Introduction to computational and systems biology

- Lecture 2: A crash course on molecular biology for computer scientists

- Ion Petre
- Department of Mathematics and Statistics
- University of Turku
- Fall 2019

Turku Centre *for* Computer Science

Life inside a cell

A view on "The Inner Life of a Cell" (Harvard University, 2006):

http://www.xvivo.net/the-inner-life-of-the-cell/

Beautiful representation of metabolite transportation, protein-protein binding, DNA replication, DNA ligase, microtubule formation/dissipation, protein synthesis, …

Acknowledgment

- The slides in this lecture are based on N.C.Jones, P.A.Pevzner – "An introduction to bioinformatics algorithms", http://bioalgorithms.info/

How Molecular Biology came about?

- Microscopic biology began in 1665
- Robert Hooke (1635-1703) discovered organisms are made up of cells
- Matthias Schleiden (1804-1881) and Theodor Schwann (1810-1882) further expanded the study of cells in 1830s

- Robert Hooke

- Matthias Schleiden

- Theodor Schwann

Major events in the history of Molecular Biology 1800 - 1870

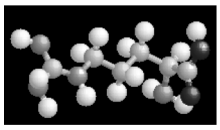
- **1865** Gregor Mendel discover the basic rules of heredity of garden pea.
  - An individual organism has two alternative heredity units for a given trait (dominant trait v.s. recessive trait)
- **1869** Johann Friedrich Miescher discovered DNA and named it nuclein.

Mendel: The Father of Genetics

Johann Miescher

Major events in the history of Molecular Biology 1880 - 1900

- **1881** Edward Zacharias showed chromosomes are composed of nuclein.
- **1899** Richard Altmann renamed nuclein to nucleic acid.
- **By 1900**, chemical structures of all 20 amino acids had been identified



Major events in the history of Molecular Biology 1900-1911

- **1902** - Emil Hermann Fischer wins Nobel prize: showed amino acids are linked and form proteins
  - Postulated: protein properties are defined by amino acid composition and arrangement, which we nowadays know as fact

Emil Fischer

- **1911 –** Thomas Hunt Morgan discovers genes on chromosomes are the discrete units of heredity

Thomas Morgan

- **1911** Pheobus Aaron Theodore Lerene discovers RNA

Major events in the history of Molecular Biology 1940 - 1950

- **1941** – George Beadle and Edward Tatum identify that genes make proteins

George Beadle

Edward Tatum

- **1950** – Edwin Chargaff find Cytosine complements Guanine and Adenine complements Thymine

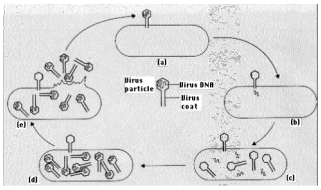
Edwin Chargaff

Major events in the history of Molecular Biology 1950 - 1952

- **1950s** – Mahlon Bush Hoagland first to isolate tRNA

Mahlon Hoagland

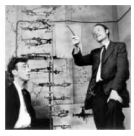
- **1952** – Alfred Hershey and Martha Chase make genes from DNA



Hershey Chase Experiment

Major events in the history of Molecular Biology 1952 - 1960

- **1952-1953** James D. Watson and Francis H. C. Crick deduced the double helical structure of DNA



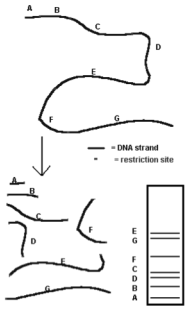
James Watson and Francis Crick

- **1956** George Emil Palade showed the site of enzymes manufacturing in the cytoplasm is made on RNA organelles called ribosomes.

George Emil Palade

Major events in the history of Molecular Biology 1970

- **1970** Howard Temin and David Baltimore independently isolate the first restriction enzyme
- DNA can be cut into reproducible pieces with site-specific endonuclease called restriction enzymes;
  - the pieces can be linked to bacterial vectors and introduced into bacterial hosts. (gene cloning or recombinant DNA technology)



Major events in the history of Molecular Biology 1970- 1977

- **1977** Phillip Sharp and Richard Roberts demonstrated that pre-mRNA is processed by the excision of introns and exons are spliced together.

