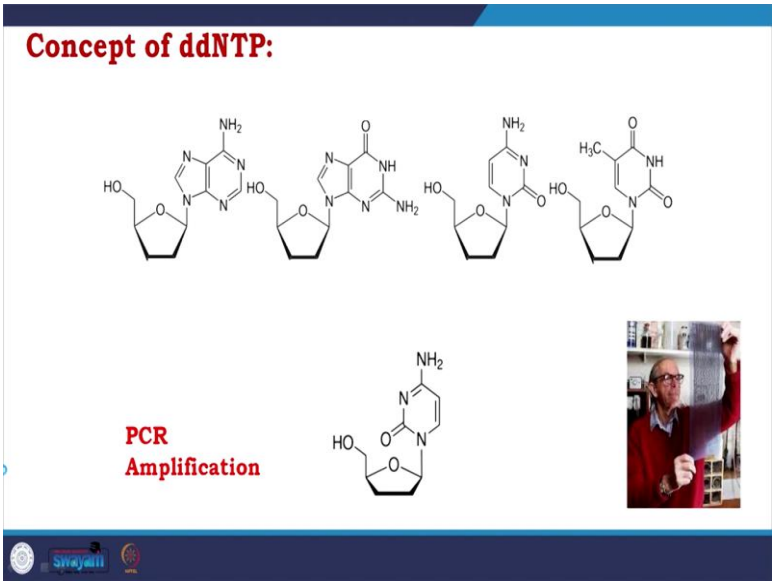
Structural Biology
Prof. Saugata Hazra
Department of Biotechnology
Indian Institute of Technology – Roorkee

Lecture 04
Introduction: Genome Sequencing

Hi, welcome again to the course on structural biology. Today we are going to class 4 and we are going to discuss genome Sequencing and the interesting phenomenon changes around that.

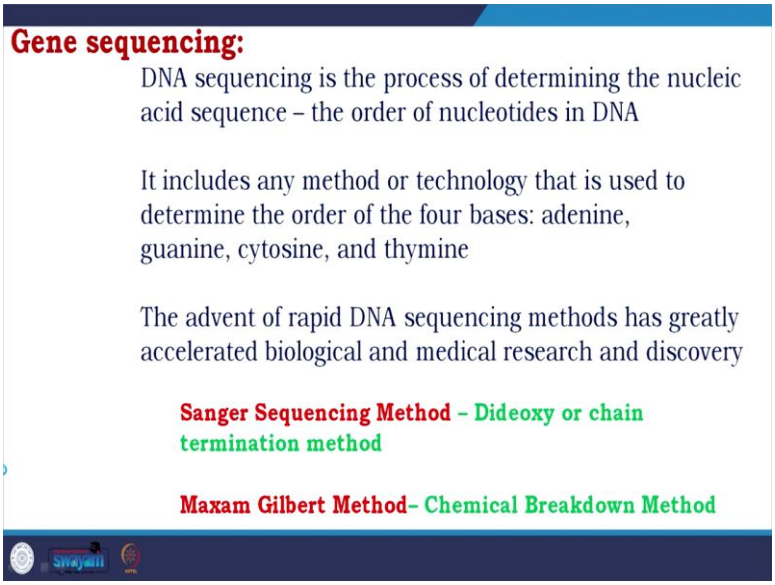
(Refer Slide Time: 00:45)



We talked about the concept of dideoxynucleotide phosphate, which helped us understand how we could go for chain termination. This, along with PCR amplification, gives us Sanger Sequencing methods. As I told earlier, have changed Biology for the first time will learn how to read the biological macromolecules in the truest sense of the term.

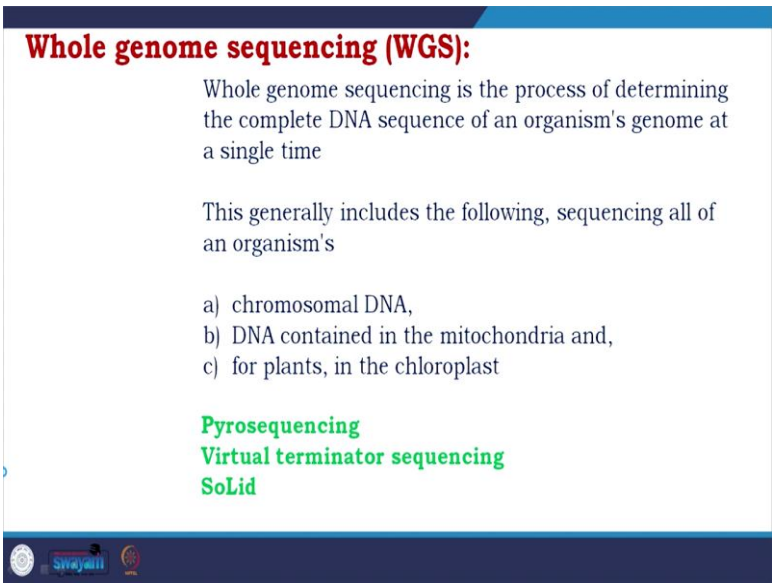
I talked about why Sanger Sequencing was revolutionary. Today, I will talk about the journey from a gene to an entire genome.

(Refer Slide Time: 01:43)



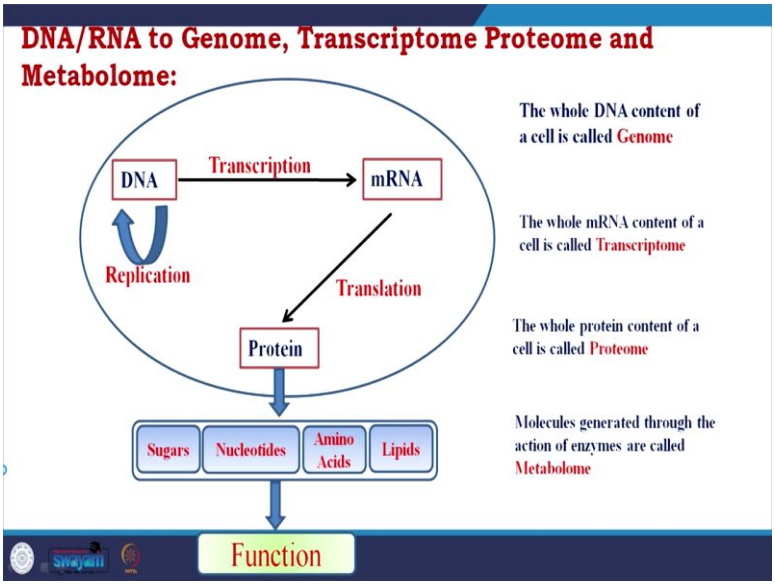
DNA Sequencing is the process of determining the nucleic acid sequence, the order of nucleotides in DNA. It includes any method of technology that is used to determine the order of the four bases ATGC. The advent of rapid DNA Sequencing methods greatly accelerated biological and medical research and discovery. There are two methods to talk about: the Sanger Sequencing method, which we talked about, and the Maxam Gilbert method, which is a Chemical breakdown method.

