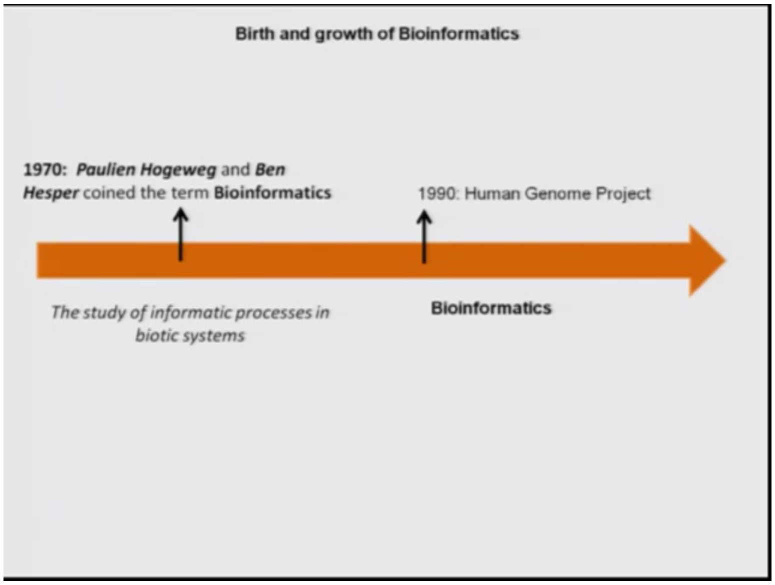
http://waynesword.palomar.edu/lmexer3.htm

In that way he was revolutionary in terms of vision and so on. Hence theses you know approaches also led to you know enormous data the wealth you know for example if you look into the human genome sequence. If you you know if you pull out the DNA from our body it can go all the way to moon and come back that is the length of the DNA. So and if you get the sequence out of it and from all the organisms, how are you going to really store the data how are you going to analyse it its humanly not possible.

So that led to a new field of research and methodology development called Bio-informatics. So it is also you know primarily you know now we talk about Bio-informatics as something which is a computational analysis of genomic data it could be DNA, it could be RNA and so on. So it is a field which has a specialization that falls between three major domains of course one is biology because these are biological data then you have computer science because in the data science what you call large data analysis come from you know approaches where by which you are able to automated the process with the computers can take analyse and give you data.

Since you are analysing the you know biology data it is called the Bio-informatics but it also involves a lot of statistics mathematical modelling and so on that is where it is a cross sectional biology, statistics and computer science and that is being Called as Bio-informatics. So is it that Bio-informatics came much later than genomics, so most of us majority even the common public you know hear about Bio-informatics.

(Refer Slide Time: 28:03)



Birth and growth of Bioinformatics

1970: ***Paulien Hogeweg*** and ***Ben Hesper*** coined the term **Bioinformatics**

The study of informatic processes in biotic systems

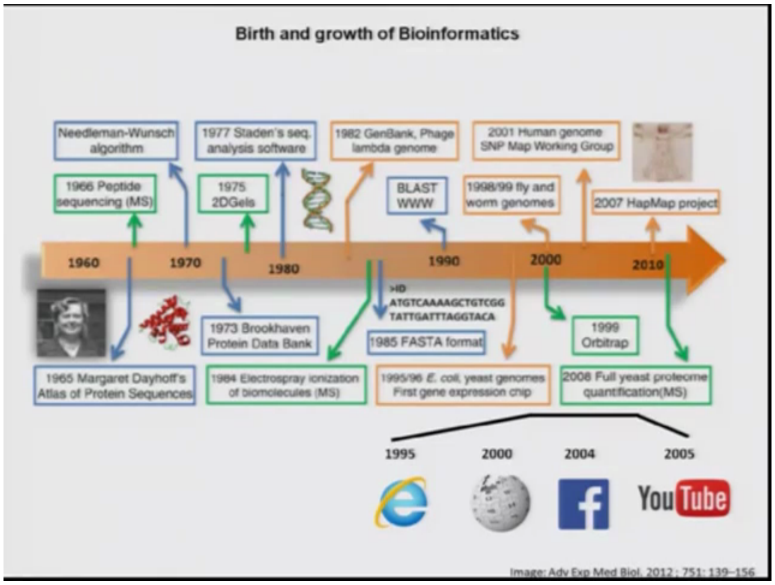
1990: Human Genome Project

Bioinformatics

You must have seen there are various institutes and universities even offer Bachelors, Masters degree in Bio-informatics there are Bio-informatics papers introduced as well in other curriculum. We talk about all these things only after the human genome project has been launched because the well the data that was generated. There could be large number of labs that could really look into and understand so the need or the demand also has gone up just like the information technology boomed in the country and other part of the world but it is not that you know in in late 90 or 2000 the Bio-informatics came in its there in existence much before. In fact the (())(28:43) Bio-informatics was coined by Hogeweg and Hesper in 1970 but the intention was not necessary to analyse the genome data only.

