CS3225v:
Combinatorial Methods in Computation Biology

Basics of Bioinformatics

Wing-Kin Sung, Ken 宋永健
ksung@comp.nus.edu.sg

Outline

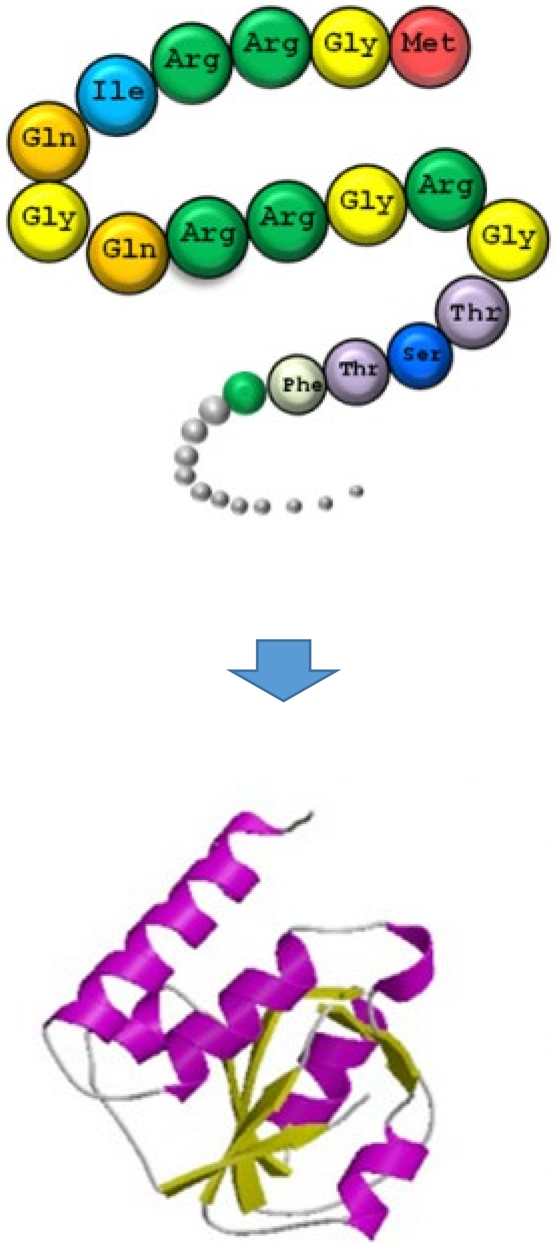
- Basic Molecular Biology
- Technologies
- Problems that can be solved by bioinformatics
- Other bioinformatics problems solved by us

Cell

- Cell performs two type of functions:
    - Perform chemical reactions necessary to maintain our life
    - Pass the information for maintaining life to the next generation
- Actors:
    - Protein performs chemical reactions
    - DNA stores and passes information
    - RNA is the intermediate between DNA and proteins

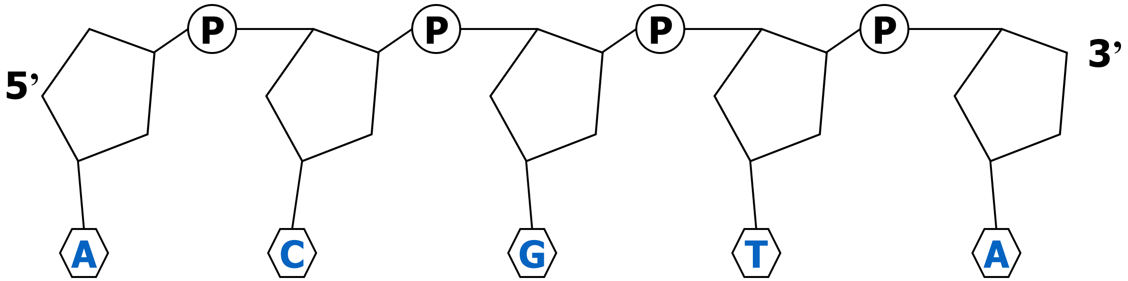
Protein

- Protein is a sequence composed of an alphabet of 20 amino acids.
  - The length is in the range of 20 to more than 5000 amino acids.
  - In average, protein contains around 350 amino acids.
- Protein folds into three-dimensional shape, which form the building blocks and perform most of the chemical reactions within a cell.
  - Structural: building blocks of cells
  - Signaling: Turn gene on or off, Pass signal between cells, Get signal from environment.
  - Catalyze reaction: Enzyme



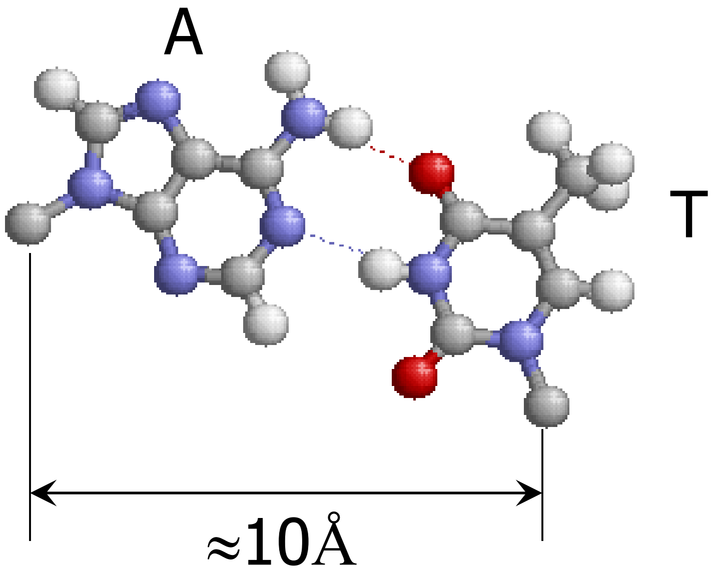
DNA

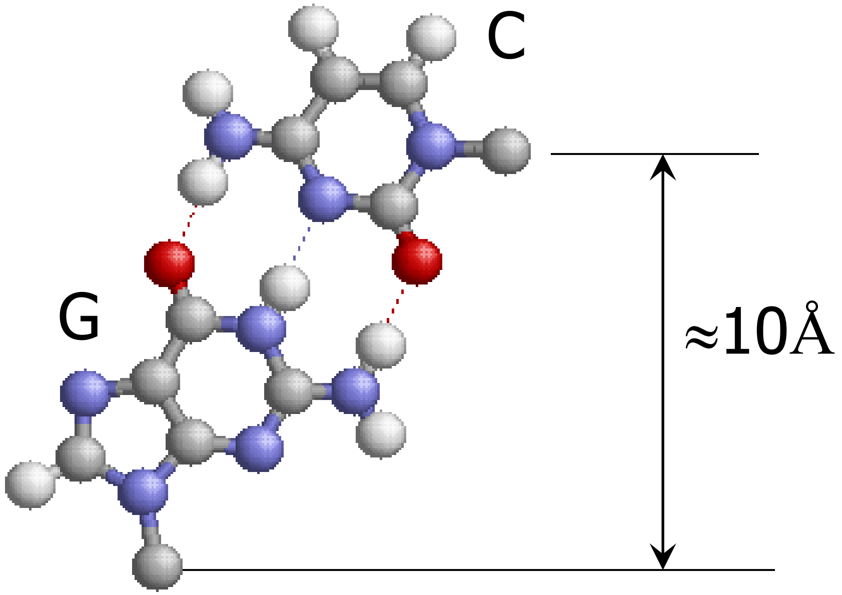
- DNA stores the instruction needed by the cell to perform daily life function.
- It consists of two strands which interwoven together and form a double helix.
- Each strand is a chain of some small molecules called nucleotides.
- There are 4 types of nucleotides: A, C, G, and T.