(Refer Slide Time: 02:39)



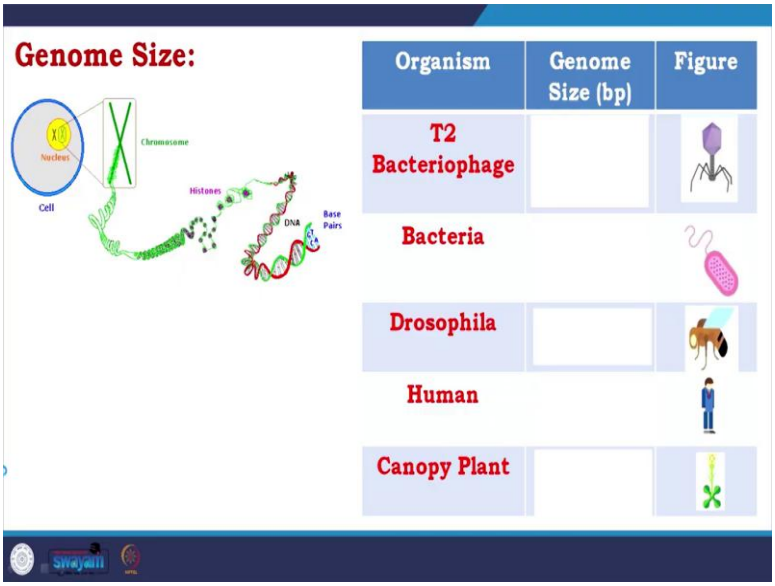
In the whole genome Sequencing, it is about determining the complete DNA sequence of an organism genome at one time. This generally includes the Sequencing of an organism's entire chromosomal DNA, DNA content in the mitochondria, and plant then in the chloroplast. There are several methods like pyrosequencing, virtual terminator sequencing, and SoLid.

(Refer Slide Time: 03:15)



I would rather come back to the central dogma. The central dogma is about DNA, RNA, especially mRNA and protein. DNA self-replicates to form DNA. DNA through transcription makes mRNA, and mRNA through translation makes protein. So these three macromolecules are interconnected. But in the body of an organism's cell, the function is done by protein majorly. Now, if you see the whole DNA content of the cell is called the genome. The whole mRNA content of the cell is called transcriptome. The whole protein content of the cell is called proteome, and the molecules generated through the action of enzymes are called the metabolome. If we could read genes and convert the ability to study genes into a genome, that means we could read life.

(Refer Slide Time: 05:07)

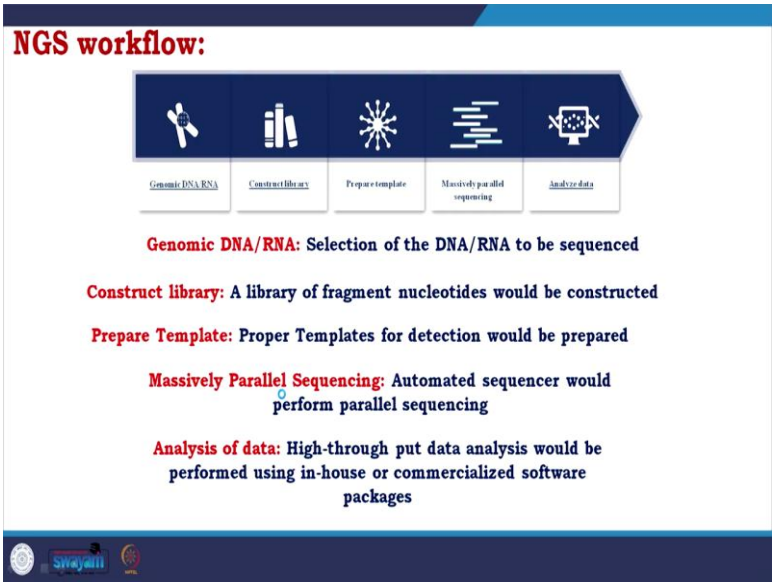


But there are challenges. If you look at the cell, there are chromosomes. And chromosomes, with the help of histones, are very closely packed, you know that there are like 23 pairs of

chromosomes in humans. If you open up the Human Genome, the distance you get, you could go to the sun and come back 60 times.

Let us look at genomes from very small organisms to plants with a bigger genome. So, first, with T2 bacteriophage, the virus and its DNA have a side of 170000 base pair. So as we were talking about in Sanger Sequencing, you could do 850 around nucleotide, which means you have to perform 200 times to sequence one of the smallest organism genome, which is T2 bacteriophage. If you look at another genome like bacteria at an average of 4.6 million base pairs. Drosophila 130 million, human 3.2 billion, Canopy plant 150 billion. So now you understand that probably a massive parallel action you cannot perform even if it is perfect or very accurate. You cannot utilize Sanger singly to convert your readability from gene to genome.

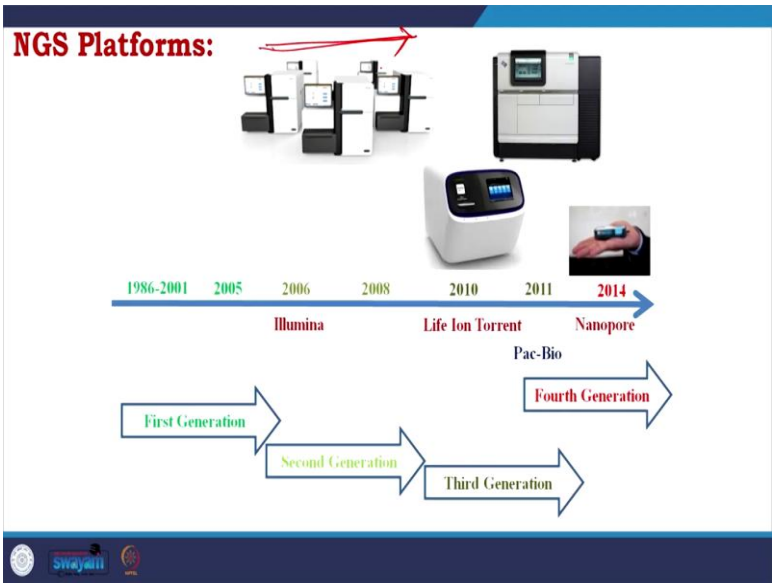
(Refer Slide Time: 08:07)



And that is why the next generation Sequencing appeared. So, what were the basic characteristics of next-generation sequencing? One the generation of many millions of short reads in parallel. As I was talking about, instead of one Sanger, it would be much better if you perform parallel 1,000 Sanger. The speedup of the sequencing process compared to the first generation. The low cost of sequencing and the sequencing output are directly detected without electrophoresis. When it is 1000 nucleotides, you could read it by looking at that, but when you are going for such a big number, you need automation and direct reading. So these are the basic characteristics of NGS. And how do NGS works? First, you have to develop the genomic library of DNA or RNA.

So, select the DNA RNA you want to sequence to get the target. Then you go for constructing a library, a library of fragments nucleotide that you have to construct. Then you go for preparing template proper template for detection. Like you are breaking the genome into several fragments, you have to reassemble them after the sequencing to develop a proper template. Then as I told you, massively parallel Sequencing with automated sequencers. In the end, you are analyzed high throughput data analysis should be performed either in-house if you have in-house software or commercialized software packages.

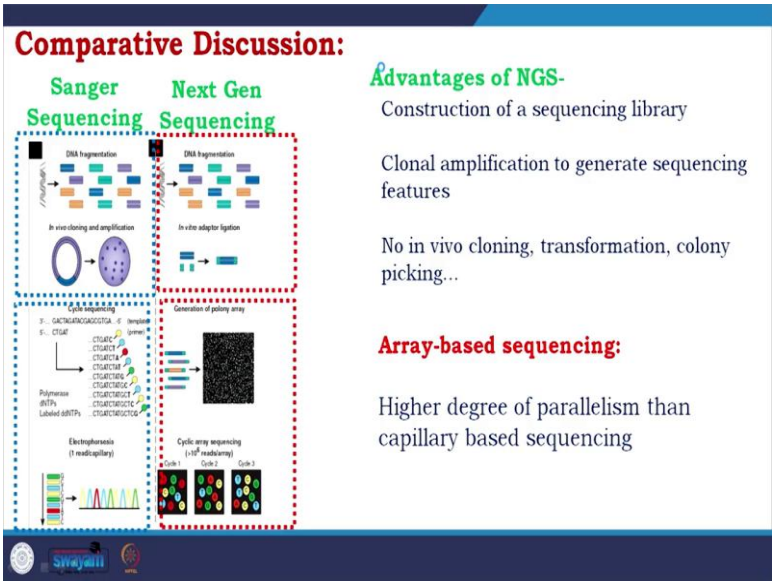
(Refer Slide Time: 10:05)



These are the NGS platforms ABI Sanger, Roche, Illumina, life solid, Life Ion Torrent, Pac-Bio, and recently Nanopore. If you see the Sanger and Roche are called the first generation. Illumina and life solid started from the second generation, Then 3rd Generation Life Ion Torrent, Pac-Bio kind of between third generation and the next generation where nanopore is used. You could look at the size of the nanopore. It is that small, and with that, Pac-Bio and Nanopore now in the Sequencing cost are getting lower day by day.