Phillip Sharp

Richard Roberts

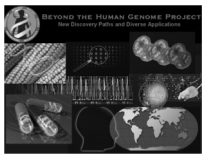
- Joan Steitz determined that the 5′ end of snRNA is partially complementary to the consensus sequence of 5′ splice junctions.

Joan Steitz

Major events in the history of Molecular Biology 1986 - 1995

- **1986** Leroy Hood: Developed automated sequencing mechanism
- **1986** Human Genome Initiative announced
- **1990** The 15 year Human Genome project is launched by congress
- **1995** Moderate-resolution maps of chromosomes 3, 11, 12, and 22 maps published (These maps provide the locations of "markers" on each chromosome to make locating genes easier)


Leroy Hood



Major events in the history of Molecular Biology 1995-1996

- **1995** John Craig Venter: First bacterial genomes sequenced
- **1995** Automated fluorescent sequencing instruments and robotic operations
- **1996** First eukaryotic genome-yeast-sequenced


John Craig Venter

Major events in the history of Molecular Biology 1997 - 1999

- **1997** E. Coli sequenced
- **1998** PerkinsElmer, Inc. Developed 96-capillary sequencer
- **1998** Complete sequence of the Caenorhabditis elegans genome
- **1999** First human chromosome (number 22) sequenced

Major events in the history of Molecular Biology 2000-2001

- **2000** Complete sequence of the euchromatic portion of the Drosophila melanogaster genome
- **2001** International Human Genome Sequencing:first draft of the sequence of the human genome published

Major events in the history of Molecular Biology 2003- Present

- **April 2003** Human Genome Project Completed. Mouse genome is sequenced.
- **April 2004** Rat genome sequenced.

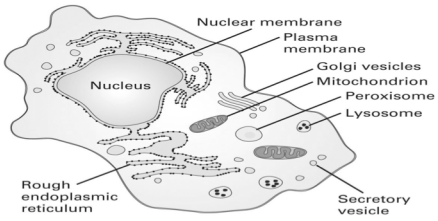
Start with something simple: viruses

- Viruses are essentially just a protein coat hosting some DNA
  - In particular they do not have the machinery to replicate themselves
  - Well-studied example: lambda-phage
  - The protein coat attaches to the membrane of a cell and inserts the viral DNA into the cell
  - Once in, the viral DNA loops on itself forming a circular molecule
  - The cell's own transcription machinery will transcribe the viral DNA as if it were its own
  - In the case of the lambda-phage, the result is a protein called *lambda integrase* that inserts the viral DNA in the host's chromosomal DNA
  - The cell and all its descendants are from now on carriers of the viral DNA
  - Some external event may trigger the virus to become active: excise its DNA from the host's chromosome, multiply itself, create protein coats, assemble many copies of the virus, destroy the cell's membrane and release the new lambda phage to the intercellular environment

Plasmids

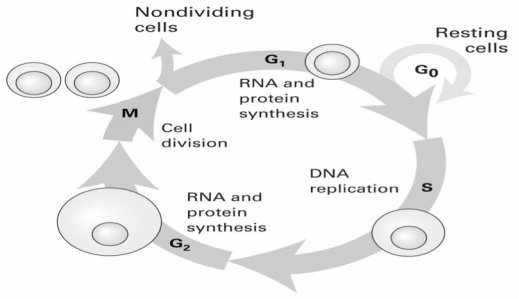
- There is nothing special about the viral DNA that makes the cell transcribe it as if it were its own
  - The same machinery will recognize any plasmid (circular DNA) and transcribe it as well
  - The basis for bioengineering: encode into DNA the "instructions" and have the cell execute the code

Life begins with Cell



- A cell is a smallest structural unit of an organism that is capable of independent functioning
- All cells have some common features