Watson-Crick rules

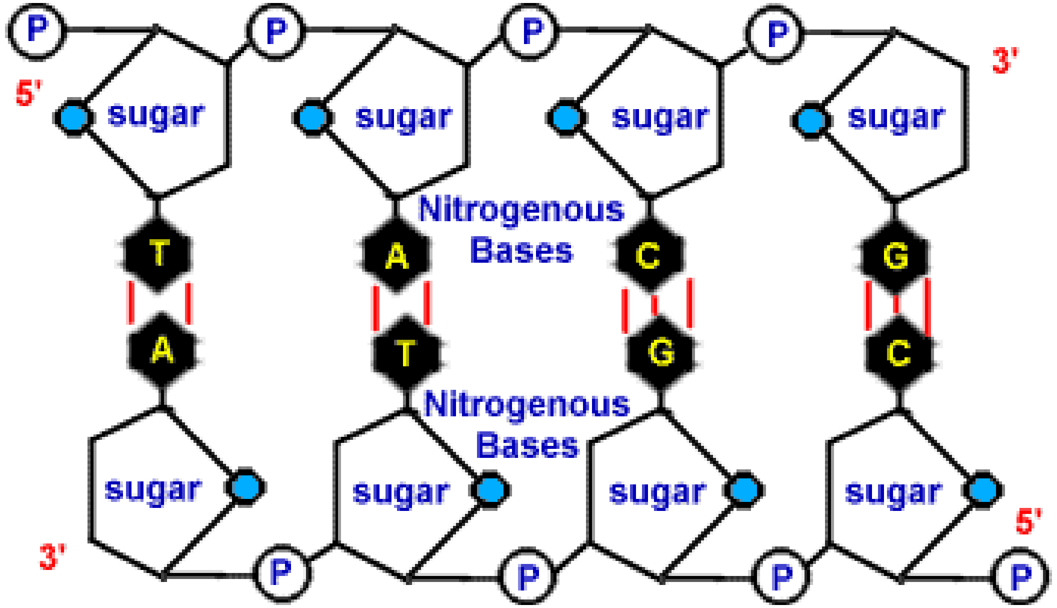
- Complementary bases:
  - A with T (two hydrogen-bonds)
  - C with G (three hydrogen-bonds)

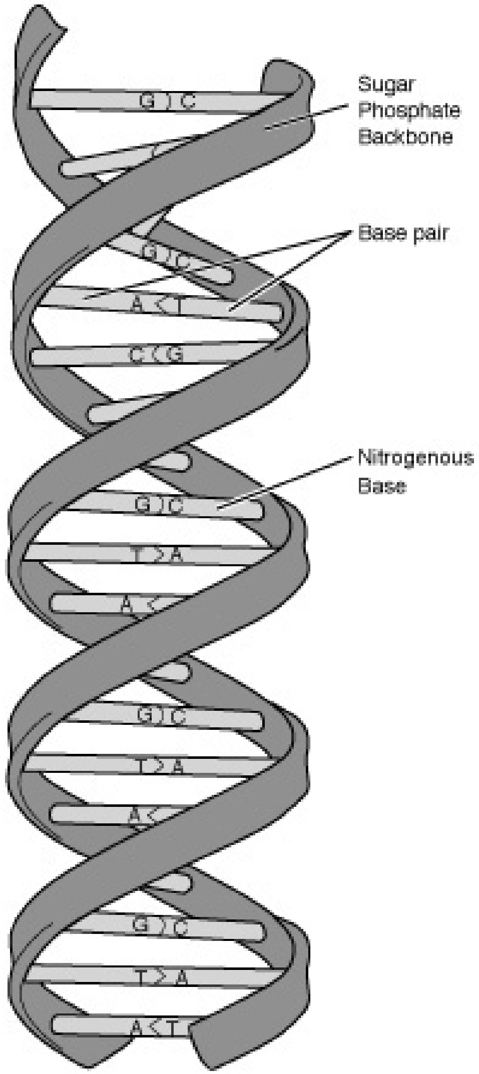




Double stranded DNA

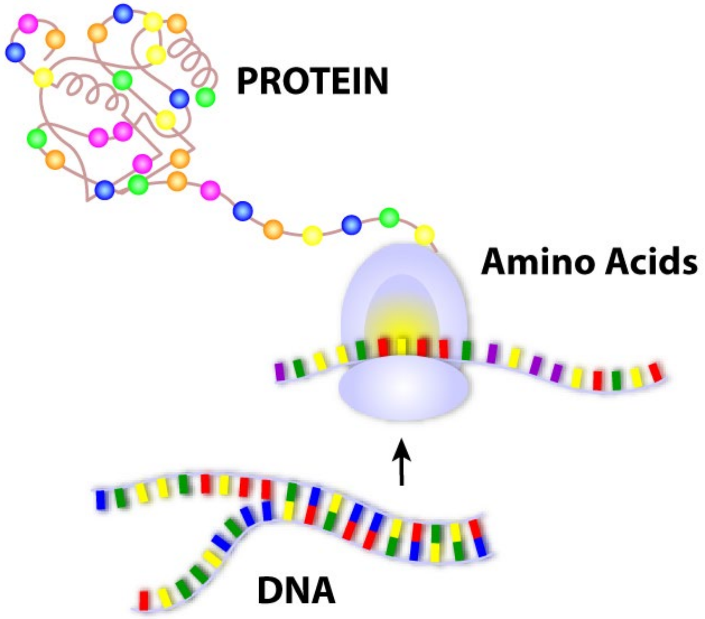
- Normally, DNA is double stranded within a cell. The two strands are antiparallel. One strand is the reverse complement of another one.
- The double strands are interwoven together and form a double helix.
- One reason for double stranded is that it eases DNA replicate.

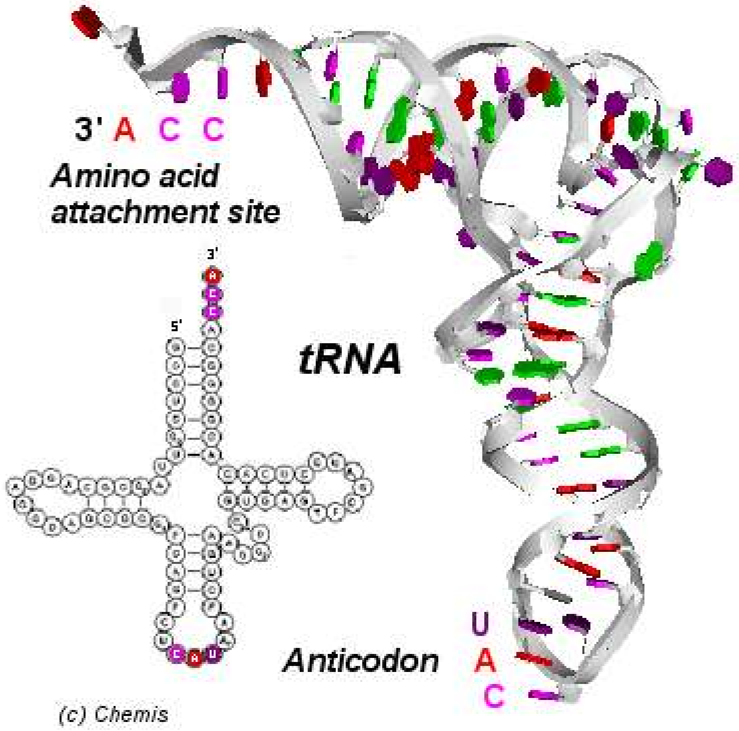




RNA

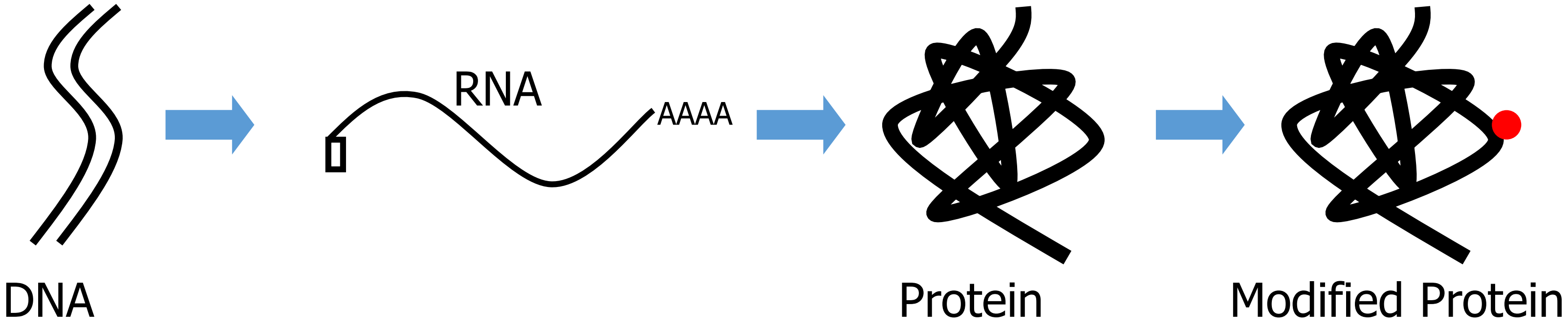
- RNA has two functions
  - As an intermediate between DNA and protein
  - Form complex 3-dimensional structure and perform some functions.