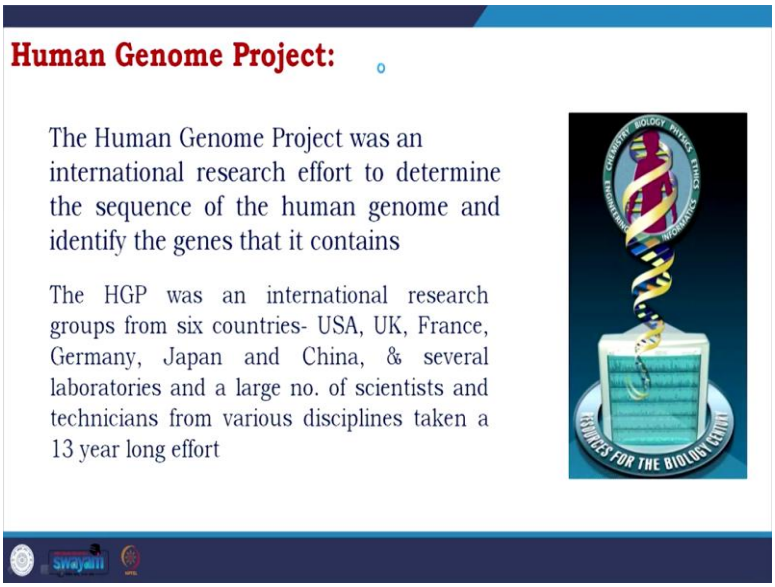
Yes, there is still the issue of accuracy. As I told Illumina, the accuracy is still best, but these guys are also coming with a higher accuracy device.

(Refer Slide Time: 11:38)



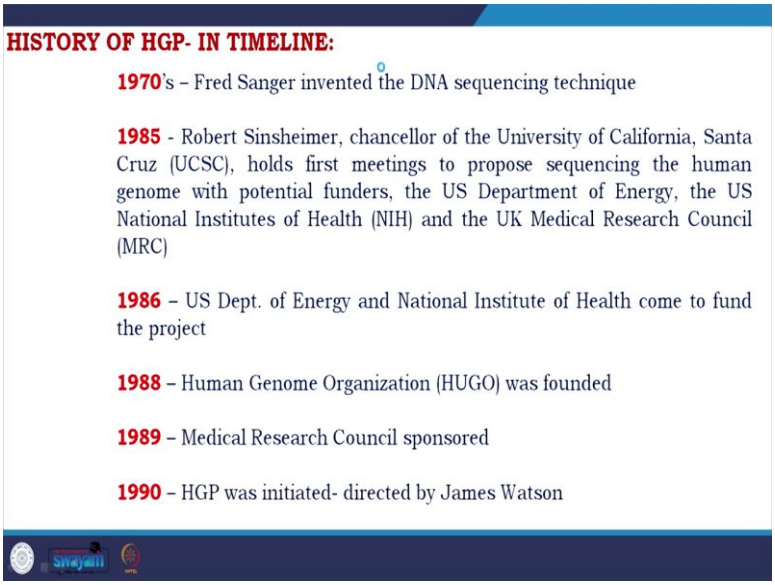
If you compare Sanger and next-generation sequencing, they have the advantage of constructing a sequencing library, Clonal amplification, to generate the Sequencing feature. In Sanger Sequencing, there is no in vivo cloning, transformation, colony picking. Then, array-based sequencing has a higher degree of parallelism than capillary-based sequencing.

(Refer Slide Time: 12:37)



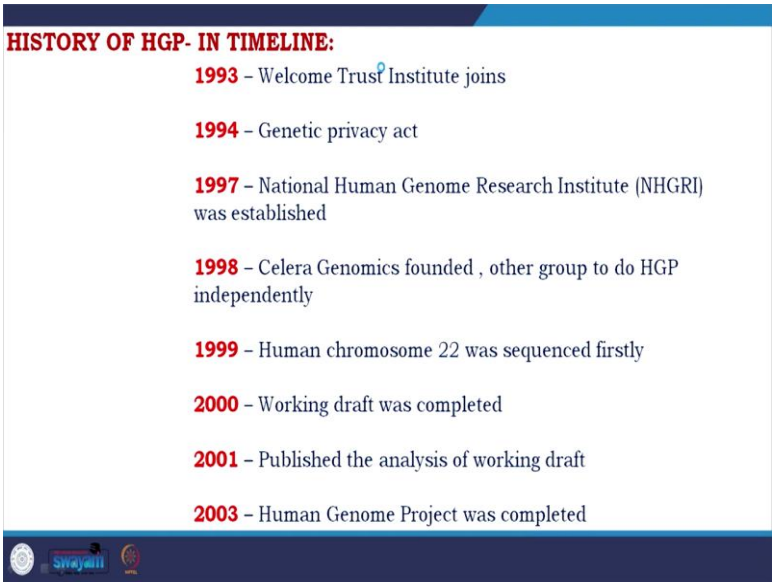
The Human Genome Project was an international research effort to determine the sequence of the Human Genome and identify the genes that it contains. The Human Genome Project was an international research group from 6 countries US, UK, France, Germany, Japan, and China, and several Laboratories and a large number of scientists and technicians from various disciplines took a 13 year-long effort to make it successful.

(Refer Slide Time: 13:05)



Let us look at the timeline of the Human Genome Project. In 1970 Fred Sanger invented the DNA Sequencing techniques we talked about that. Then in 1985, Robert Sinsheimer, chancellor of the University Of California, Santa Cruz, held the first meeting to propose Sequencing the human genomic potential funded by the US department of energy, the US National Institute of Health NIH, and UK Medical Research Council MRC. In 1986 US Department of Energy and the National Institute of Health come to fund the project. In 1988 Human Genome Organisation was founded. In 1989 Medical Research Council sponsored, in 1990 HGP was initiated and directed by James Watson, the same person who had come up with the structure of DNA.

(Refer Slide Time: 14:01)

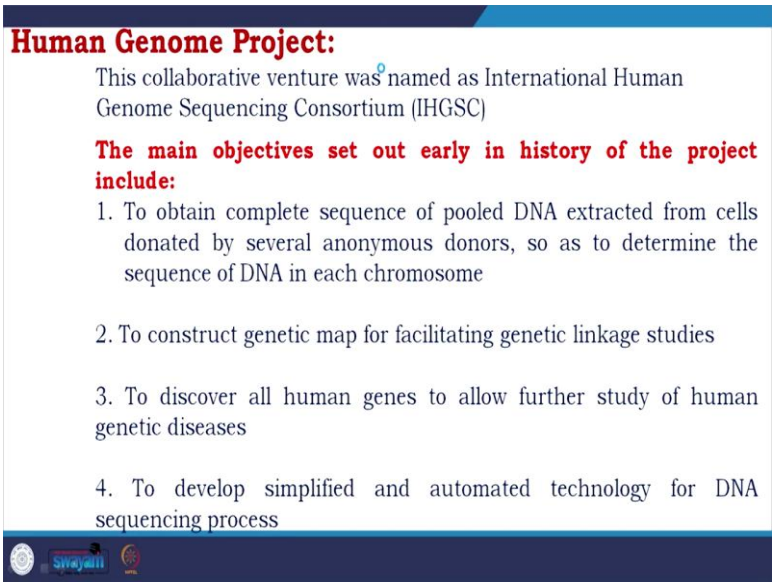


1993 Welcome Trust Institute joins. 1994 they have developed the Genetic Privacy Act because they have what about the probable misdirection when doing Human Genome Project.