All Cells have common Cycles

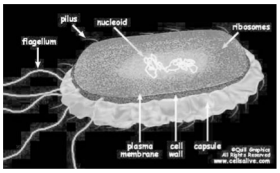


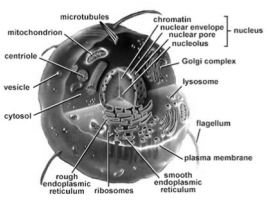
- Born, eat, replicate, and die

Cells

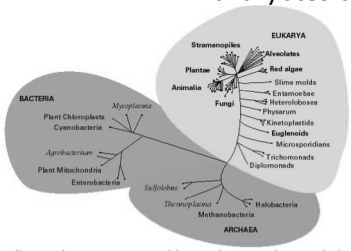
- Chemical composition-by weight
  - 70% water
  - 7% small molecules
    - salts
    - Lipids
    - amino acids
    - nucleotides
  - 23% macromolecules
    - Proteins
    - Polysaccharides
    - lipids

2 types of cells: Prokaryotes v.s.Eukaryotes





Prokaryotes and Eukaryotes



- According to the most recent evidence, there are three main branches to the tree of life.
- Prokaryotes include Archaea ("ancient ones") and bacteria.
- Eukaryotes are kingdom Eukarya and includes plants, animals, fungi and certain algae.

Prokaryotes and Eukaryotes, continued

Prokaryotes	Eukaryotes
Single cell	Single or multi cell
No nucleus	Nucleus
No organelles	Organelles
One piece of circular DNA	Chromosomes
No mRNA post transcriptional modification	Exons/Introns splicing

Cells Information and Machinery

- Cells store all information to replicate itself
  - Human genome is around 3 billions base pair long
  - Almost every cell in human body contains same set of genes
  - But not all genes are used or expressed by those cells
- Machinery:
  - Collect and manufacture components
  - Carry out replication
  - Kick-start its new offspring

  (A cell is like a car factory)

Overview of organizations of life

- **Nucleus** = **library**
- **Chromosomes** = **bookshelves**
- **Genes** = **books**
- Almost every cell in an organism contains the same libraries and the same sets of books.
- Books represent all the information (DNA) that every cell in the body needs so it can grow and carry out its vaious functions.

Some Terminology

- **Genome**: **an organism's genetic material**
- **Gene**: **a discrete units of hereditary information located on the chromosomes and consisting of DNA.**
- **Genotype**: **The genetic makeup of an organism**
- **Phenotype**: **the physical expressed traits of an organism**
- **Nucleic acid**: **Biological molecules(RNA and DNA) that allow organisms to reproduce;**

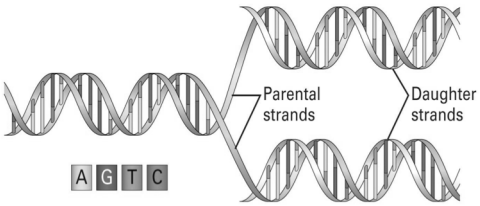
More Terminology

- The **genome** is an organism's complete set of DNA.
  - a bacteria contains about 600,000 DNA base pairs
  - human and mouse genomes have some 3 billion.
- human genome has 24 distinct chromosomes.
  - Each chromosome contains many **genes**.
- **Gene**
  - basic physical and functional units of heredity.
  - specific sequences of DNA bases that encode instructions on how to make **proteins**.
- **Proteins**
  - Make up the cellular structure
  - large, complex molecules made up of smaller subunits called **amino acids.**

All Life depends on 3 critical molecules

- DNAs
  - Hold information on how cell works
- RNAs
  - Act to transfer short pieces of information to different parts of cell
  - Provide templates to synthesize into protein
- Proteins
  - Form enzymes that send signals to other cells and regulate gene activity
  - Form body's major components (e.g. hair, skin, etc.)

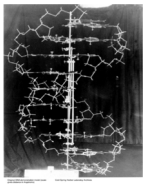
DNA: The Code of Life



- The structure and the four genomic letters code for all living organisms
- Adenine, Guanine, Thymine, and Cytosine which pair A-T and C-G on complimentary strands.