Central Dogma

- Central Dogma tells us how we get the protein from the gene. This process is called gene expression.
- The expression of gene consists of two steps
  - Transcription: DNA → mRNA
  - Translation: mRNA → Protein
  - Post-translation Modification: Protein → Modified protein



Replicate or Repair of DNA

- DNA is double stranded.
- When the cells divide,
  - DNA needs to be duplicated and passes to the two daughter cells.
  - With the help of DNA polymerase, the two strands of DNA serve as template for the synthesis of another complementary strands, generating two identical double stranded DNAs for the two daughter cells.
- When one strand is damaged,
  - it is repaired with the information of another strand.

What is bioinformatics?
(from computer science point of view)

- [wiki] Bioinformatics is an interdisciplinary field that develops methods and software tools for understanding biological data.

- Bioinformatics combines
  - biology,
  - computer science,
  - information engineering,
  - mathematics and
  - statistics

  to analyze and interpret biological data.

The Promises of Bioinformatics

- To the patient:
  - Better drug, better treatment

- To the pharma:
  - Save time, save cost, make more $

- To the scientist:
  - Better science

Pervasiveness of Bioinformatics

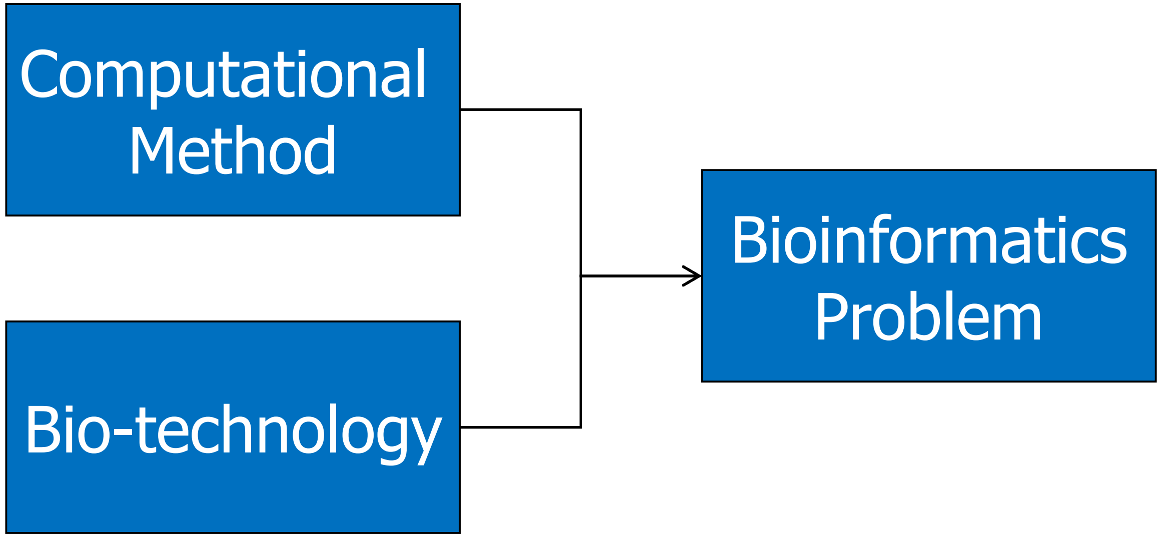
- Bioinformatics is mandatory for large-scale biology
    - e.g., High-throughput, massively-parallel measurements, or “lab on a chip” miniaturization

- Computational data analysis is mandatory for indirect experimental methods
    - e.g., reconstruction haplotype from genotype data

- Limitless opportunities!

What do we study?

- We study the application of computer science and bio-technology to solve bioinformatics problems



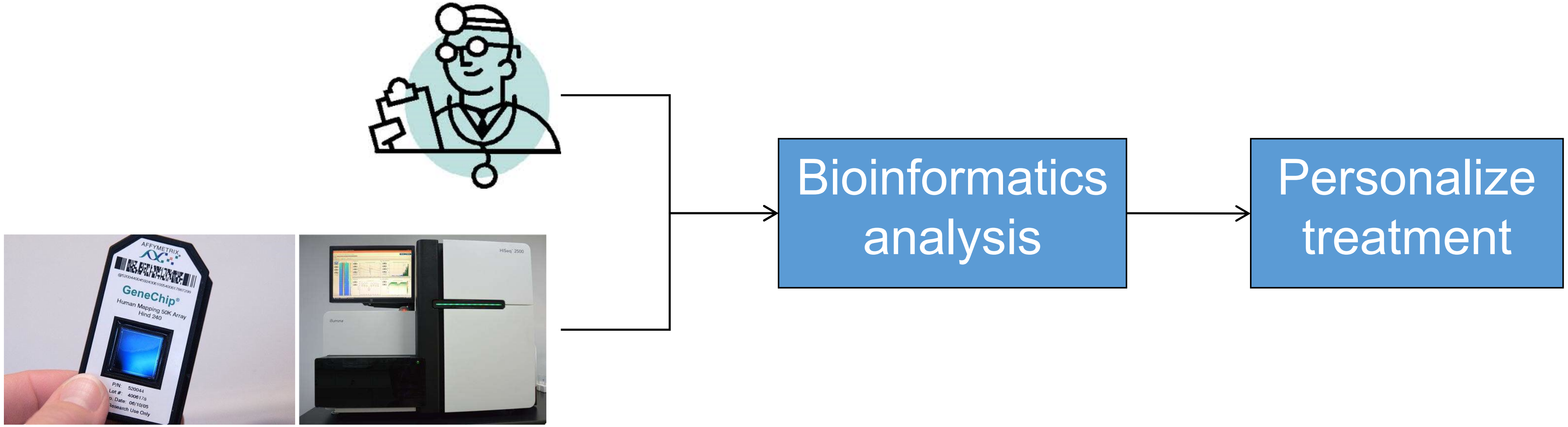
Computational Method

Bio-technology

Bioinformatics Problem

Why these problems are important?

- Personalize sequencing is a big market.



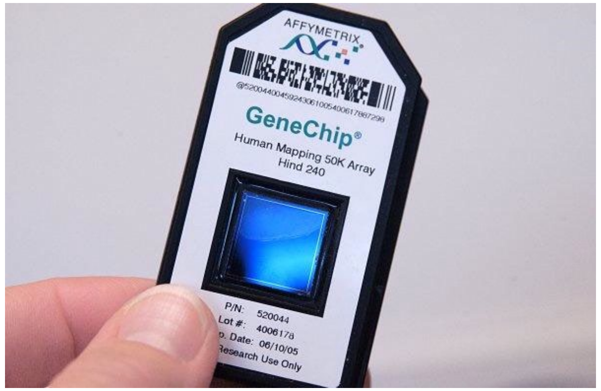
Bioinformatics analysis

Personalize treatment

- A number of big companies and start-up companies.
  - Bioinformatics is the main driving force.

