Bioinformatics

Prepared by: Nojood Altwaijry

Office: Building 5, 3rd floor, 197

Grades distribution

- Midterm 1: 15 marks
- Midterm 2: 15 marks
- Lab: 30 marks
- Final: 40 marks

References

- Bioinformatics and functional Genomics

Exam dates

- Midterm 1: Sunday 5th February (30 Jumada' I)
- Midterm 2: Sunday 6th March (29 Jumada' II)
- From 12-1 pm

General Objectives

1) Understand the meaning of bioinformatics.
2) How to be a confidant user of the available tools and websites related.

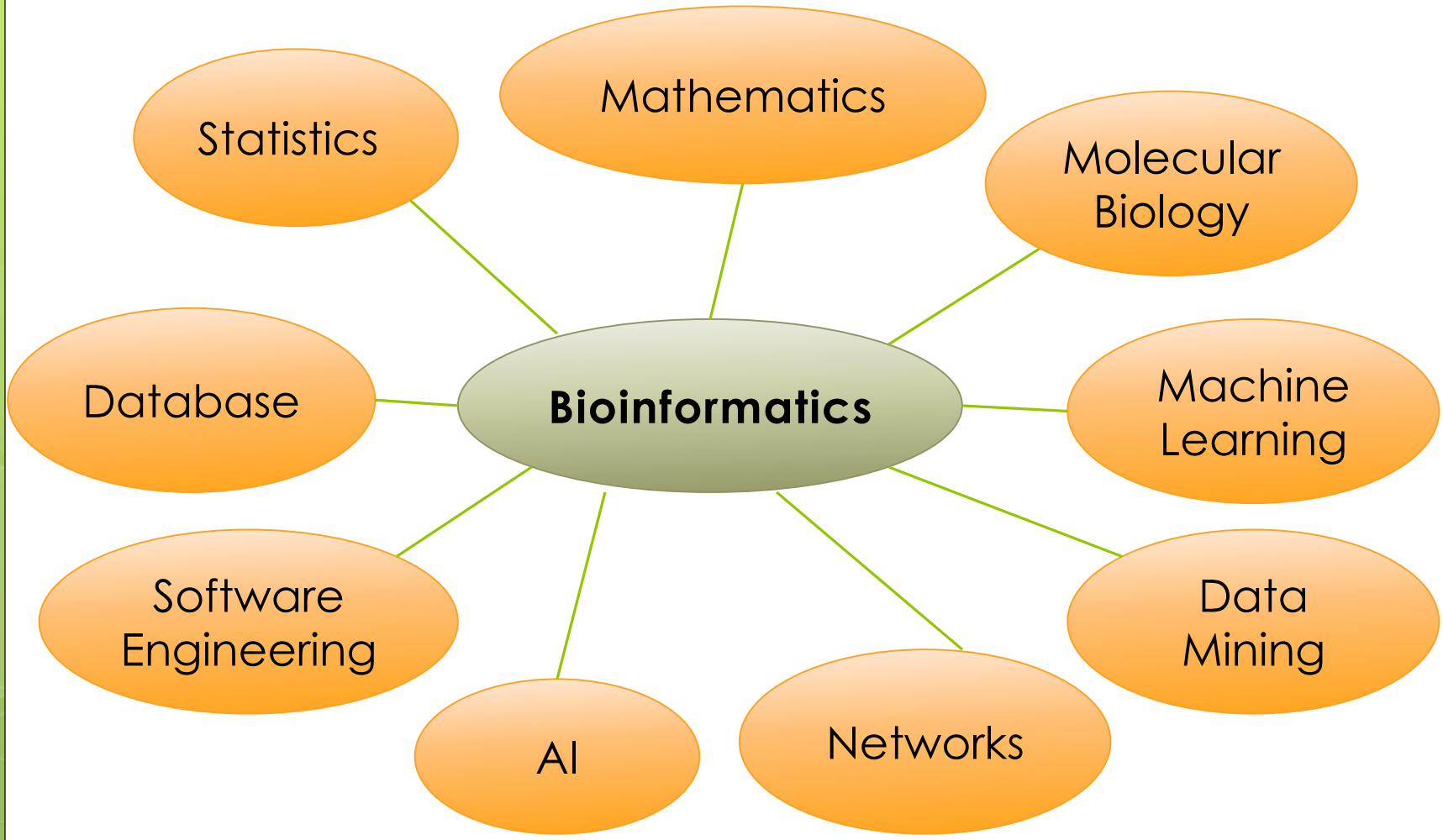
What is Bioinformatics?

- It is a computational approach to solve a biological problem.
  - *i.e. the use of computers for the acquisition, management and analysis of biological information*

What is Bioinformatics cont'ed?

- It is a multi-disciplinary field including: computer systems management networking, database design, computer programming, molecular biology.
- It draws upon the strengths of **computer sciences**, **mathematics**, and **information technology** to determine and *analyze* genetic information.