In 1997 National Human Genome Research Institute NHGRI was established. In 1998 the Celera Genomic groups also did Human Genome Project independently.

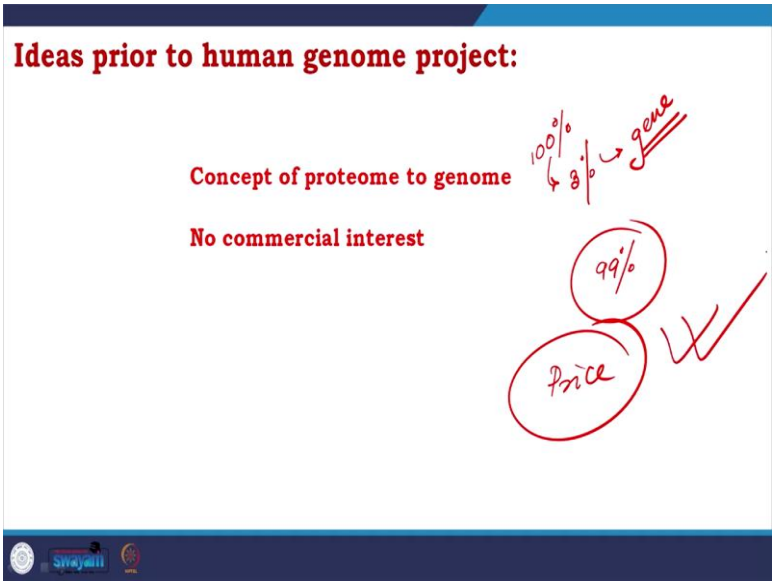
In 1999 human chromosome 22 was sequenced firstly that the first entire chromosome sequencing. 2000 working draft of the genome was completed. 2001 they published the analysis of the working draft, and in 2003 Human Genome Project was completed.

(Refer Slide Time: 14:55)



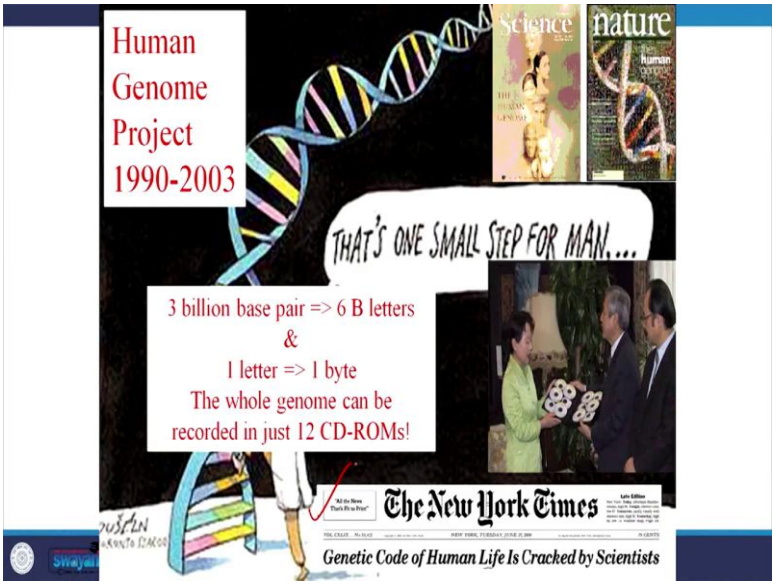
So this is a collaborative venture and named the international Human Genome Sequencing Consortium or IHGSC. The main objectives among many, I am just talking about obtaining a complete sequence of full DNA extracted from cells donated by several anonymous donors was to determine the sequence of DNA in each chromosome to construct a genetic map of facilitating genetic linkage studies. To discover all human genes to further study human genetic diseases and develop simplified and automated technology for the DNA Sequencing process.

(Refer Slide Time: 16:43)



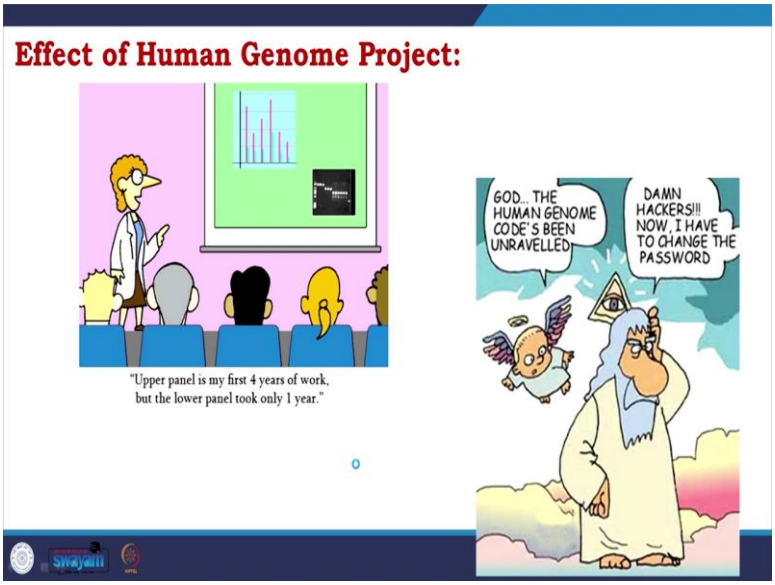
For example, humans have a 100% genome around 3% of the gene, so that concept was missing. Another thing was people have the idea that, especially between one organism, most of the genomes a pretty identical.

(Refer Slide Time: 17:55)



What happened after Human Genome Project, as I told it to take 1990 to 2003 after it was published in all the like esteemed dailies like here, I am giving you a good example of New York Time genetic code of human life is cracked by the scientist. It was regularly published in Science and nature. It was a 3 billion base pair with 6 billion letters, and the whole genome was recorded in just 12 CD-Roms. And think about how important it was, the American President was present to receive the 12 CD of the first draft of the Human Genome.

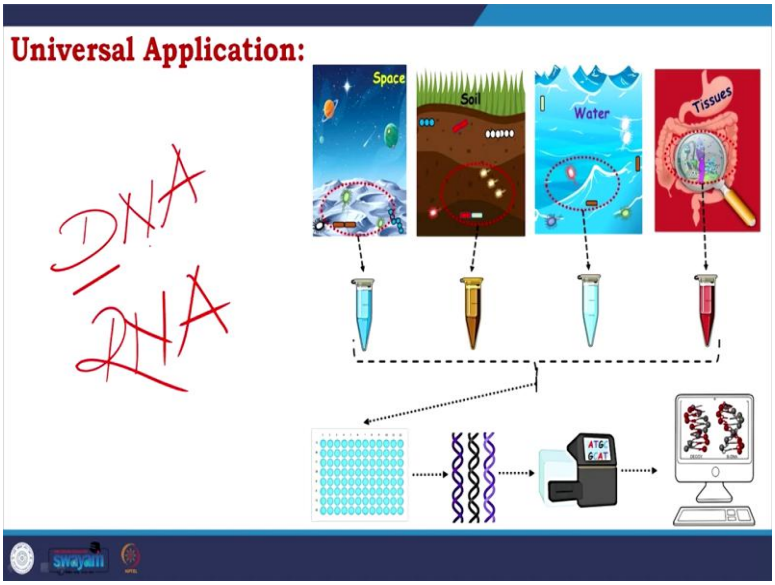
(Refer Slide Time: 19:12)