Technologies

DNA array

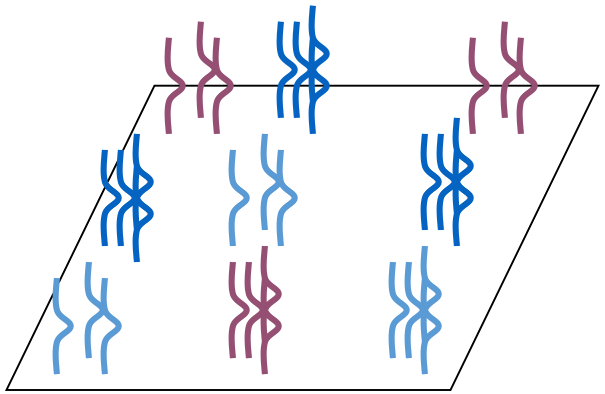


- The idea of hybridization leads to the DNA array technology.

- In the past, “one gene in one experiment”
- Hard to get the whole picture
- DNA array is a technology which allows researchers to do experiment on a set of genes or even the whole genome.

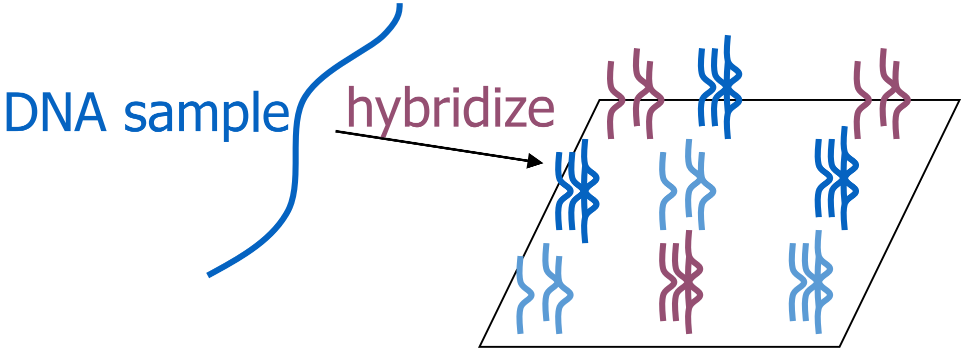
DNA array’s idea (I)

- An orderly arrangement of thousands of spots.
- Each spot contains many copies of the same DNA fragment.



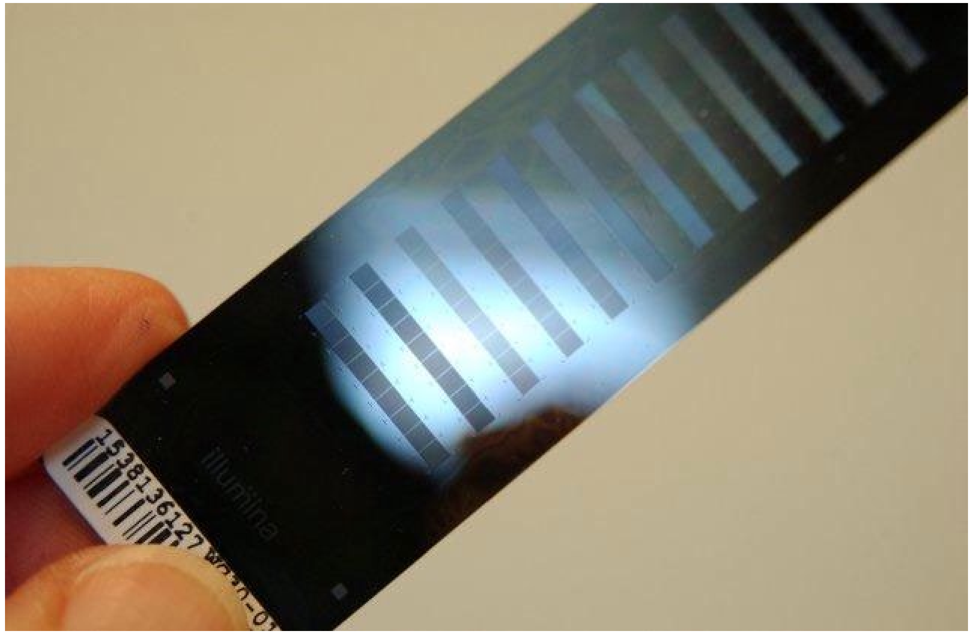
DNA array's idea (II)

- When the array is exposed to the target solution, DNA fragments in both array and target solution will match based on hybridization rule:
  - A=T, C≡G (hydrogen bond)
- Such idea allows us to do thousands of hybridization experiments at the same time.



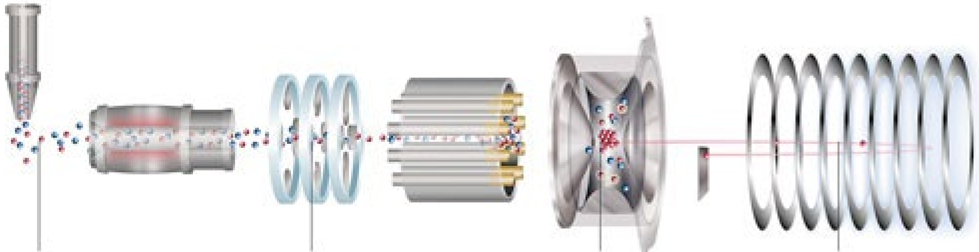
Genotyping chip

- Based on microarray technology.
- Allows us to know the genotype for millions of positions in our genome.



Mass Spec

- Measure mass of different molecules accurately



ESI Interface
Ionization is promoted using drying gas.
Orthogonal source geometry for noise reduction.

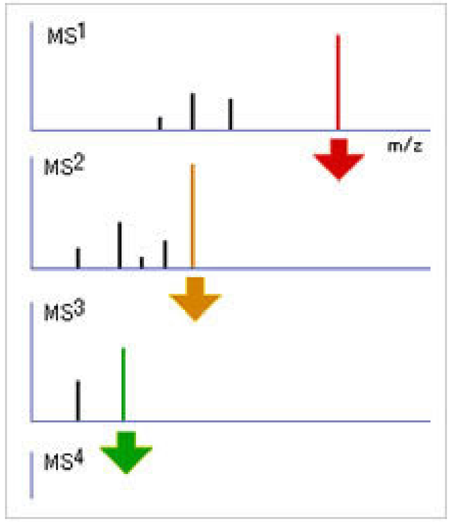
Lens System
Pulsed ion introduction into trap for high sensitivity.

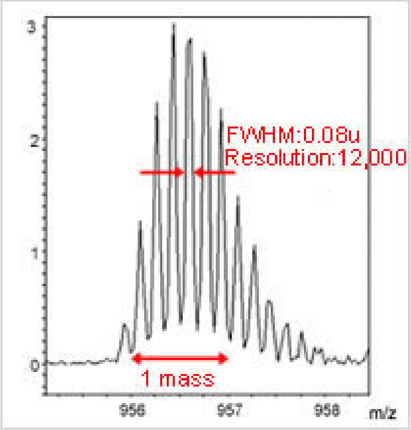
Quadrupole Ion Trap (QIT)
Utilizes pulsed Ar gas introduction for efficient CID.

Time-of-Flight (TOF)
Dual-stage reflectron guarantees excellent resolution and mass accuracy.

MS^n measurement:
One peak acquired by MS^1 is performed MS^2 , and one peak acquired by MS^2 is performed MS^3.
LCMS-IT-TOF can perform by MS^{10}.
This function supports structural analysis strongly.

High resolution and accuracy
This data shows a Mass spectra of Insulin Hexavalent Ion.
Resolution of >12,000 was achieved.
6 peaks are separated clearly in one mass difference.





Main unit: 1685mm , LC unit (by module): 260mm

Sequencing Technology

- **Next-generation sequencing (NGS)** can generate tens of billions of DNA bases efficiently.
- These machines can generate large amount of data per day.
- For example, Illumina sequencer can sequences 60G DNA bases per run.