Bioinformatics and Other Fields



Insilico areas of bioinformatics

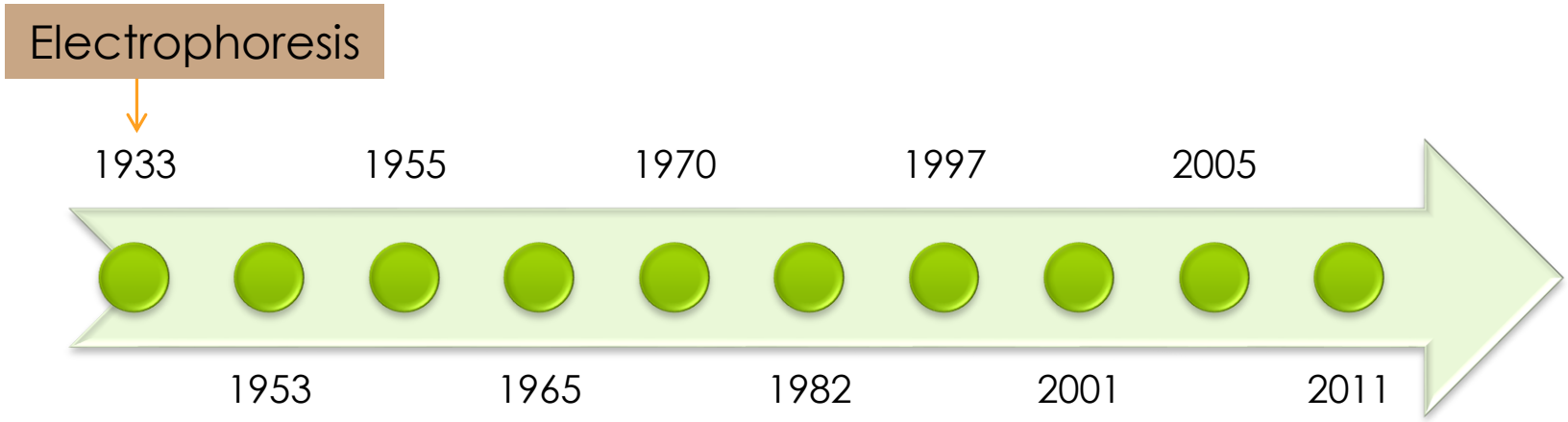
Computational biology

- Comparative homology modeling
- Phylogenetic analysis
- Protein structure prediction
- Protein folding prediction
- Micro array analysis
- Docking approaches &new drug discovery

Applications of Bioinformatics

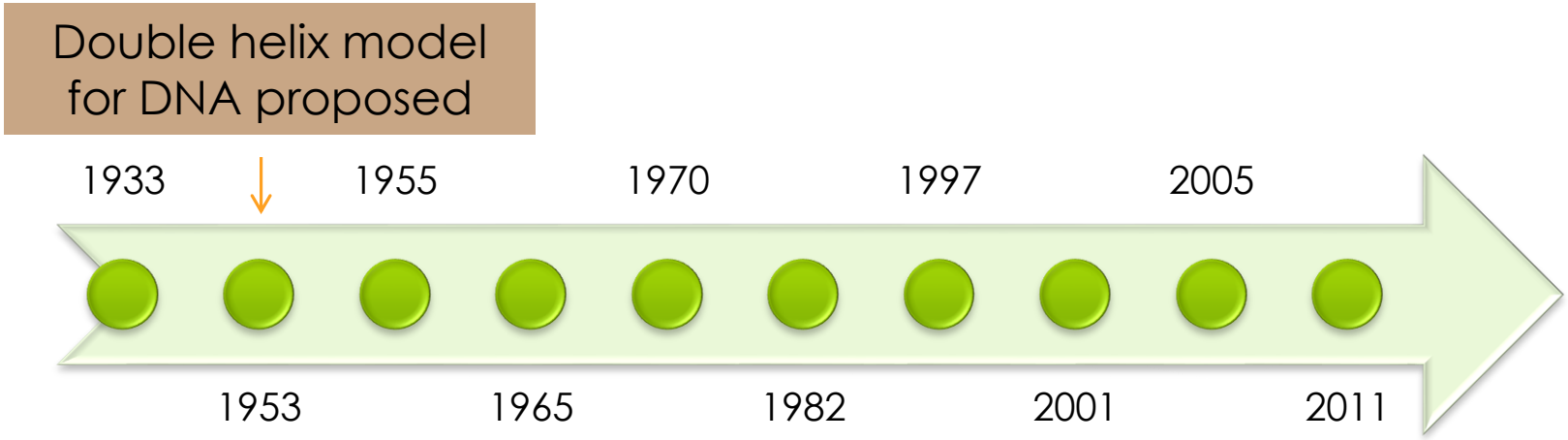
- Molecular, personal and preventative medicine.
- Gene therapy.
- Drug development.
- Microbial genome applications.
- Climate change studies.

Short History



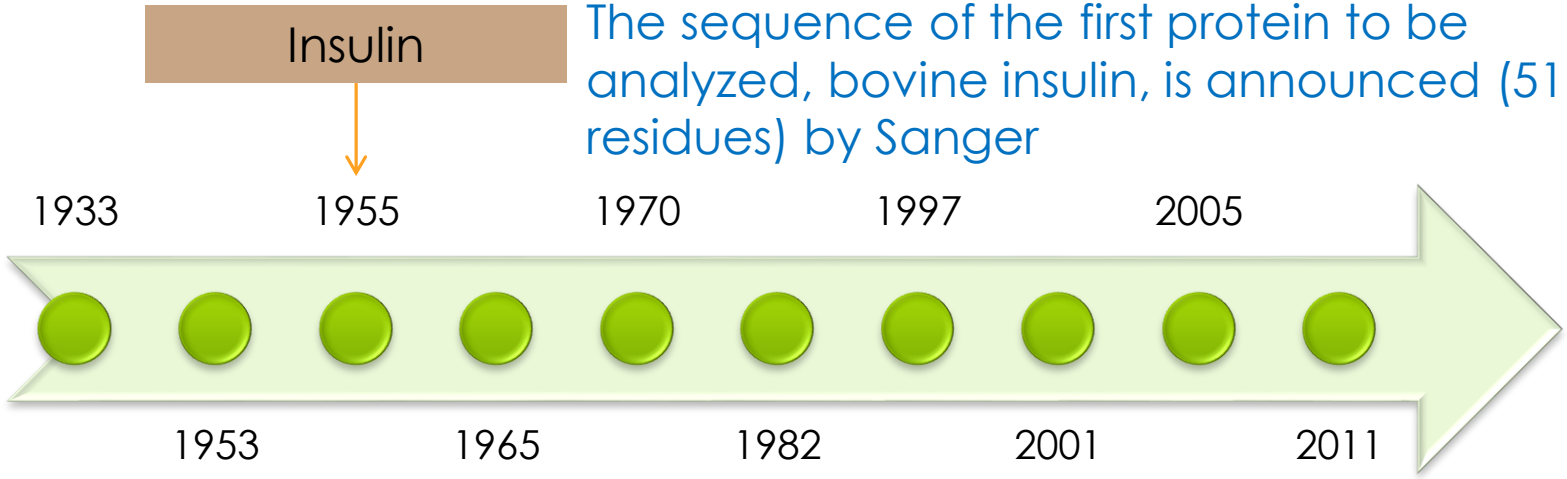
A new technique, **electrophoresis**, is introduced by Tiselius for separating proteins in solution.