Discovery of DNA

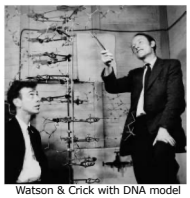
- DNA Sequences
  - Chargaff and Vischer, 1949
    - DNA consisting of A, T, G, C
      Adenine, Guanine, Cytosine, Thymine
  - Chargaff Rule
    - Noticing #A≈#T and #G≈#C
      A "strange but possibly meaningless" phenomenon.
- Wow!! A Double Helix
  - Watson and Crick, *Nature,* April 25, 1953
  - 1 Biologist
    1 Physics Ph.D. Student
    \+ 900 words
    = Nobel Prize
  - Rich, 1973
    - Structural biologist at MIT.
    - DNA's structure in atomic resolution.



Crick Watson

Watson & Crick – *"…the secret of life"*

- Watson: a zoologist, Crick: a physicist
- "*In 1947 Crick knew no biology and practically no organic chemistry or crystallography..*" – www.nobel.se
- Applying Chagraff's rules and the X-ray image from Rosalind Franklin, they constructed a "tinkertoy" model showing the double helix
- Their 1953 *Nature* paper: *"It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material."*



Watson & Crick with DNA model

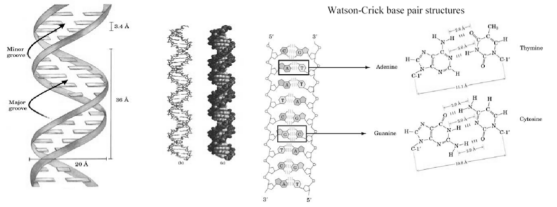
Rosalind Franklin with X-ray image of DNA

Double helix of DNA

- James Watson and Francis Crick proposed a model for the structure of DNA.
  - Utilizing X-ray diffraction data, obtained from crystals of DNA)
- This model predicted that DNA
  - as a helix of two complementary anti-parallel strands,
  - wound around each other in a rightward direction
  - stabilized by H-bonding between bases in adjacent strands.
  - The bases are in the interior of the helix
    - Purine bases form hydrogen bonds with pyrimidine.

DNA: The Basis of Life

- Deoxyribonucleic Acid (DNA)
  - Double stranded with complementary strands A-T, C-G
- DNA is a polymer
  - Sugar-Phosphate-Base
  - Bases held together by H bonding to the opposite strand