Illumina HiSeq

Pacific BioSciences

Oxford Nanopore

Short read machine

Long read machine

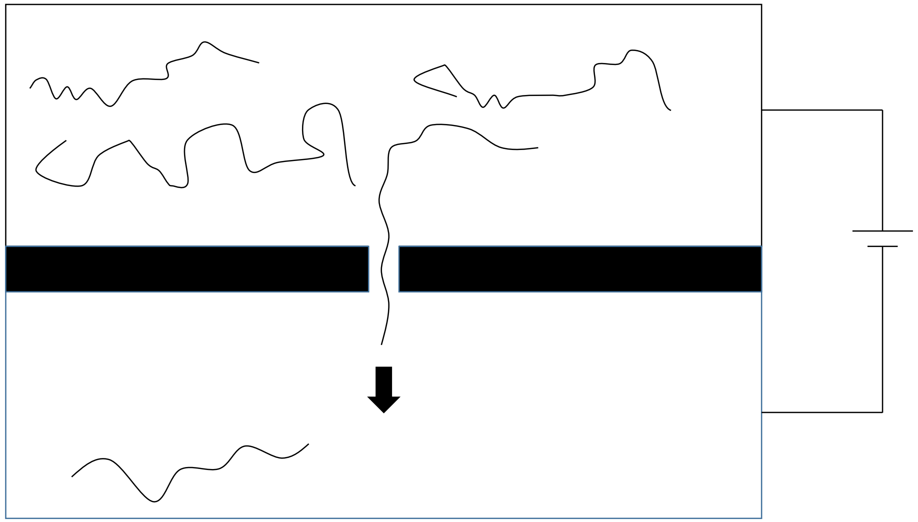
Illumina machine sequences short reads

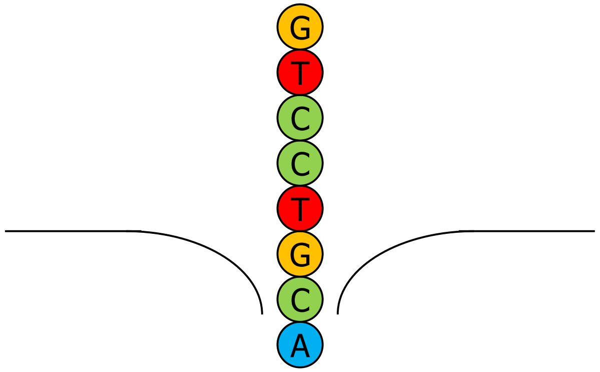
Illumina HiSeq

gatggcccaggagaaccccaagatgcacaactcggagatcagcaagcgcctgggcgccga

Nanopore sequences long reads

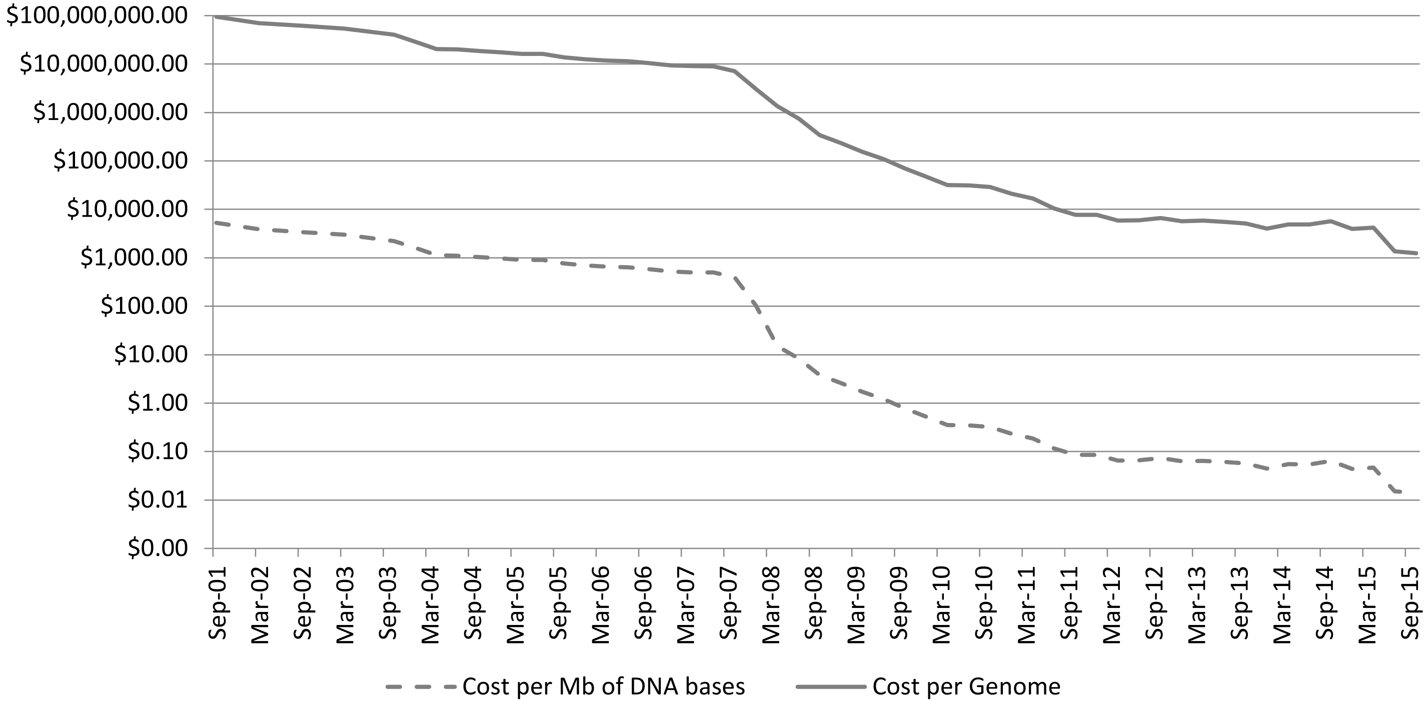
- This technology detect nucleotides by measuring the ionic current flowing through the pore.





The cost of high-throughput sequencing is continue to reduce

- Below figure shows the cost of sequencing.
- Now, to sequence an individual genome, the cost is about US$1000.
- The cost is expected to reduce dramatically in the near future.
- We expect sequencing is popular in the future. (E.g. every individual may sequence their genome.)



Computational techniques

Computational techniques

- Algorithm
  - Greedy algorithm
  - Dynamic Programming
  - EM algorithm
- Data-structure
  - Perfect hashing
  - Suffix tree
- Machine learning
  - SVM
  - k-mean
  - Neural network
- Statistics
  - Normal distribution, etc

Bioinformatics Problems

Example biology problems that can be solved by algorithm

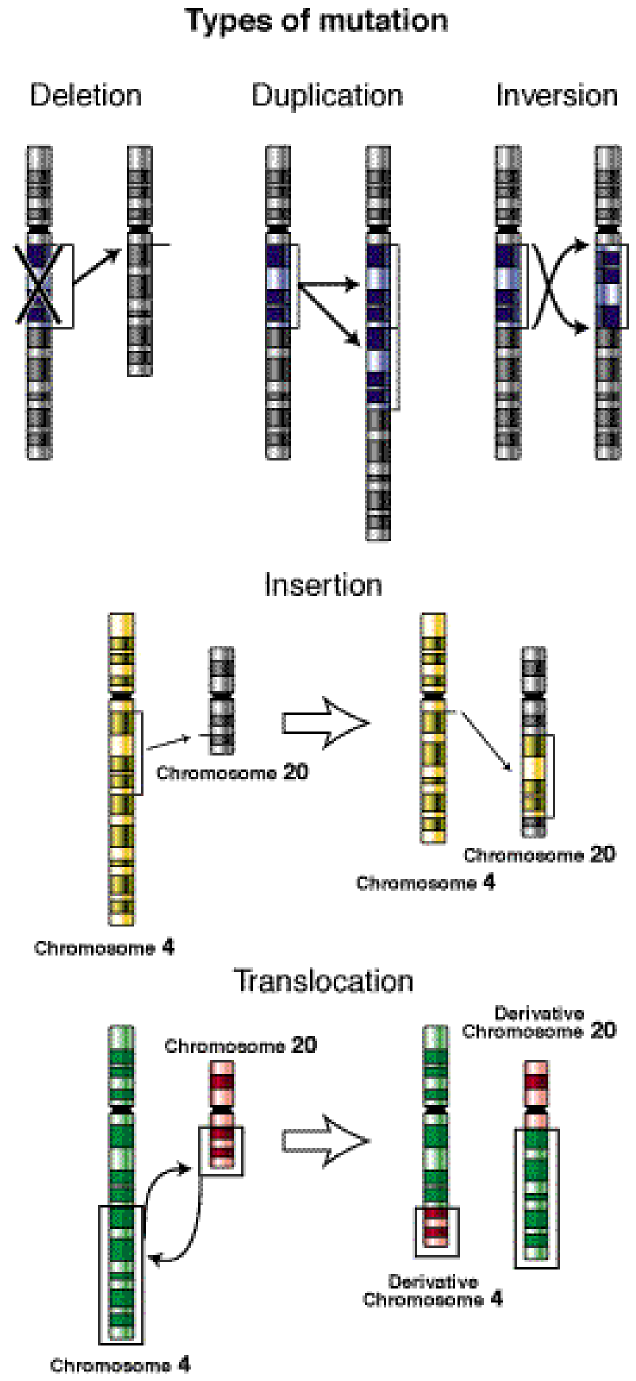
- Learn the mutations in our genome
- Construct and comparing phylogenetic trees
- Whole genome alignment
- Genome rearrangement
- Population genetics
- RNA secondary structure prediction
- Peptide sequencing
- Virus sequencing using microarray