(Refer Slide Time: 20:02)

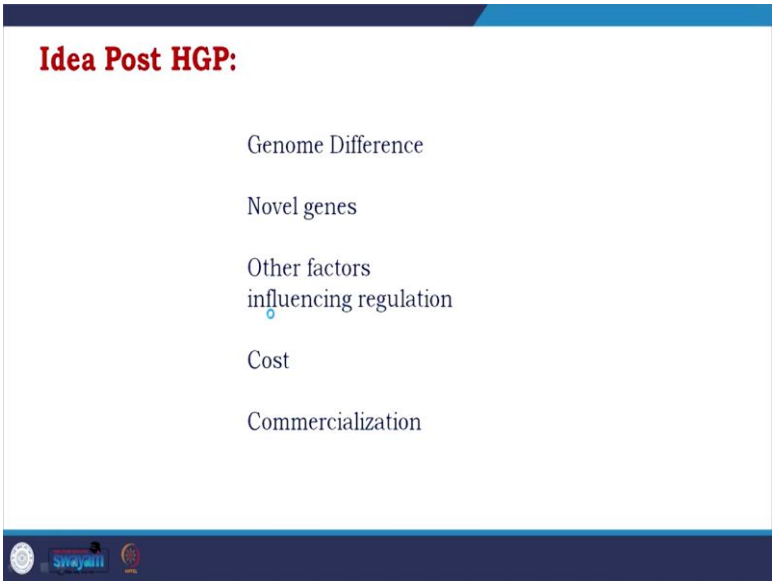
Before that, the biochemist scientist used to work on a single gene, like BRCA, ACTIN, RAS, ANSERGENE important structural genes and all. After genome sequencing, there is a development of departments, like genomics and proteomics, and they are not talking about gene talking about the effect of many genes in single chromosome to the level of prediction goes a level higher and in that time as I already told the whole world was coined so much genome, proteome, transcriptome.

(Refer Slide Time: 21:29)



It could have sequence of anything and everything in space, soil, water, and tissues. Just what you need is DNA or RNA. So you have DNA and RNA you could get information from anything and everything.

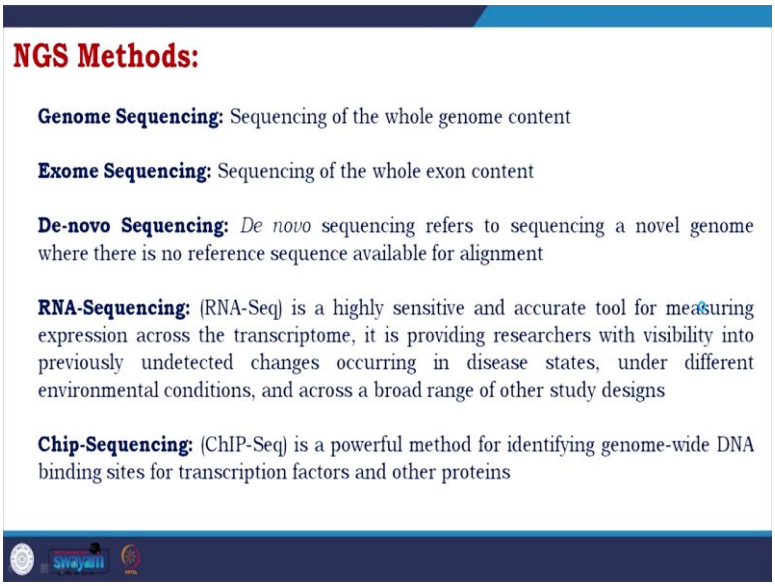
(Refer Slide Time: 21:54)



So, I talked about the Human Genome Project. After the Human Genome Project, the genome difference I told about gene and non-gene parts is clear. So now people know that the difference, predicted before between one human to another human between human to Chimpanzee, is only based on the gene part and the non-gene part is extremely different and divergent. New novel genes are coming out. Other factors include influencing the regulation of control in the replication level transcription level and translation level. Cost I have already talked about and now commercialization. So if the patients have a problem in part of the tissue that takes cells from there and from the healthy part, then do genome sequencing and

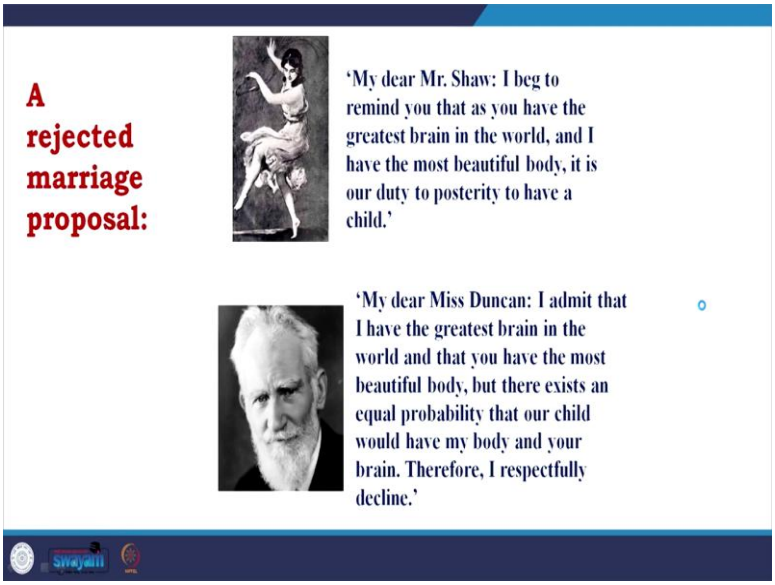
compare, and that comparative genomics gives them the real cause. So this is changing the concept of biotechnology, biomedical science, and many others.

(Refer Slide Time: 24:41)



And standing here now, I ask whether NGS is a solution to everything. So if you look at the NGS method applies to genome sequencing, sequencing the whole genome content. Exome Sequencing, sequencing the whole exon part. De novo Sequencing, de novo sequencing refers to the sequencing of the novel gene where there is no reference sequence available for alignment that is also possible now. RNA Sequencing is a highly sensitive and accurate tool for measuring expression across the transcriptome. It provides researchers with visibility into previously undetected changes occurring in disease states under different environmental conditions and across a broad range of other study designs. Chip Sequencing is a powerful method for identifying genome-wide DNA binding sites for transcription factors and other proteins, which will give you the regulation. It is the complete picture of the regulation through the transcriptional factor.

(Refer Slide Time: 26:16)



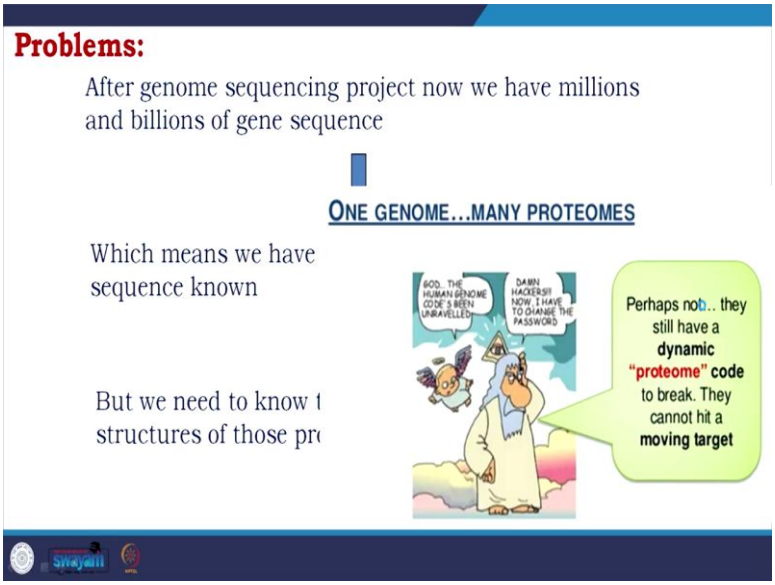
There was in the medieval age. There was a beautiful dancer, actress Miss Duncan who once have sent a marriage proposal to that time considered one of the most intellectual person novelist actor-director George Bernard Shaw.