Short History

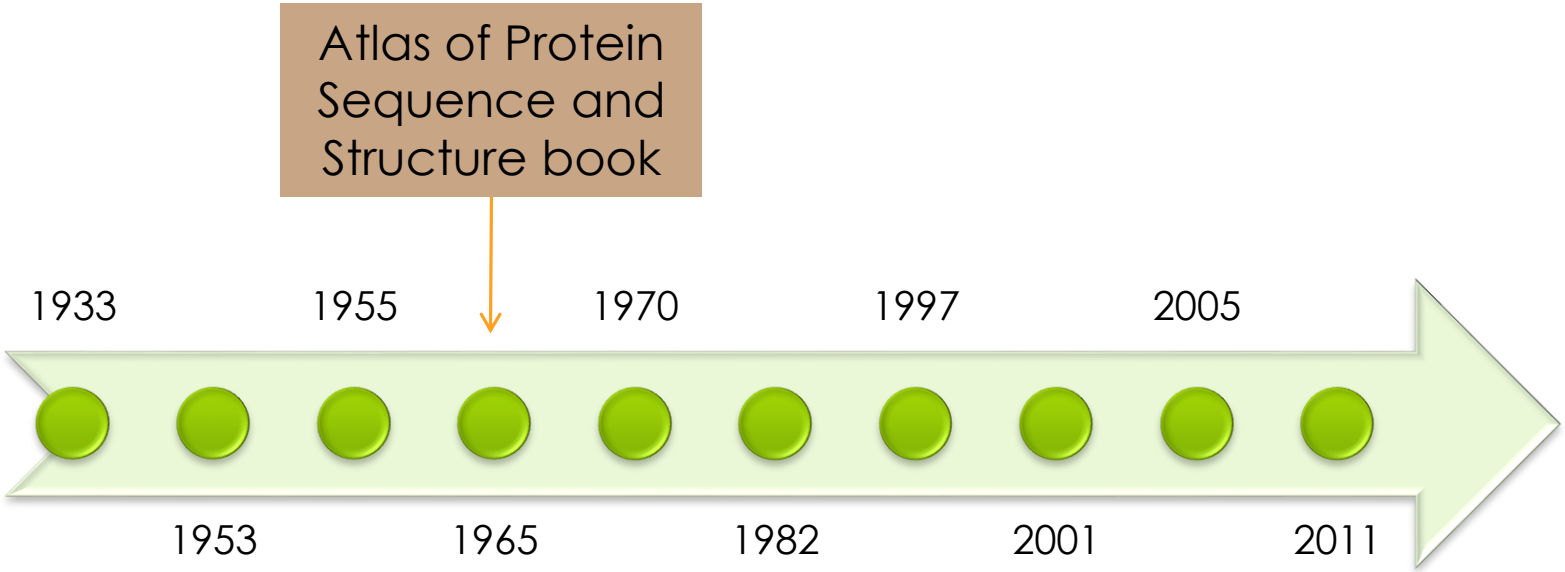


Watson and Crick propose the double helix model for DNA based on x-ray data obtained by Franklin and Wilkins (*Nature*, **171**: 737-738, 1953).

Short History



Short History



1965 **Margaret Oakley Dayhoff (**mother and father of bioinformatics) the Atlas of Protein Sequence and Structure, a book collecting all known protein sequences+ pioneered the use of computer capabilities to determine amino acid sequences of protein molecules.

In the mid-1970s, it would take a laboratory at least two months to sequence 150 nucleotides.

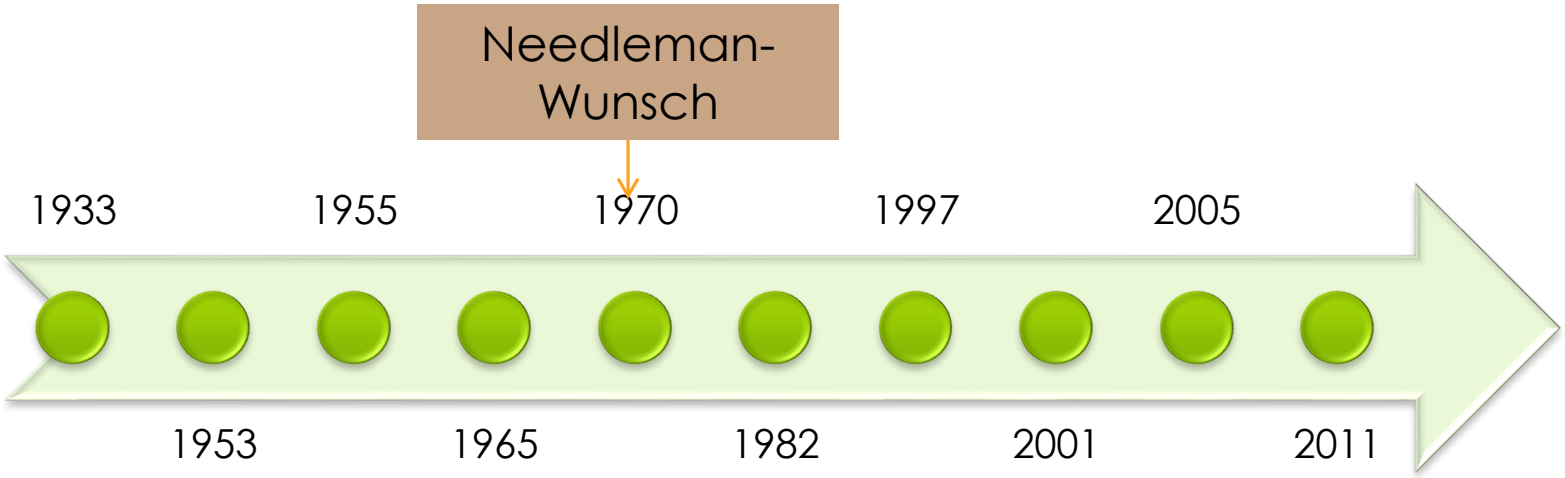
Margaret Dayhoff

"There is a tremendous amount of information regarding the evolutionary history and biochemical function implicit in each sequence and ***the number of known sequences is growing explosively****. We feel it is important to collect this significant information, correlate it into a unified whole and interpret it"*