Learning mutation

- Despite the near-perfect replication, infrequent unrepaired mistakes are still possible.
    - Those mistakes are called mutations.

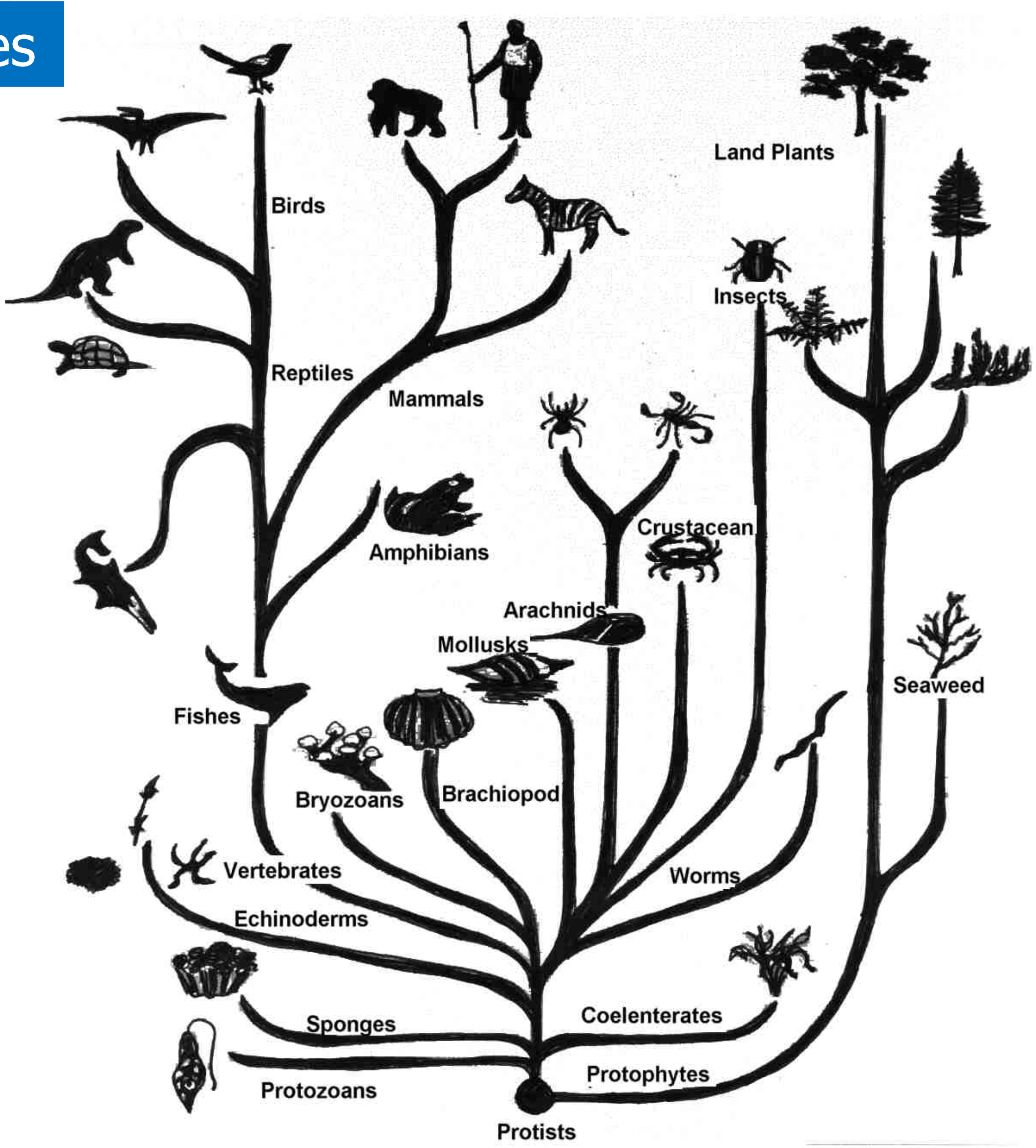
- The most common type of mutation is point mutation.
- Other mutations are structural variations.

- Note: mutation can occur in DNA, RNA, and Protein



Evolutionary tree

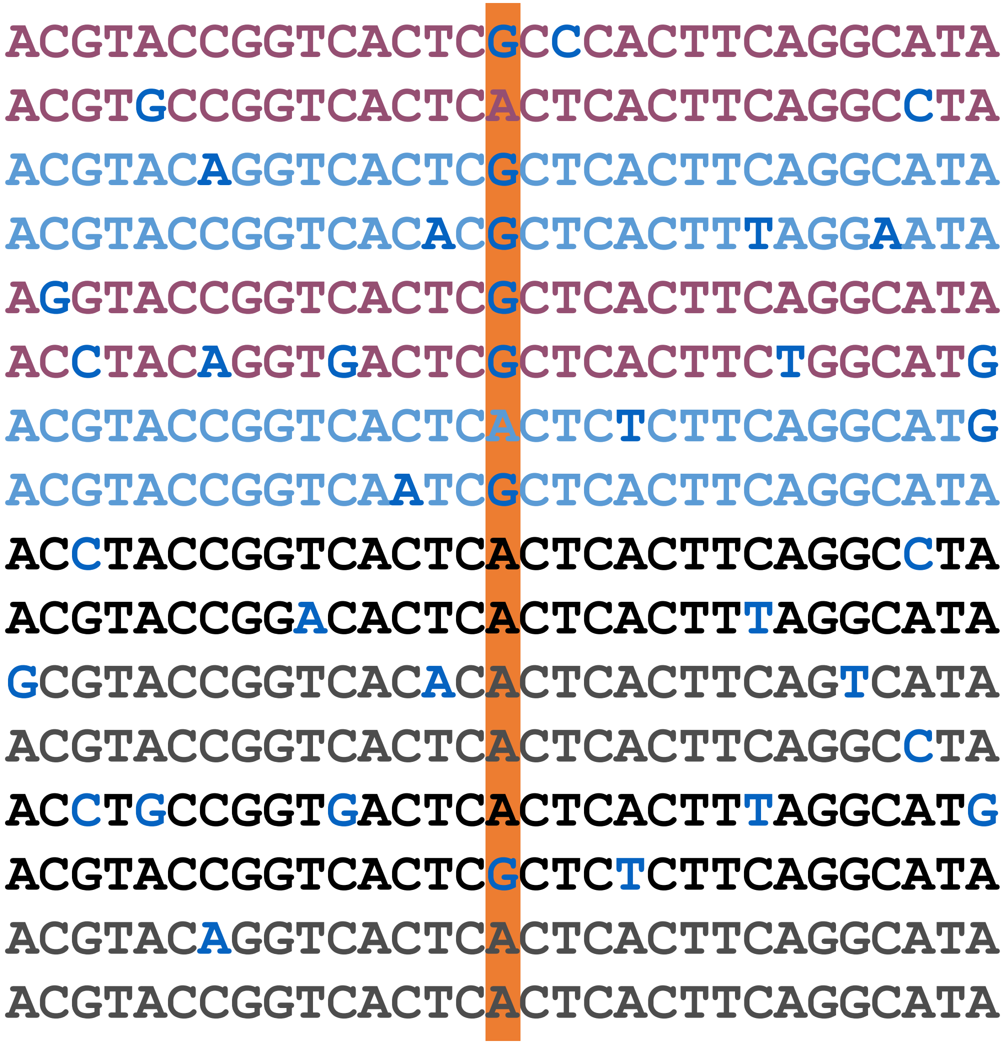
- Occasionally, mutations make the cells or organisms survive better in the environment.
  - The selection of the fittest individuals to survive is called natural selection.
- Mutation and natural selection have resulted in the evolution of a diversified organisms.
- Given the mutations, we can study the evolutionary tree of the individuals.
- Note that mutation is also the cause of diseases (like cancer, flu). We can study diseases by analyzing evolutionary tree.