It was like the study of informatics processes in biotic systems that is their definition. Any information that comes from the Bio-field needs to be analysed using certain automated processes because it could demand such an analysis, it is not humanly possible that is how it was proposed but not very well practised by this team but we can look at it is not only in 90s came in.

(Refer Slide Time: 28:32)



Birth and growth of Bioinformatics

Needleman-Wunsch algorithm
1977 Staden's seq. analysis software
1982 GenBank, Phage lambda genome
2001 Human genome SNP Map Working Group
1966 Peptide sequencing (MS)
1975 2DGels
BLAST WWW
1998/99 fly and worm genomes
2007 HapMap project
1960
1970
1980
1990
2000
2010
>ID
ATGTCAAAAGCTGTCGG
TATTGATTTAGGTACA
1973 Brookhaven Protein Data Bank
1985 FASTA format
1999 Orbitrap
1965 Margaret Dayhoff's Atlas of Protein Sequences
1984 Electrospray ionization of biomolecules (MS)
1995/96 E. coli, yeast genomes First gene expression chip
2008 Full yeast proteome quantification(MS)
1995
2000
2004
2005
You Tube

Image: Adv Exp Med Biol. 2012 ; 751: 139–156

It came in much before, I am just putting some timescale here. It all started probably with the development of what is called as Atlas of the protein sequence in early 1965 or so. So that is when people started you know analysing the protein if you remember I said after Morgan his students started working on proteins and protein became main tool to understand how possibly the cell functions, bio-chemistry field evolved and then they started sequencing by you know cutting the protein into smaller peptide and then cleaving them into individual amino acids and so on.

So when they did the process we need to you know how do you arrive at the sequence they have used certain simpler algorithms kind of you know mission reading tools and to come up with the sequence so that is the kind of approach that was taken in early 1965 using you know computational approach which pretty much being used now in peptide sequencing since then onwards. As I said there are in 1970s there are algorithms in a computational tools came in to understand the protein and structure and so on. And then with this you know there are large number of proteins being sequenced like we talk about gene bank which came in much later.

There was a protein data bank which was developed and all the protein sequence was deposited over there so that you know so in that way that is also a Bio-informatics in the sense that they are storing data from biological origin for analysis. So that obviously Bio-informatics existed even before the DNA sequence came in and then there are many other softwares came in until 1980s to understand the proteins and 84-85 that is when the DNA sequence came because the sequencing technology or the tool approach was established with Gilbert and the off course you have 1982 the gene bank and so on. The first forge, lambda forge genome was sequenced and so on.

So like from there you have wealth of data that is more of genomic whether it is from the Drosophila or a human and so on. You will find that you know this genomic data overtook the protein sequence and no longer people are sequencing the protein because the moment you have sequenced the DNA or the messenger RNA via a cDNA you are able to predict what is a coding sequence and if you are able to predict what is a coding sequence with no time you can predict what is the protein sequence.

So the protein data bank after which all derived from the sequence of the mRNA meaning cDNA so now unless for identifying what protein it is like for example 2D (())(32:22) people really do not go for protein sequence because they are just sequence are bit to know what is a sequence and go back and look into the protein sequence derived from the mRNA and match this is a protein, so the protein sequence is pretty much now shelved people do not do but this came up with you know much more challenges because the volume of data that comes from genome sequencing data really grown. So there are new technologies but many of them are based on earlier principles as to how to analyse the DNA sequence but you know with advent of RNA sequence coming up, expression analysis coming up and the variation projects.

The demand for newer technologies and newer approaches to understand the difference have come in so that is you know that is what you will see little later. But you cannot say that you know the field Bio-informatics meaning classically you know putting a data base where in information regarding biological systems are stored if you define strictly that way is not that that new, not even 1960 (())(33:33) if you talk about classification of organism which give subject people hate is very very important because that is the way you are able to relate organism, classify them.

You know that India for example as these many plants, species or animal species very very important and that is nothing but a data base wherein you are putting the data as to what are the plants exists what are the different species what is a characteristics species what is a relatedness. This again a data base and which can go back in centuries you know you can trace back but not only that even you started with domestication of the animals. Domestication of the plants you know if you go back you know even our own ancient literatures you know whether it is Ayurveda or whatever.