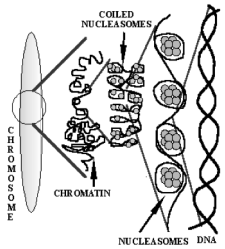


Hepatitis delta virus complete genome

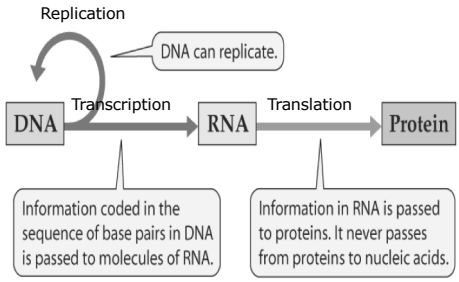
1 atgagccaag ttccgaacaa ggattcgcgg ggaggataga tcagcgcccg agaggggtga
61 gtcggtaaag agcattggaa cgtcggagat acaactccca agaaggaaaa aagagaaagc
121 aagaagcgga tgaatttccc cataacgcca gtgaaactct aggaagggga aagagggaag
181 gtggaagaga aggaggcggg cctcccgatc cgaggggccc ggcggccaag tttggaggac
241 actccggccc gaaggttga gagtacccca gaggaggaga gccacacgga gtagaacaga
301 gaaatcacct ccagaggacc ccttcagcga acagagagcg catcgcgaga gggagtagac
361 catagcgata ggaggggatg ctaggagttg ggggagaccg aagcgaggag gaaagcaaag
421 agagcagcgg ggctagcagg tgggtgttcc gccccccgag aggggacgag tgaggcttat
481 cccgggaac tcgacttatc gtccccacat agcagactcc cggacccct ttcaaagtga
541 ccgaggggg tgactttgaa cattgggac cagtggagcc atggatgct cctcccgatt
601 ccgcccaagc tccttcccc caagggtcgc ccaggaatgg cgggacccca ctctgcaggg
661 tccgcgttcc atcttttctt acctgatggc cggcatggtc ccagcctcct cgctggcgcc
721 ggctgggcaa cattccgagg ggaccgtccc ctcggtaatg gcgaatggga cccacaaatc
781 tctctagctt cccagagaga agcgagagaa aagtggctct cccttagcca tccgagtgga
841 cgtgcgtcct ccttcggatg cccaggtcgg accgcgagga ggtggagatg ccatgccgac
901 ccgaagagga aagaaggacg cgagacgcaa acctgcgagt ggaaacccgc tttattcact
961 ggggtcgaca actctgggga gaggagggag ggtcggctgg gaagagtata tcctatggga
1021 atccctggct tcccctatg tccagtccct ccccggtccg agtaaagggg gactccggga
1081 ctccttgcat gctggggacg aagccgcccc cgggcgctcc cctcgttcca ccttcgaggg
1141 ggttcacacc cccaacctgc gggccggcta ttcttctttc ccttctctcg tcttcctcgg
1201 tcaacctcct aagttcctct tcctcctcct tgctgaggtt ctttcccccc gccgatagct
1261 gctttctctt gttctcgagg gccttccttc gtcggtgatc ctgcctctcc ttgtcggtga
1321 atcctccccct ggaaggcctc ttcctaggtc cggagtctac ttccatctgg tccgttcggg
1381 cccttctcgc cggggagcc ccctctccat ccttatcttt ctttccgaga attccttga
1441 tgtttcccag ccagggatgt tcatcctcaa gtttcttgat tttcttctta accttccgga
1501 ggtctctctc gagttcctct aacttctttc ttccgctcac ccactgctcg agaacctctt
1561 ctctcccccc gcggtttttc cttcttcgg gccggctcat cttcgactag aggcgacggt
1621 cctcagtact cttactcttt tctgtaaaga ggagactgct ggcctgtcg cccaagttcg
1681 ag

DNA: The Basis of Life

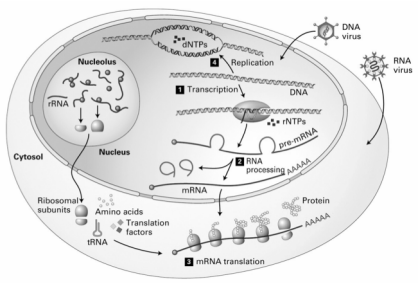
- Humans have about 3 billion base pairs.
  - How do you package it into a cell?
  - How does the cell know where in the highly packed DNA to start transcription?
    - Special regulatory sequences
  - DNA size does not mean more complex
- Complexity of DNA
  - Eukaryotic genomes consist of variable amounts of DNA
    - Single Copy or Unique DNA
    - Highly Repetitive DNA



DNA, RNA, and the Flow of Information



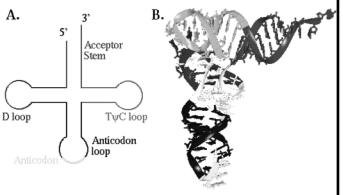
Central dogma of biology: DNA→RNA→Protein



- A gene is expressed in two steps
  1) Transcription: RNA synthesis
  2) Translation: Protein synthesis

RNA

- RNA is similar to DNA chemically. It is usually only a single strand. T(hyamine) is replaced by U(racil)
- Some forms of RNA can form secondary structures by "pairing up" with itself. This can change its properties dramatically.
- DNA and RNA can pair with each other.



tRNA A. linear and B. 3D view

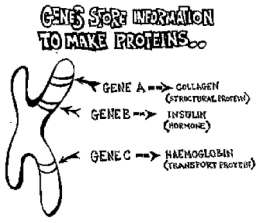
http://www.cgl.ucsf.edu/home/glasfeld/tutorial/trna/trna.gif

Gene transcription

- Transcription is highly regulated. Most DNA is in a dense form where it cannot be transcribed.
- To begin transcription requires a promoter, a small specific sequence of DNA to which polymerase can bind (~40 base pairs "upstream" of gene)
- Finding these promoter regions is a partially solved problem that is related to motif finding.
- There can also be repressors and inhibitors acting in various ways to stop transcription. This makes regulation of gene transcription complex to understand.