Population genetics: Finding causal variants

Case
(Disease sample)

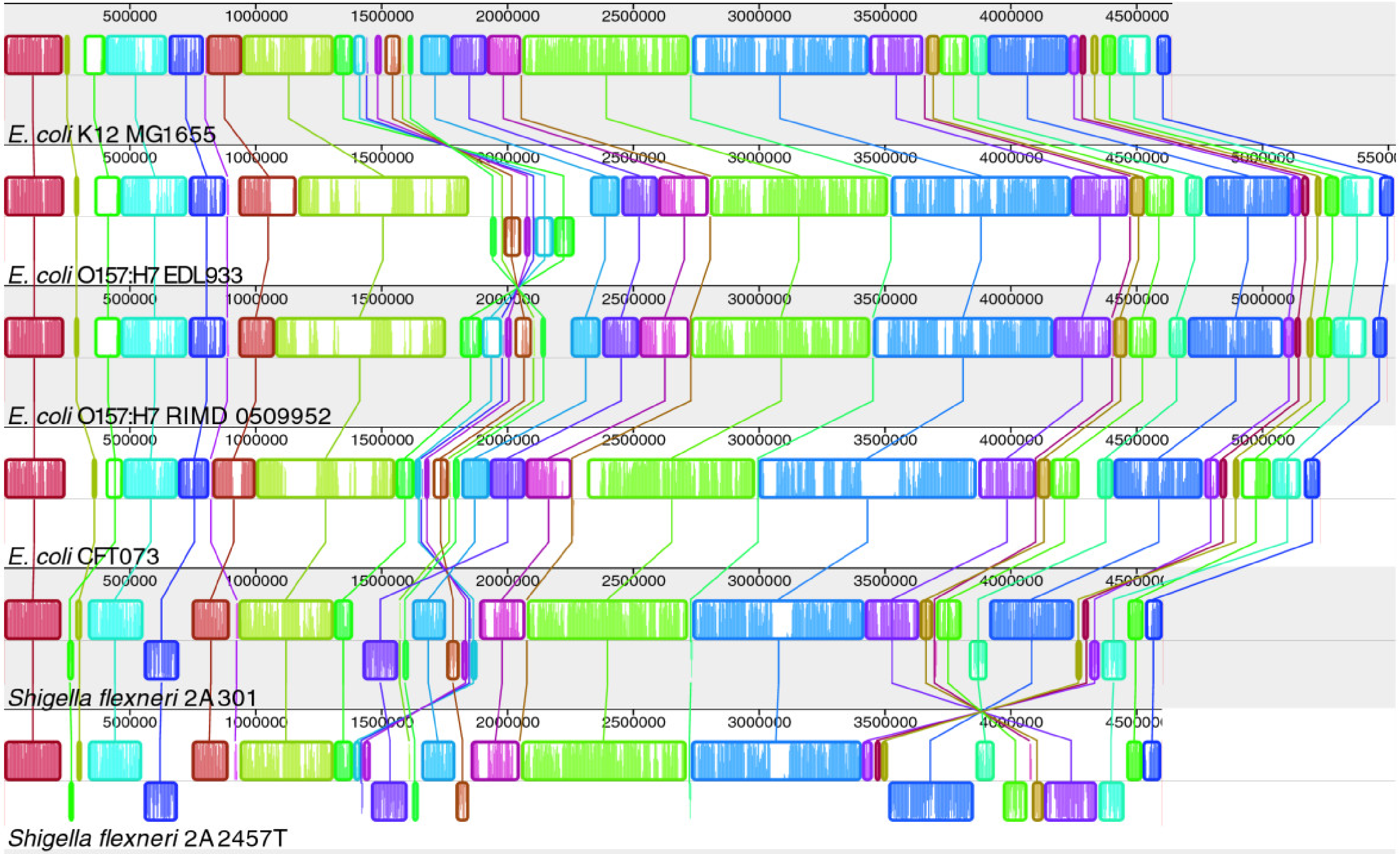
Control
(Normal sample)



```
ACGTACCGGTCACTCGCCCACTTCAGGCATA
ACGTGCCGGTCACTCACTCACTTCAGGCCTA
ACGTACAGGTCACTCGCTCACTTCAGGCATA
ACGTACCGGTCACACGCTCACTTTAGGAATA
AGGTACCGGTCACTCGCTCACTTCAGGCATA
ACCTACAGGTGACTCGCTCACTTCTGGCATG
ACGTACCGGTCACTCACTCTCTTCAGGCATG
ACGTACCGGTCAATCGCTCACTTCAGGCATA
ACCTACCGGTCACTCACTCACTTCAGGCCTA
ACGTACCGGACACTCACTCACTTTAGGCATA
GCGTACCGGTCACACACTCACTTCAGTCATA
ACGTACCGGTCACTCACTCACTTCAGGCCTA
ACCTGCCGGTGACTCACTCACTTTAGGCATG
ACGTACCGGTCACTCGCTCTCTTCAGGCATA
ACGTACAGGTCACTCACTCACTTCAGGCATA
ACGTACCGGTCACTCACTCACTTCAGGCATA
```