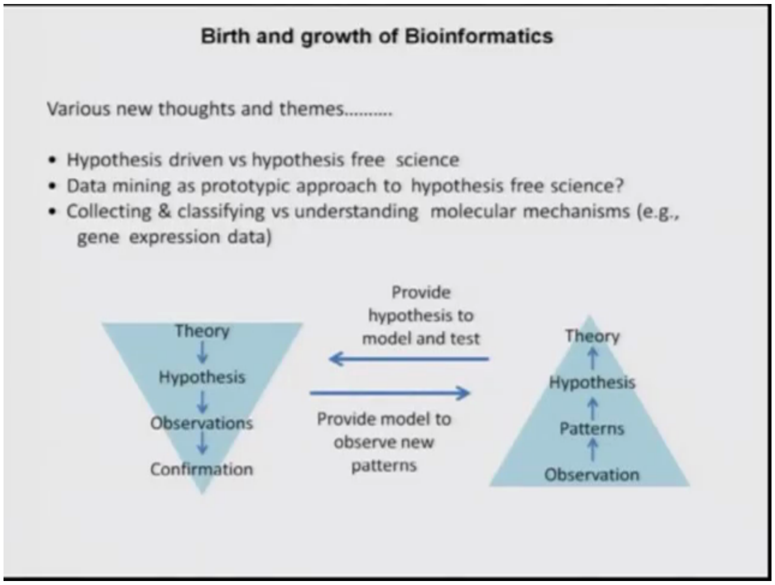
It talks about medicinal plants with certain names and within certain characteristics features these are the plants with these properties available in this part of the country can heal this-this process again it is a data base. So you know it is Bio-informatics is not something new, it exists as long as the human civilization came into place you know there is a process where in the data is stored-retrieved you know ever since the language evolved you have and there are even before that you know there are you know in since as old as human civilization itself.

So if you relate that but what has become important in recent years is the development of computers and computing tools that really changed the way the data being analysed. Earlier it was like stored and analysed by humans but now the data is so voluminous that humans cannot really analyse the sequence. So that is again you see that there are some timelines given here 1951, the first commercial computer came in that is around when the peptide sequence being sequenced and analysed.

So obviously they have used that tool IBM came up with storage device because this is important because you need to store data because the data you know size increases. The removal storage drive really held because you can put that fill it and take it another put one and that really helped us to store the data for later analysis 1962 and 81 came the work station, so they work hard most of the data analysis applications Apple and Sans Microsystems and in 1983 the Apple personal computers came and 1990 is when the internet or web servers came and as changed the way the data being analysed so anybody sitting from any part of the world can analyse the sequence using remote login.

So that was possible because of the web servers. Just to relate with the contemporary media tools that you guys use 95 is when the famous browser internet explorer was released, 2000 Wikipedia which is a free source for most of the information. 2004 is the Facebook and 2005 YouTube, so all these used data that are stored in different part of the world and to query analyse whatever you want. It is a very similar tool that being used for even Bio-informatics but much more complex data and complex tools to study them.

(Refer Slide Time: 36:28)



Birth and growth of Bioinformatics

Various new thoughts and themes..........

- Hypothesis driven vs hypothesis free science
- Data mining as prototypic approach to hypothesis free science?
- Collecting & classifying vs understanding molecular mechanisms (e.g., gene expression data)

Theory ↓ Hypothesis ↓ Observations ↓ Confirmation

Provide hypothesis to model and test

Provide model to observe new patterns

Theory ↑ Hypothesis ↑ Patterns ↑ Observation

So the Bio-informatics mainly the one that is involving the genomics information has changed the way science was done. So there are new thoughts that came in with the birth of the post genomic Bio-informatics field so earlier people used to use what is called as the hypothesis driven science now you know with a data science you can do hypothesis free science, I will talk about what is the major difference between these two and data mining as a prototype approach to hypothesis you know data really helped to start what is called as a open exploratory science without really thinking that you know you have a model you want to taste the model that is the hypothesis.

And collecting and classifying that is something that is normally done in classic mode of science and to understanding the molecular mechanism for example you know understanding how the gene expressed and so on. This could not have been done without the Bio-informatics field because the challenges much more that is something we are going to discuss. Let us see what it is so the hypothesis driven research is something that is shown on the left side ok.

So you have what is called as a theory so you have a theory saying that humans evolved from monkeys this is a theory based on observation because we have a pair of hands which are used for purpose other than walking, monkeys do the same and you have pair of legs that are dedicated for walking monkeys do. So we can say that and but we have some abilities that are beyond what monkeys can do so we would say that we are more advance, more recent as compared to monkeys therefore humans evolved from monkey, it is an hypothesis alright so you know then you strengthen hypothesis, you `do few more test and then you confirm that possible that is true.