M.O.Dayhoff to C.Berkley, February 27, 1967

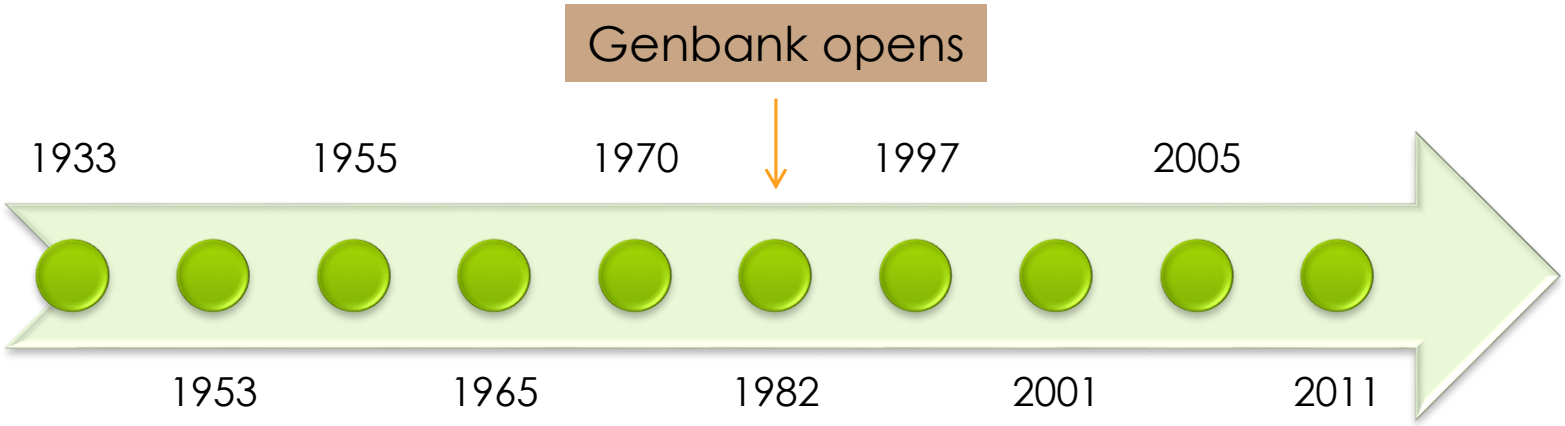
T. K. Attwood, A. Gisel, N.-E. Eriksson, and E. Bongcam-Rudloff, "Concepts, Historical Milestones and the Central Place of Bioinformatics in Modern Biology: A European Perspective," in Bioinformatics - Trends and Methodologies, M. A. Mahdavi, Ed. InTech, 2011.

A Short History

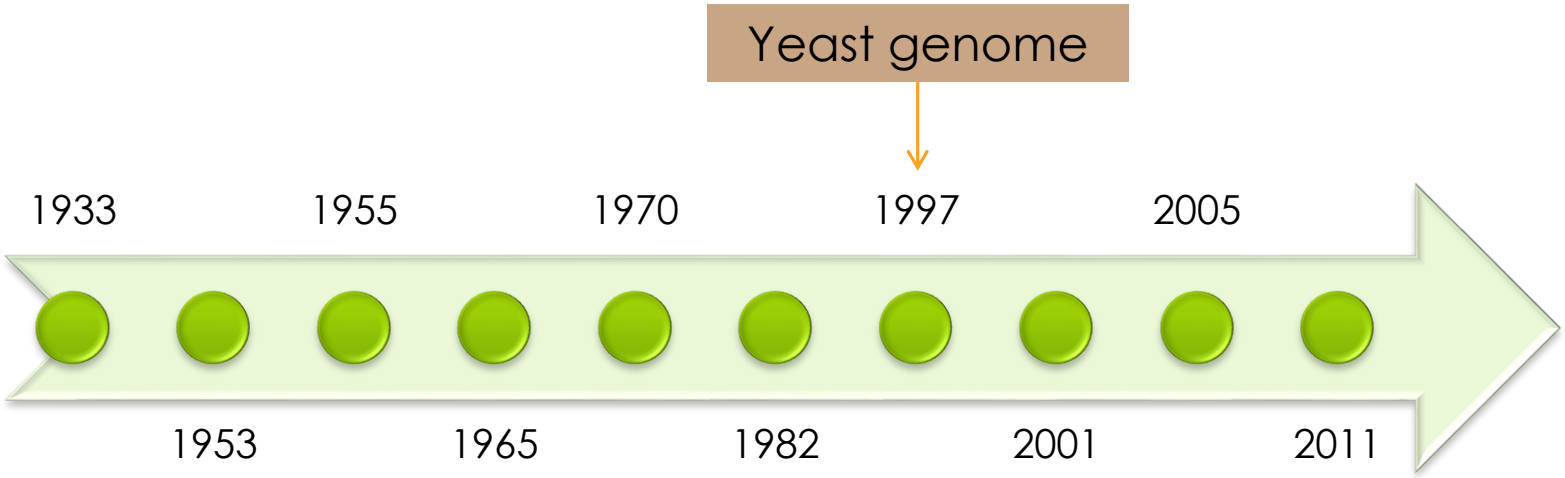


1970 the Needleman-Wunsch algorithm for sequence comparison are published.