She wrote a letter to my dear Mr. Shaw. I beg to remind you that you have the greatest brain in the world, and I have the most beautiful body. It is our duty to posterity to have a child together. In reply to that, Bernard Shaw replied, " My dear miss Duncan, I admit that I have the greatest brain not being very polite in the world and that you have the most beautiful body. But there is an equal probability that child have my body and your brain, so I respectfully decline.

So nobody knows about the truthfulness of the story, but that tells about nature's unpredictability. And the unpredictability of nature starts with the process where we cannot say NGS would be successful in every situation, first thing. When we have a lot of results now, we have a lot of data now, but we do not have the standard to compare in many cases.

So we have seen a change, but you do not understand why these changes are happening. And that is one of the reasons we have to go for a further level, which is the protein, and go for the assemblies. The molecular messenger is to form and develop.

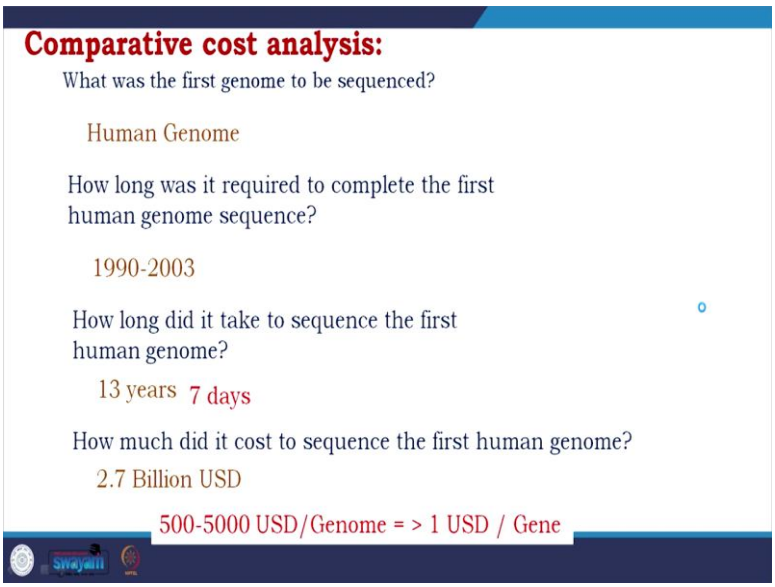
(Refer Slide Time: 28:55)



So problems after the Genome Sequencing project, we have billions and billions of gene sequences, which means we have that many protein sequences known. But we need to know the three-dimensional structure to understand its function

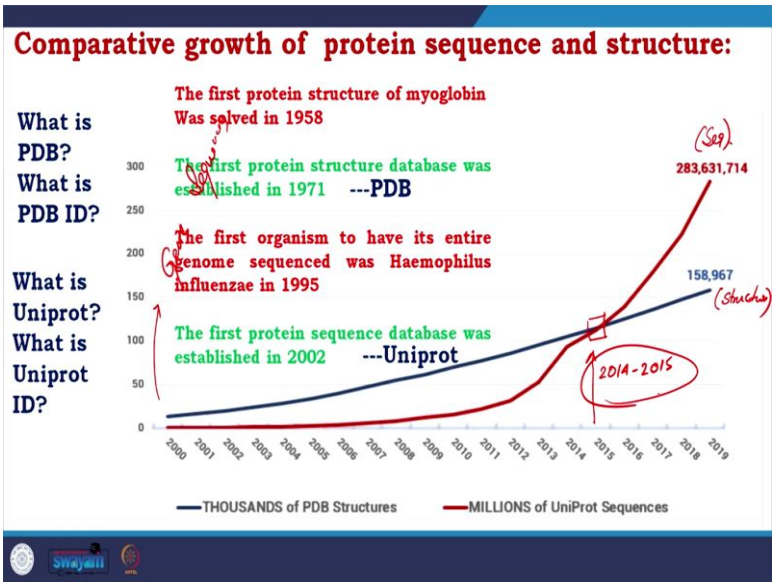
We cannot hit a moving target. So NGS gives us a lot of opportunities. Open a lot of areas we can now read the sequence. But our journey starts from here, from sequence to structure and in the structure assemblies, which are you to be done.

(Refer Slide Time: 29:56)



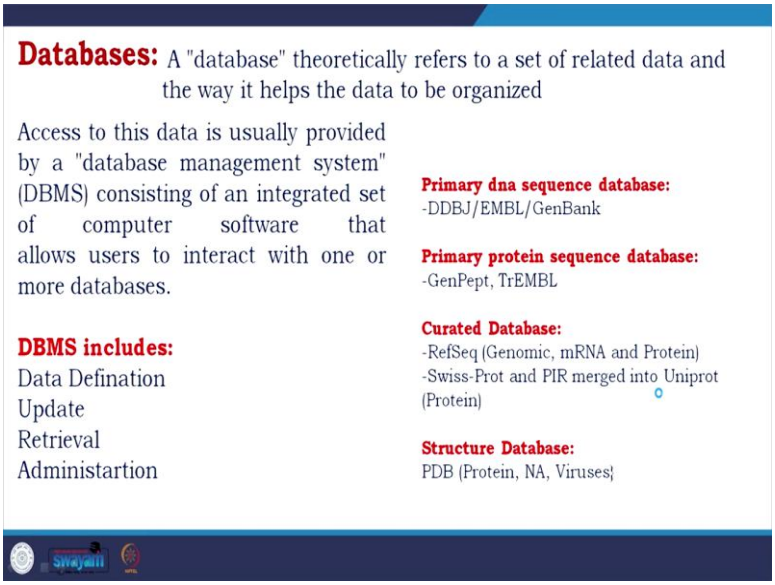
So, if we compared the cost analysis, what was the first genome to be sequenced? First Human Genome Project: How long was it required to complete the first Human Genome sequence 13 years 1990 to 2003. How much did it cost to seek the first Human Genome, 2.7 billion US dollars? Now it takes seven days in less than seven days, 500 to 5000 USD per genome, which means one USD per gene and less than that.

(Refer Slide Time: 31:11)



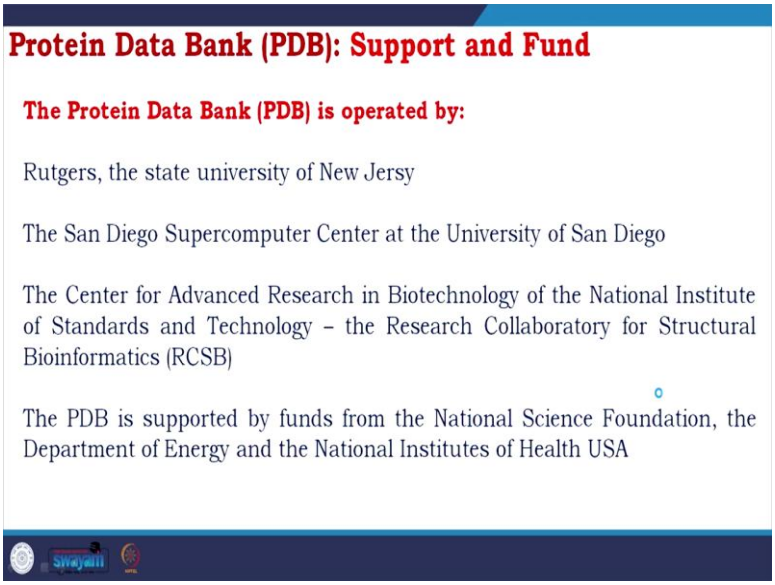
But if you could think about solving the protein structure, it still costs 3 to 5000 USD. Why are structural Biology techniques so expensive? Also, if you see here instead of 7 days, it takes two to three years. Now, look at the comparative data, growth of a number of protein sequences to protein structure. If you see initially, the sequence was much lower when there is the time of gene sequencing the automation was not there. And gradually, the number of sequences is enhanced, and this is this point which is marked around 2014-15. Sequence pass and how it passes? It goes on and on and on. The first protein structure of myoglobin was solved in 1958. The first protein structure database, PDB, was established in 1971. The first organism to have its entire genome sequence was Haemophilus influenza in 1995. The first protein sequence database was established in 2002. So I talked about this is PDB, and this is Uniprot. I will talk about this database. This database is about the entire big subject but through a few slides. I want to take a glimpse of that so that you can understand.