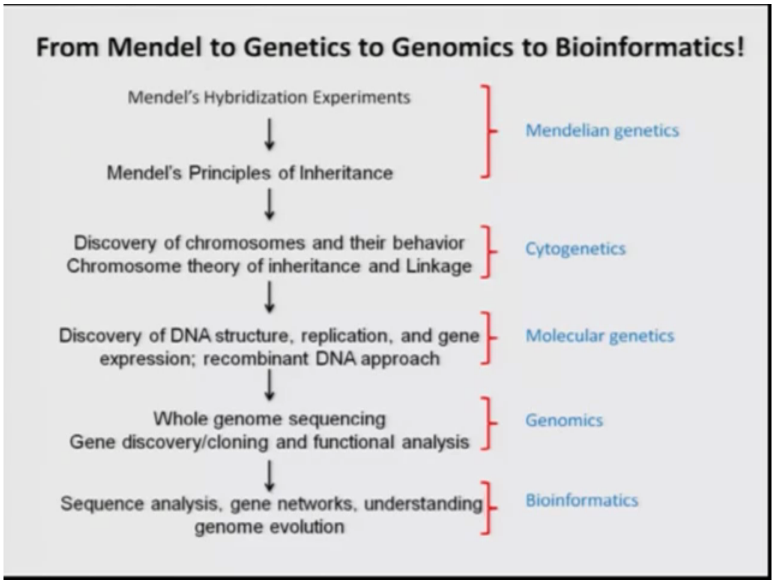
We cannot prove it but you can say that is most likely. So the way you can do is you can go back and look at human genome and monkey genome and take for example the dog genome or a snake genome and look at how similar or how different we are and you find that the monkeys are more closer to your genome sequence than say the dog or a snake then likely that we evolved from monkey. This is the way to you know prove that that is called hypothesis joint research.

Now the data science with with all the Bio-informatics tools it has changed so what you have is huge amount of data, so have one file in which you have the human, the other file you have for example monkey, the other file you have frog, you have fish genome and so on. I do not do anything I do not have no hypothesis all I do is I am going to compare all this five different files in terms of how similar the sequence are.

I do this and then my patterning we basically look at certain patterns and then its going to tell you there are more pattern that are similar between human and monkey as compared to say a monkey and dog, a monkey and snake and monkey and frog. Whereas frog has more similar patter you know with snake as compared to a human or a dog. So this kind of you know patterns led to a hypothesis now we can say that dogs evolved sorry the frogs evolved earlier than snakes and and dogs and monkeys and humans.

So you can come up with an theory now, so this is hypothesis free science that is possible because you have a data which you can mine and look at but these two are not really independent of each other. The hypothesis driven research that is conventional science provides more models to test you even using data science and new theory that is generated from the hypothesis free science using data science can give an hypothesis which again can be tested again using the conventional approach so these are not something not over lapping. They are in fact complementing each-other but there are various ways of doing research.

(Refer Slide Time: 41:38)



From Mendel to Genetics to Genomics to Bioinformatics!

Mendel's Hybridization Experiments
↓
Mendel's Principles of Inheritance
} Mendelian genetics

↓

Discovery of chromosomes and their behavior
Chromosome theory of inheritance and Linkage
} Cytogenetics

↓

Discovery of DNA structure, replication, and gene expression; recombinant DNA approach
} Molecular genetics

↓

Whole genome sequencing
Gene discovery/cloning and functional analysis
} Genomics

↓

Sequence analysis, gene networks, understanding genome evolution
} Bioinformatics

So if you look into entire discussion so far we had from Mendel to the sequence Bio-informatics how does it really there is a comparison so Mendels hybridization experiments led to theories of inheritance, dominance and segregation and then we came up with the chromosomes and linkage wherein we say there are the distinct entities. Some genes are present on the same chromosome and off course the chromosome nothing but DNA and you have their replicate make copies and you have gene sequences and the genomics sequenced and you analyse and so on.

So this is how you sort of understand the evolution so really how does it all these different steps or stages are of understanding can be categorised, so this is how we call now. So Mendels hybridization experiments and principles you call Mendelian genetics. The chromosomes behaviour during self-division either Mitosis miosis basically called as Cytogenetics and you look at DNA molecules manipulate them change them. Its molecular genetics when you sequenced the entire genome and you go for a large high through put sequencing large number of data if you are generating its normally called as genomics and the large data he would generated used computational tools to understand so that is called as Bio-informatics.

So these are manmade distinctions just to say that each you know sub domain requires different skill sets to perform and to understand and to master so what it is but it is all same assembly line in fact even people do tests like Mendel did. All the eye yielding varieties of menu that paddy, rice or wheat that you find all based on passive reading experiments so it helps you in today it is not the Mendel did and your done and forgotten and we understand how different these varieties are, what gives that you know a particular advantage to a particular variety that it gives you more yield.

We can go at the molecular level and then understand so that comes molecular genetics or genomics or Bio-informatics, so everything is integrated so you cannot leave them but this not the end of the road with every new approach every new tool and our understanding of the genome we can go and look at something beyond what people have done because we have seen that you know what you talk about molecular genetics was known to Mendel or Morgan they did not know what DNA is but we now know. So the next how you know this particular information led to emergence of a new field is something that we are going to discuss in the next class.