A Short History

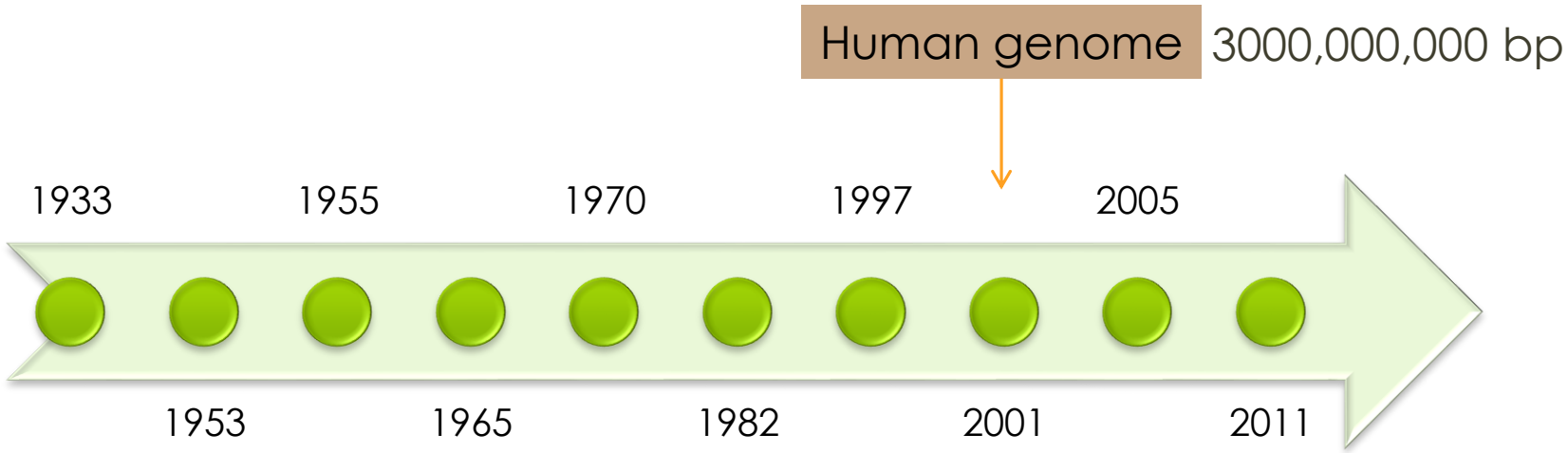


A Short History

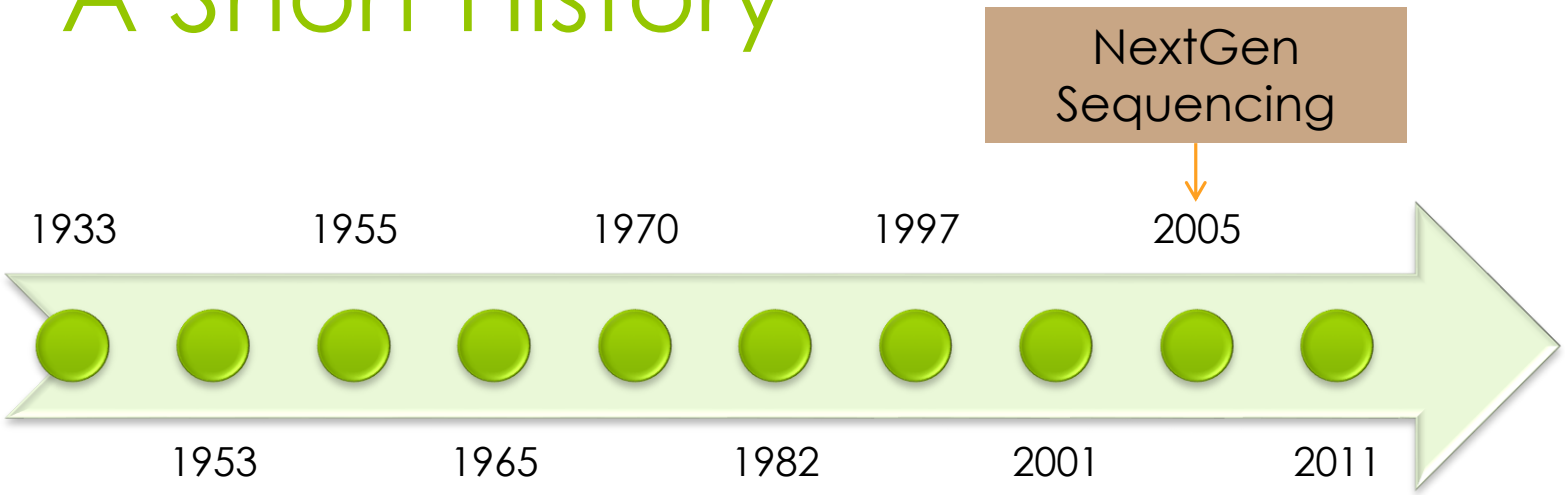


1997 The genome for *E. coli* (4.7 Mbp) is published.
4,700,000

A Short History



A Short History

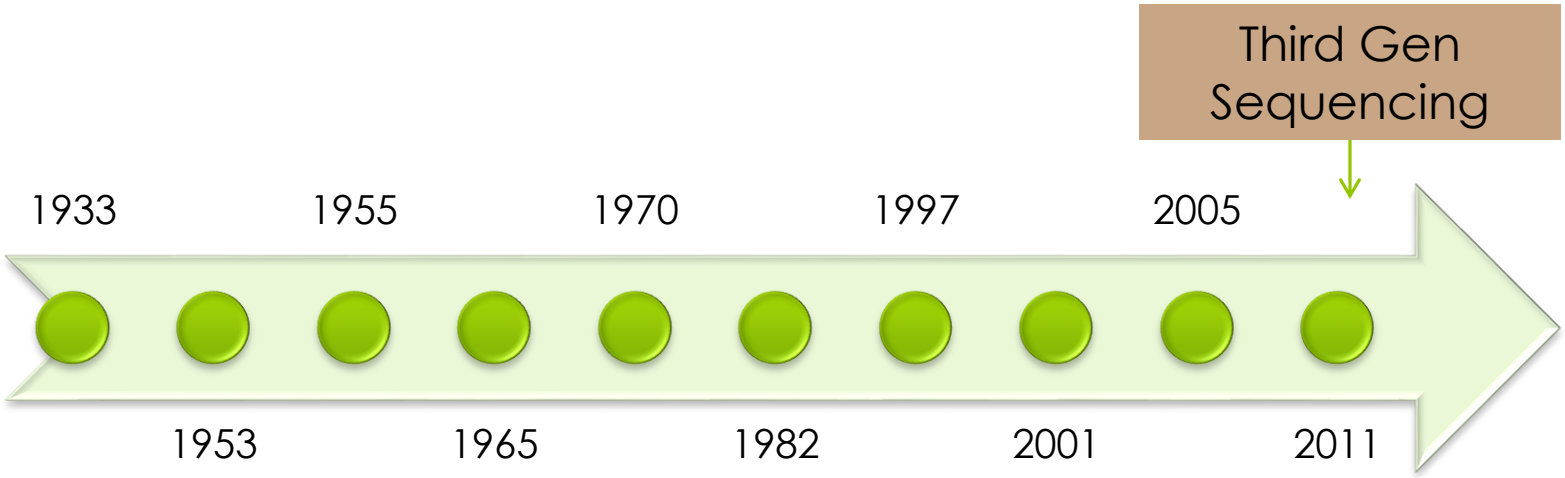


Next-generation sequencing (NGS), also known as high-throughput sequencing, is the catch-all term used to describe a number of different modern sequencing technologies including:

- Illumina (Solexa) sequencing
- Roche 454 sequencing
- Ion torrent: Proton / PGM sequencing
- SOLiD sequencing

These recent technologies allow us to sequence DNA and RNA much more quickly and cheaply than the previously used Sanger sequencing, and as such have revolutionised the study of genomics and molecular biology.

A Short History