(Refer Slide Time: 35:25)



What is a database? A database theoretically refers to a set of related data and how it helps organize the data. Access to this data is usually provided by a database management system or DBMS consisting of an integrated state of computer software that allows users to interact with one or more databases. So a database management system or DBMS includes data definition.Few examples of databases, primary DNA sequence database, DDBJ, EMBL GenBank, Primary protein sequence database GenPept. TrEMBL, Curated databases like RefSeq for genomic mRNA and protein data, Swiss Prot and PIR merged into Uniprot, and structure database PDB where protein, nucleic acid, and viruses are there.

(Refer Slide Time: 36:48)

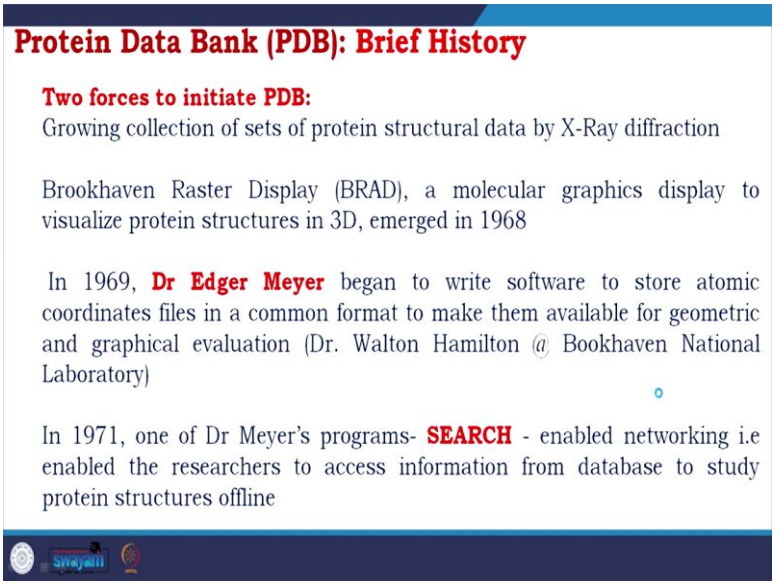


Protein data bank: Single worldwide database for experimentally determined structures. All the experimental solved structures must be deposited in protein data bank or PDB. A key resource in the area of structural Biology stores 3D structural data of biological

macromolecules like protein, nucleic acid, and viruses. Data is submitted by structural biology, biophysicist and biochemist. It is freely accessible on the internet and is updated quickly. The mission is to maintain a single protein data bank archive of macromolecular structural data. The database is operated by Rutgers, the State University of New Jersey, the San Diego supercomputer center at the University of Santiago, the center for advanced research in biotechnology of the National Institute of Standards and Technology, the research collaboratory for structural Bioinformatics which is called RCSB.

The PDB protein data bank is supported by funds from National Science Foundation, the department of energy, and the National Institute of Health US.

(Refer Slide Time: 38:24)



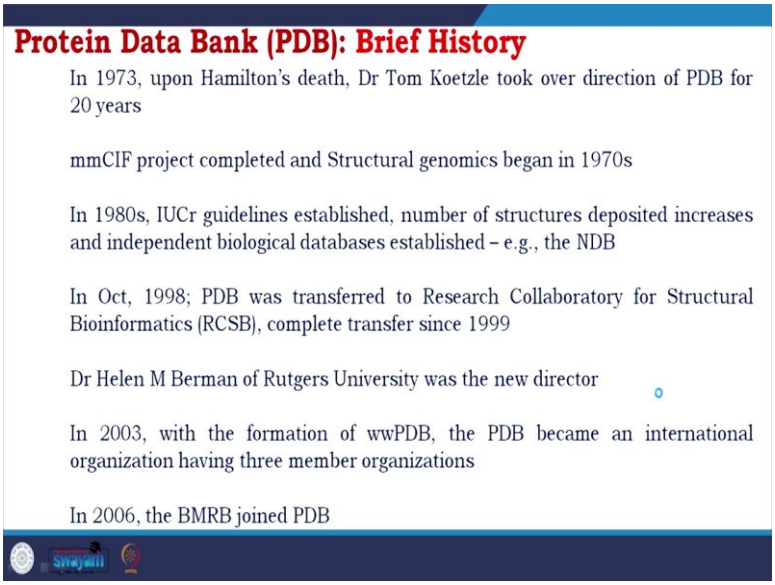
There are two forces to initiate PDB: the growing collection of state of protein structural data by x-ray diffraction. So they thought that it would be good if we had a database where the structures would be deposited, and everyone could access the data.

Brookhaven Raster display or BRAD, a molecular graphics display to visualise protein structures in 3D, emerged in 1968.

In 1969 doctor Edger Meyer begin to write software to store atomic coordinate files in a common format to make them available for geometric and graphical evaluation. He was working with Walton Hamilton at Brookhaven National laboratory.

In 1971 one of Doctor Meyers's programs SEARCH, enabled networking that enabled the researcher to access information from the database to study protein structures online.

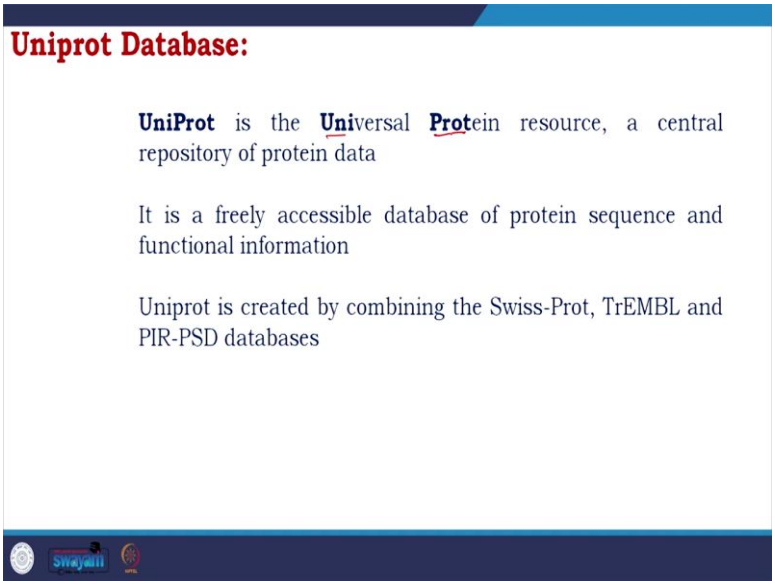
(Refer Slide Time: 40:27)



In 1973 upon Hamilton's death, doctor Tom Koetzle took over the direction of PDB for 20 years. MMCIF project was completed, and structural genomics began in 1970. In 1980, IUCr (International Union of Crystallography research) guidelines were established, the number of structures deposited increased, and independent biological databases were established.