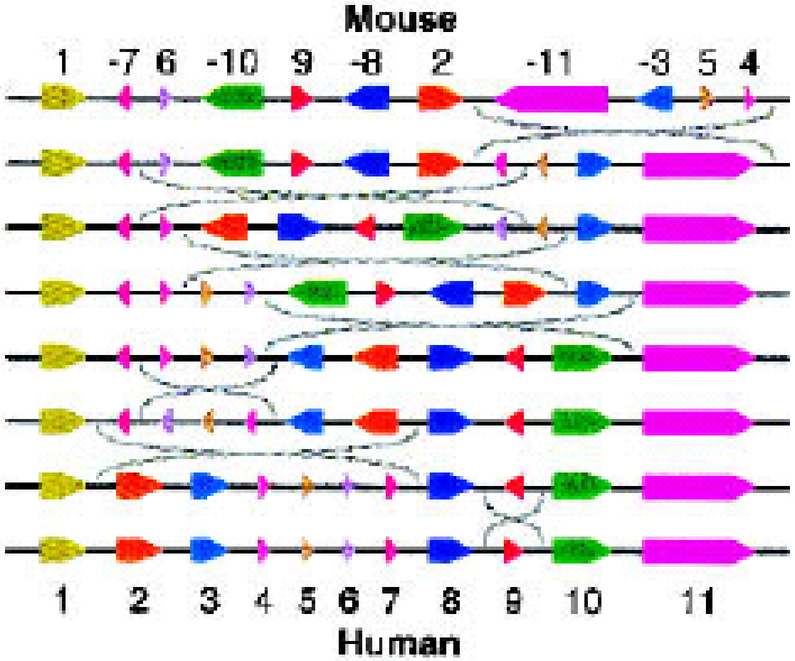
Technology: Sequencing of geome

Whole genome alignment



Genome rearrangement

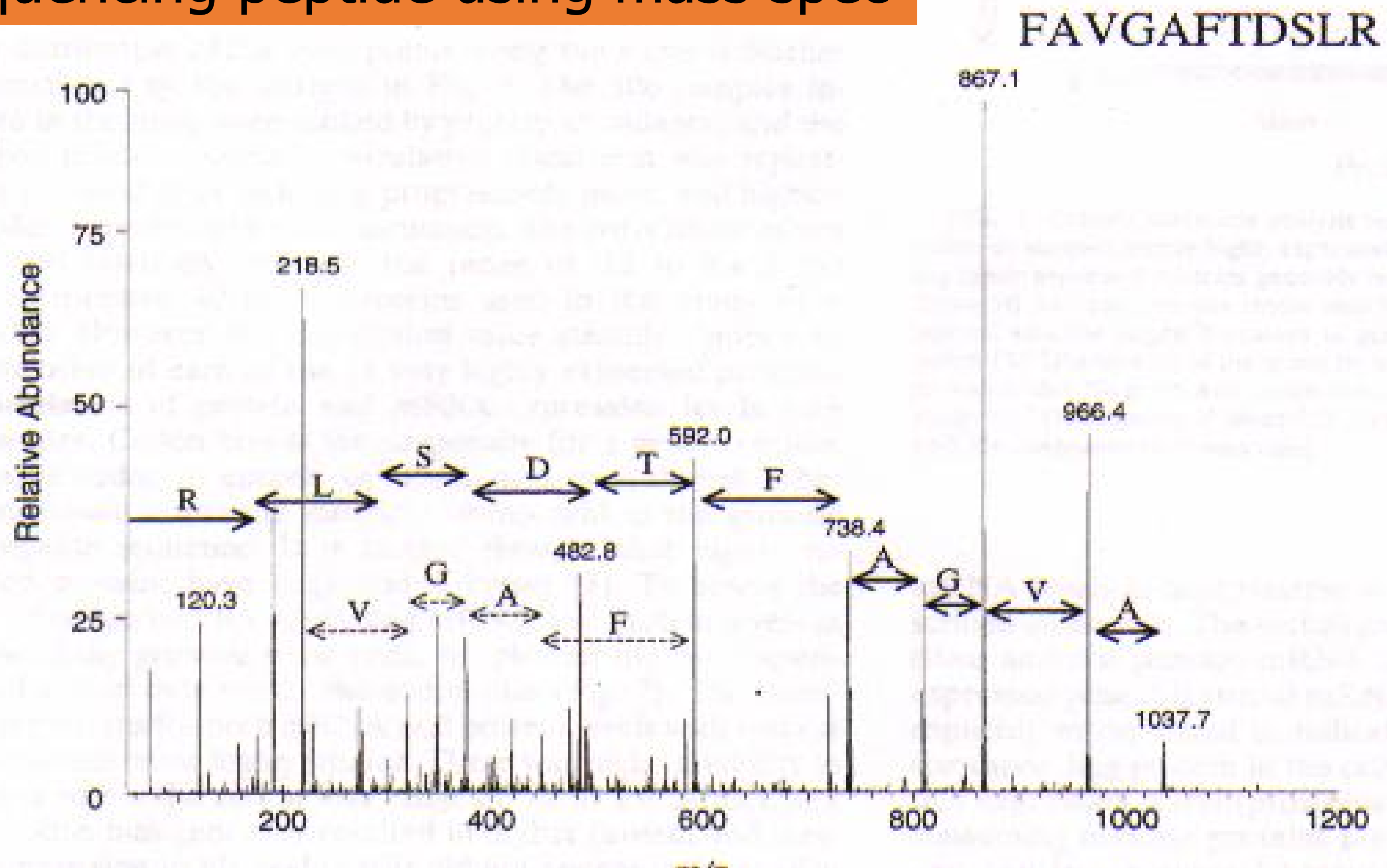
- chromosome X of human can be transformed to chromosome X of mouse using 7 reversals



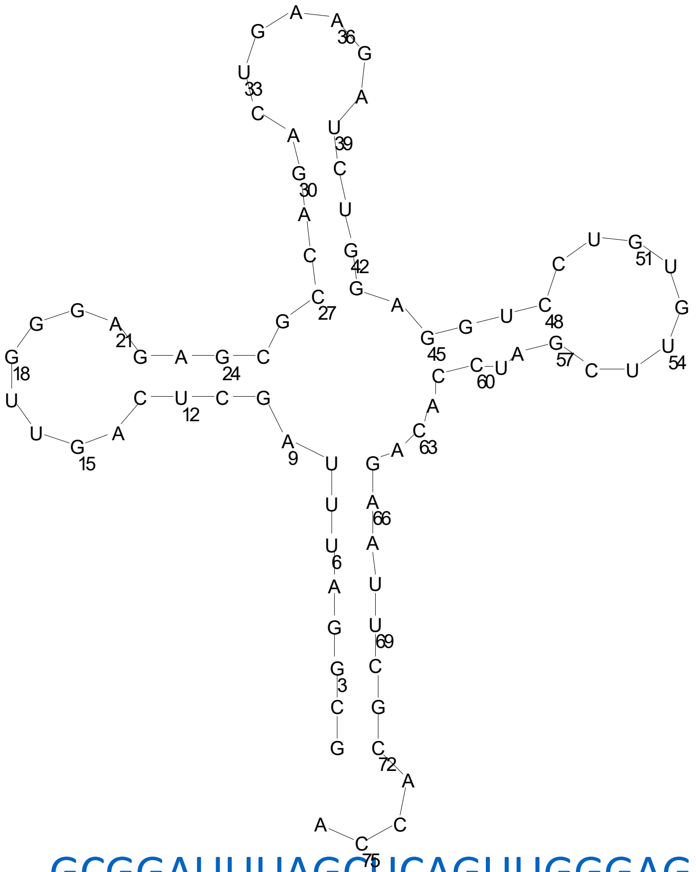
Technology: Mass spec

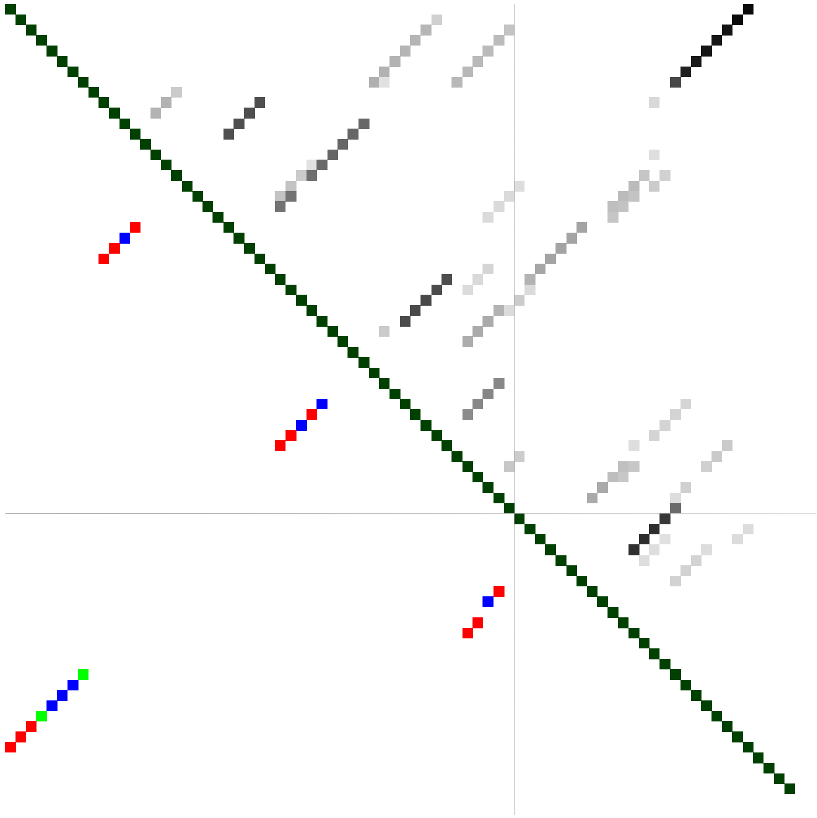
Peptide sequencing

Sequencing peptide using mass spec



Example (Secondary structure for phenylalanyl-tRNA)





GCGGAUUUAGCUCAGUUGGGAGAGCGCCAGACUGAAGAUCUGGAGGUCCUGUG
UUCGAUCCACAGAAUUCGCACCA