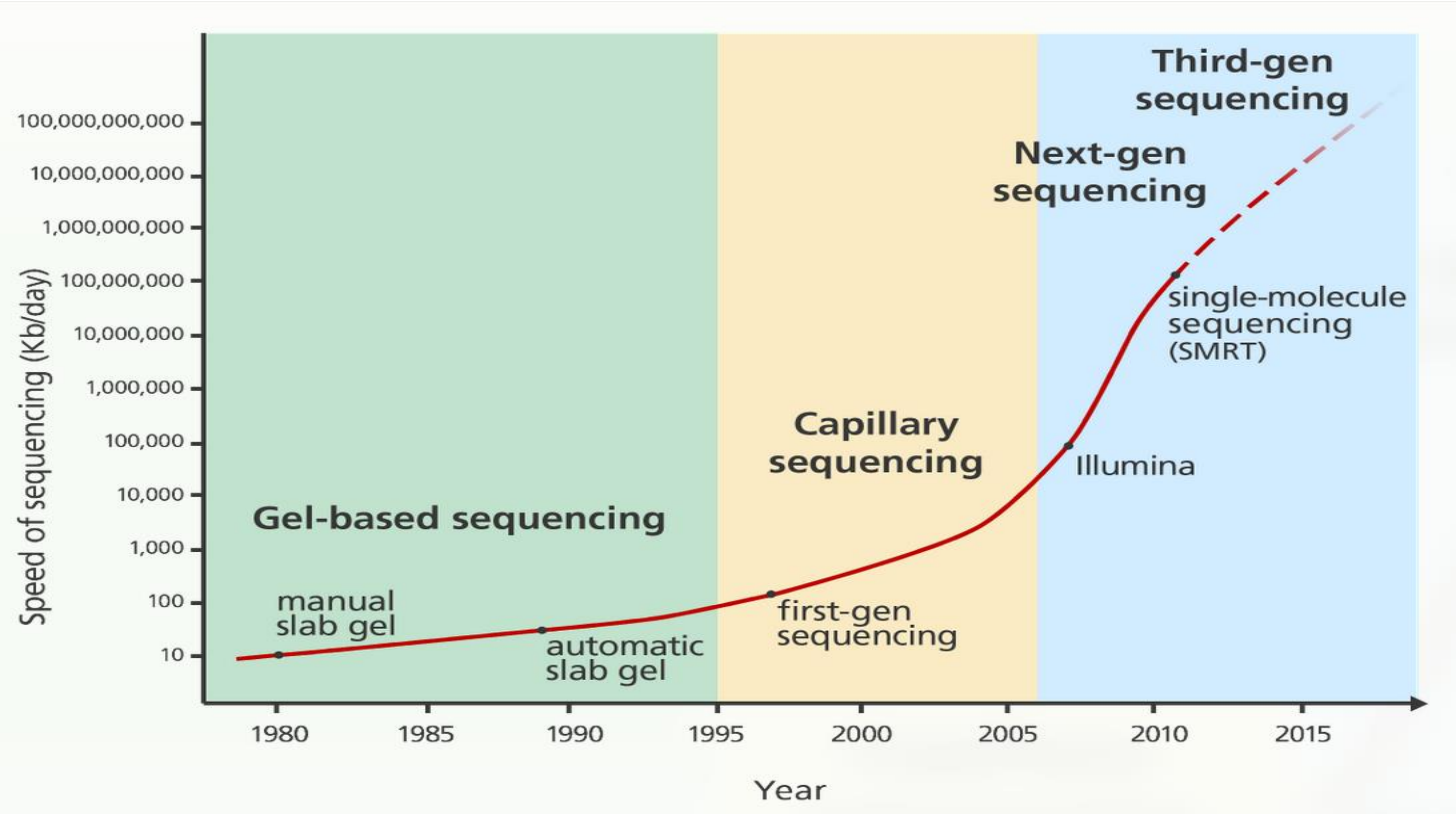


What is sequencing?

- **Sequencing**: telling the order of bases in a given biological sequence.
- Unfortunately, we do not have technologies that can take a DNA and determine its sequence from one end to another.
  - So, what happens is that DNA is chopped and put back together.
- 2001 The human genome (3,000 Mbp) is published: taking roughly 10 years and three billion dollars.
- Using 1st generation sequencing.
  - This method results in a read length that is ~800 bases on average, but may be extended to above 1000 bases
  - Small amounts of DNA could be processed per unit time (throughput), as well as high cost, resulted in it taking roughly 10 years and three billion dollars to sequence the first human genome.