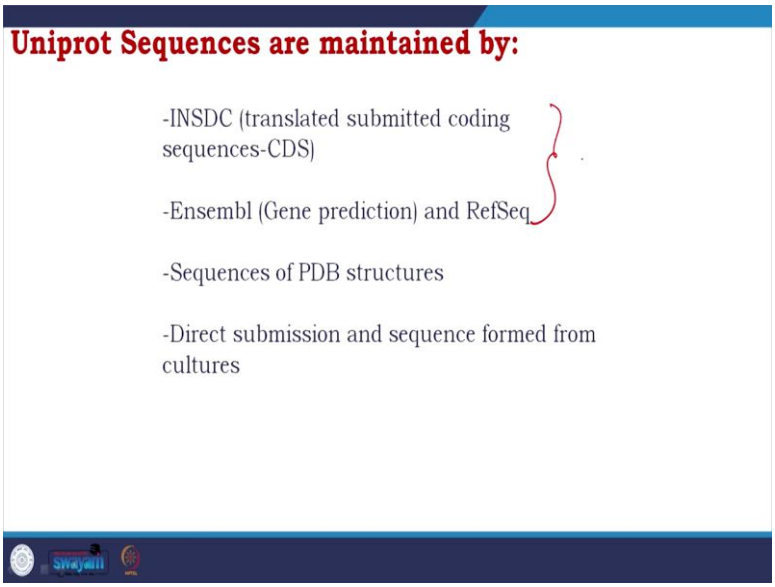
In October 1998, PDB was transferred to research collaboratory for structural Bioinformatics RCSB, and the complete transfer happened in 1999. Doctor Helen M Berman of Rutgers University was now the new director after TomKoetzle. In 2003 with the formation of wwPDB, the PDB became an international organization, having three member organizations. BMRB, which is the like depositary of NMR data, joined PDB in 2006.

(Refer Slide Time: 41:44)



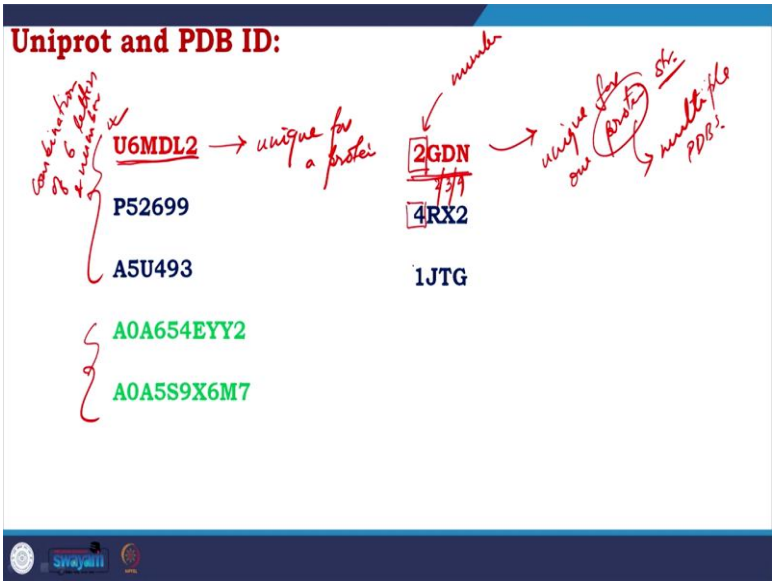
Uniprot database: Uniprot is the universal protein resource, a central repository of protein data. It is a freely accessible database of protein sequence and functional information. Uniprot is created by combining three Swissprot, TrEMBL, and PIR PSD protein information resources databases. It is updated every four weeks.

(Refer Slide Time: 42:25)



So uniprot sequences are maintained by INSDC where translated data is submitted with CDS coding sequences. Ensemble gene prediction and RefSeq, sequences from PDB structures and direct submission and sequence formed from cultures like new sequence.

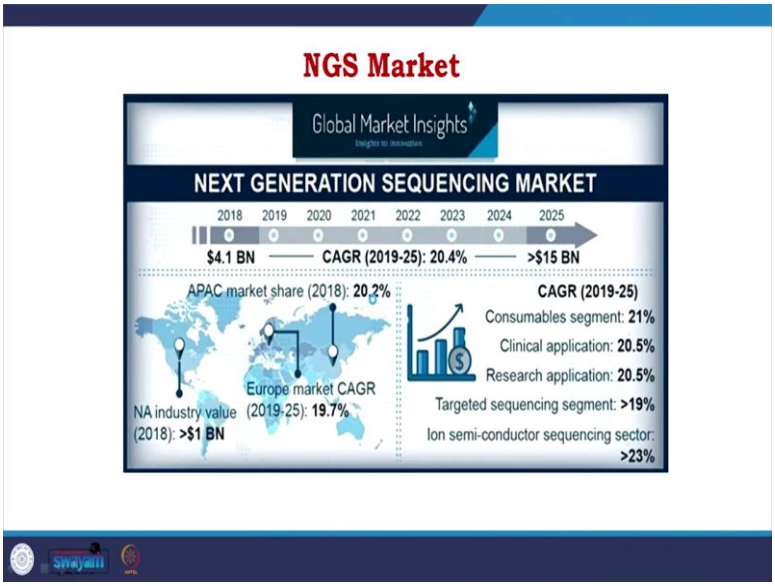
(Refer Slide Time: 43:01)



Now, coming to uniprot ID and PDB ID, they are very important for further working in the world of protein sequence and protein structure. Uniprot Id is a combination of 6 letters and numbers. So you could see this is Uniprot ID is unique for a protein. So you take this Uniprot ID, you go to uniprot.org and put this ID, and you get the information about one exclusive protein.

PDB ID is unique for one protein structure but remembers that protein might have multiple PDB's. Interestingly these numbers have one indication. Higher the number, the more recent the structure is. Again, you go to PDB, provide a PDB ID, and get the unique structure and information.

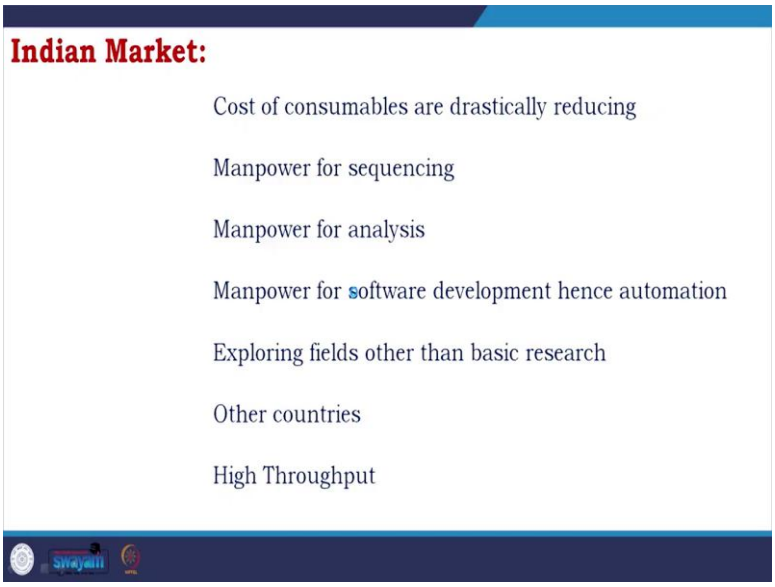
(Refer Slide Time: 46:48)



So the NGS market is one thing to know about. If you see the market is increasing, it will be 15 billion in 2025. More importantly, the market share of the Asian countries in 2018 was

around 20% and personal being associated with some of the startups and some of the inspiring startups.

(Refer Slide Time: 47:43)



The reasons are the cost of consumables at drastically reducing now. So now it depends on the manpower for sequencing, which makes us stronger to explore because the workforce cost is lower in our country. The manpower for analysis is already standing on the Boom of the software having many Software Experts in our hands. Exploring fields other than basic research: So mostly sequencing was continued or explored by people like us for basic research, but gradually it is shifting towards other applications as I told doctors are regularly prescribing now.