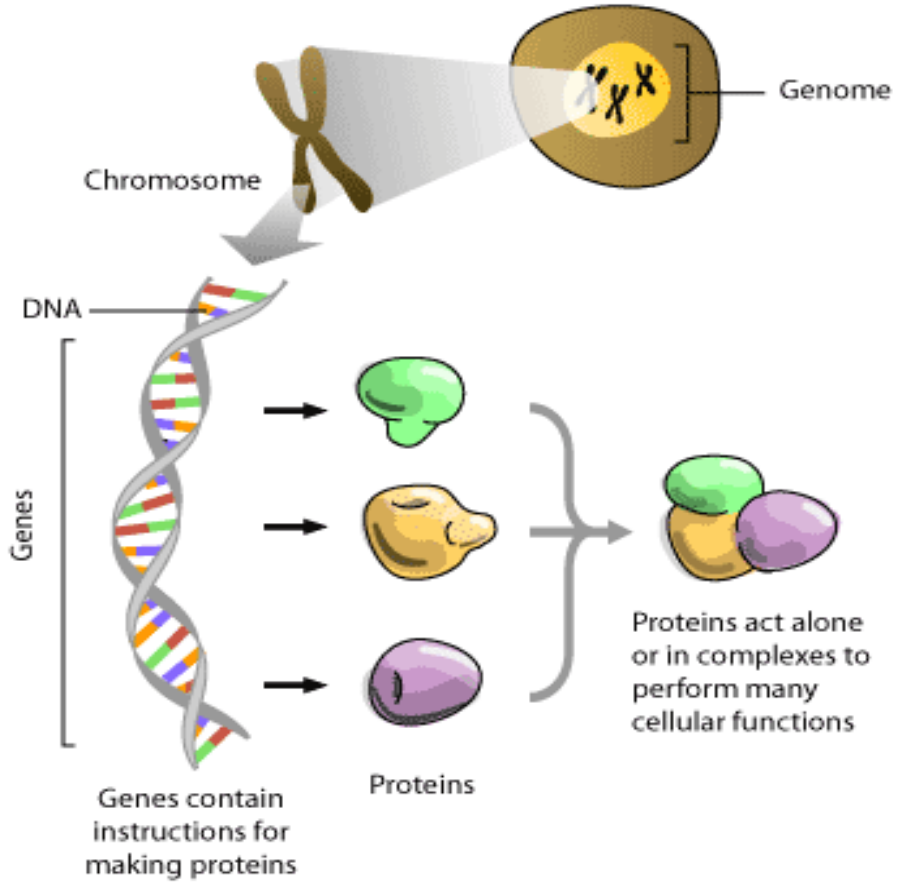
http://hmg.oxfordjournals.org/content/19/R2/R227.full

Speed of DNA Sequencing



A graph showing how the speed of DNA sequencing technologies has increased since the early techniques in the 1980s.
Image credit: Genome Research Limited
Source: http://www.yourgenome.org/stories/third-generation-sequencing

Human Genome Project



Human Genome Project

- Construct a high resolution genetic map of the human genome
- Produce physical maps of all chromosomes
- Determine genome sequence of human and other model organisms
- Develop capabilities (technologies) for collecting, storing, distributing and analyzing data

Background

International Human Genome Sequencing Consortium:

- Lunched in 1990
- 20 governmental groups
- Public project
- Last chromosome sequence published in 2006
- Involved a small number of individual, finished project is a mosaic not representing one individual

Craig Venter & Celera Genomics:

- Founded 1998.
- Sequence in 3 years
- Technology: automation, computers
- Had access to public project's data
- Used shotgun strategy
- Mostly of Craig Venter

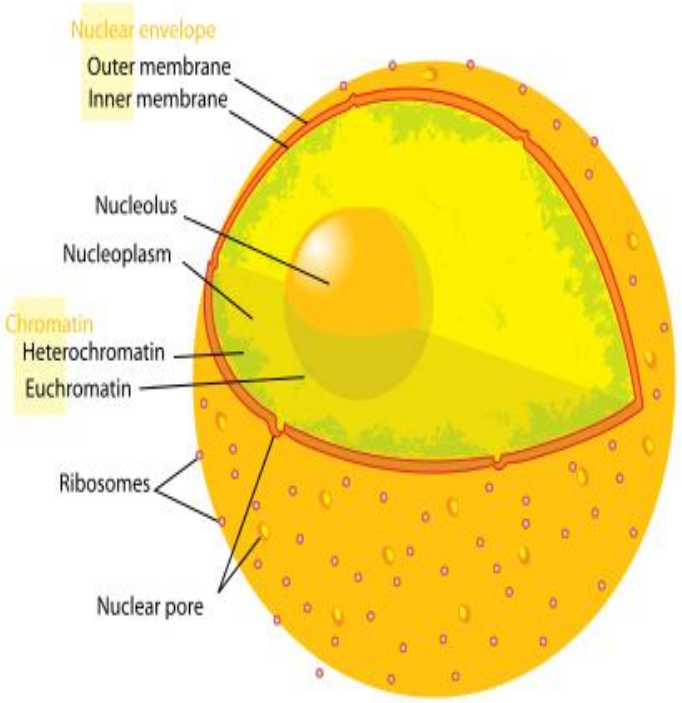
Published in *Science* and *Nature.*

International Human Genome Sequencing Consortium

- Approach was conservative and methodical.
- First produced a **clone-based physical map** of the genome
  - Broke genome into chunks of DNA whose position on chromosome was known from maps, clone into bacteria using bacterial artificial chromosome (BACs).
  - Digest BAC-inserted clonal chunks of DNA into small fragments.
  - Sequence small fragments.
  - Stitch together BAC clones to assemble sequence.
  - Assemble genome sequence from BAC clone sequences, using clone-based physical map.
  - Sequenced only *euchromatic* regions of the genome.

Euchromatin Region

- It is a lightly packed form of chromatin (DNA, RNA, and protein) that is enriched in genes, and is often (but not always) under active transcription.
- Euchromatin comprises the most active portion of the genome within the cell nucleus. 92% of the human genome is euchromatic.

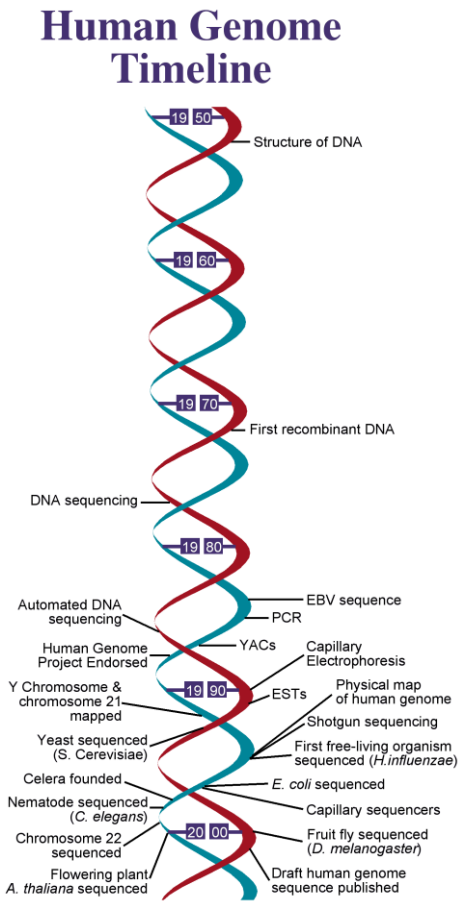


Celera

- Approach using "shotgun sequencing" (no organized map).
- Shreds genome **randomly** into small fragments with no idea of where they are physically located.
- Clones and sequences fragments.
- Uses computer to stitch together genome by matching overlapping ends of sequenced fragments.

Human genome timeline

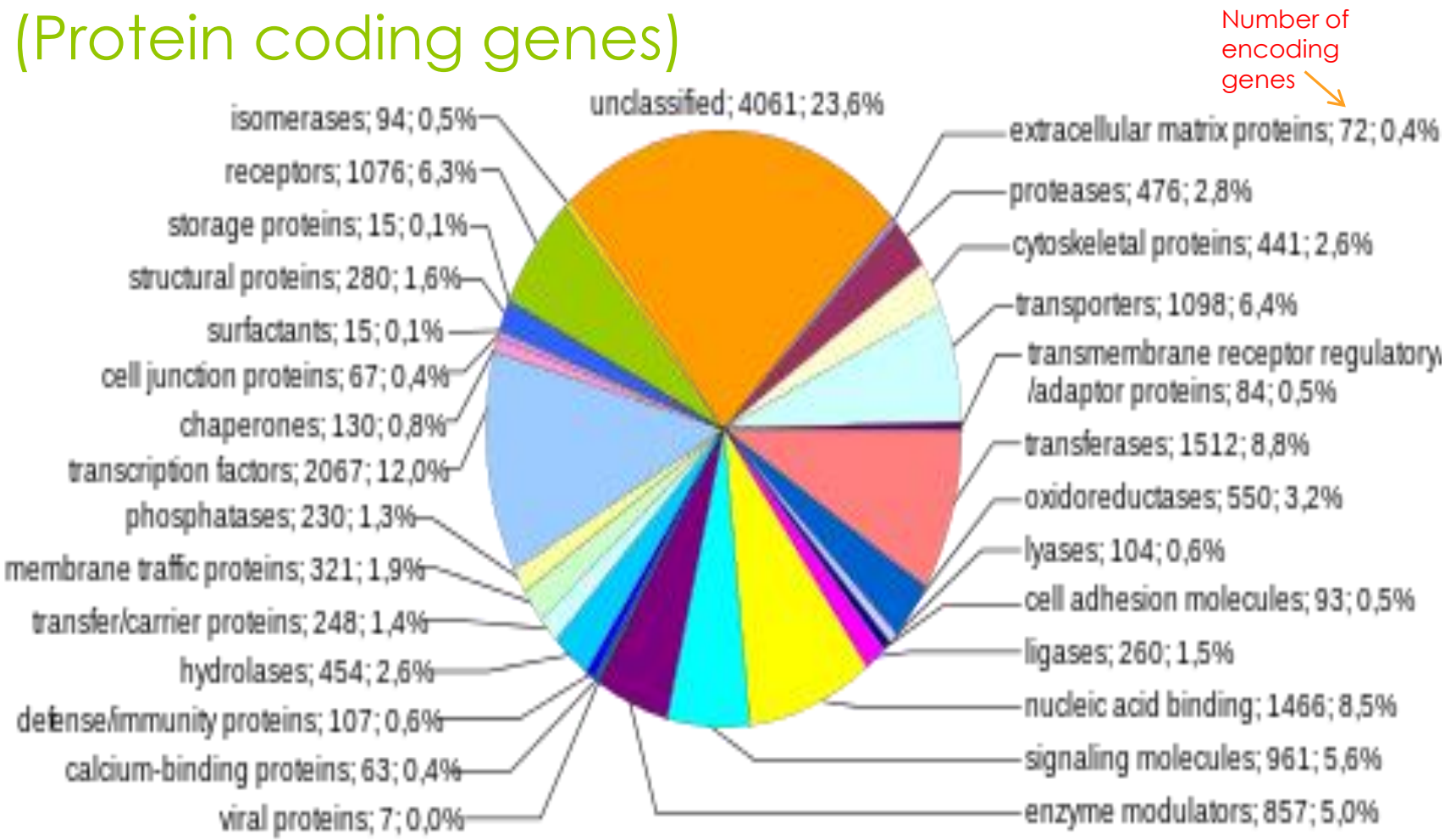
- Genome sequencing driven by technology.
  - Year 1985: 500 base pairs per day by hand.
  - Year 1985-86: PCR and automated DNA sequencing.
  - Year 2000: 1000 bases per second.



Coding vs. Noncoding DNA

- < 2% of genome actually encodes protein
- > 98% of genome is non-coding
  - More junk DNA
  - **Fewer genes!!!!**

Coding sequences
(Protein coding genes)



Human genes categorized by function of the transcribed proteins

Wiki

Access to Information

- All public project data on the Internet.
- NCBI Website: www.ncbi.nlm.nih.gov.
  - Human genome database.
  - Sequence and mapping tools.