Genes Make Proteins

- genome-> genes ->protein(forms cellular structural & life functional)->pathways & physiology



Proteins: Workhorses of the Cell

- 20 different **amino acids**
  - different chemical properties cause the protein chains to fold up into specific three-dimensional structures that define their particular functions in the cell.
- Proteins do all essential work for the cell
  - build cellular structures
  - digest nutrients
  - execute metabolic functions
  - Mediate information flow within a cell and among cellular communities.
- Proteins work together with other proteins or nucleic acids as "molecular machines"
  - structures that fit together and function in highly specific, lock-and-key ways.

Proteins

- Complex organic molecules made up of amino acid subunits
- 20* different kinds of amino acids. Each has a 1 and 3 letter abbreviation.
- http://www.indstate.edu/thcme/mwking/amino-acids.html for complete list of chemical structures and abbreviations.
- Proteins are often enzymes that catalyze reactions.
- Also called "poly-peptides"

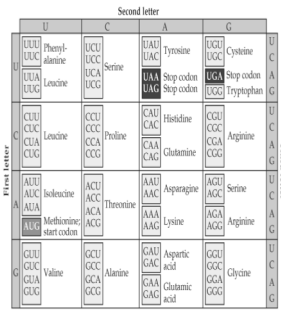
*Some other amino acids exist but not in humans.

Uncovering the code

- Scientists conjectured that proteins came from DNA; but how did DNA code for proteins?
- If one nucleotide codes for one amino acid, then there'd be 4^1 amino acids
- However, there are 20 amino acids, so at least 3 bases codes for one amino acid, since $4^2 = 16$ and $4^3 = 64$
  - This triplet of bases is called a "codon"
  - 64 different codons and only 20 amino acids means that the coding is degenerate: more than one codon sequence code for the same amino acid

Cell Information: Instruction book of Life

- DNA, RNA, and Proteins are examples of strings written in either the four-letter nucleotide of DNA and RNA (A C G T/U)
- or the twenty-letter amino acid of proteins. Each amino acid is coded by 3 nucleotides called codon. (Leu, Arg, Met, etc.)



MUtAsHONS

- The DNA can be thought of as a sequence of the nucleotides: C,A,G, or T.
- What happens to genes when the DNA sequence is mutated?

Normal DNA sequence: ATCTAG

↓

Mutated DNA sequence: ATCGAG

The Good, the Bad, and the Silent

- Mutations can serve the organism in three ways:
- **The Good :** A mutation can cause a trait that enhances the organism's function: Mutation in the sickle cell gene provides resistance to malaria.
- **The Bad :** A mutation can cause a trait that is harmful, sometimes fatal to the organism: Huntington's disease, a symptom of a gene mutation, is a degenerative disease of the nervous system.
- **The Silent:** A mutation can simply cause no difference in the function of the organism.

Campbell, Biology, 5th edition, p. 255

How Do Individuals of a Species Differ?

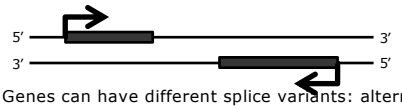
- Genetic makeup of an individual is manifested in traits, which are caused by variations in genes
- While 99.9% of the 3 billion nucleotides in the human genome are the same, small variations can have a large range of phenotypic expressions
- These traits make some more or less susceptible to disease, and the demystification of these mutations will hopefully reveal the truth behind several genetic diseases

The Diversity of Life

- Not only do different species have different genomes, but also different individuals of the same species have different genomes.
- No two individuals of a species are quite the same – this is clear in humans but is also true in every other sexually reproducing species.
- Imagine the difficulty of biologists – sequencing and studying only one genome is not enough because every individual is genetically different!

Genes and alleles

- A gene can have different variants
  - The variants of the same gene are called alleles
- Genes can be found on both strands



- Genes can have different splice variants: alternative splicing

How Do Different Species Differ?

- As many as 99% of human genes are conserved across all mammals
- The functionality of many genes is virtually the same among many organisms
- It is highly unlikely that the same gene with the same function would spontaneously develop among all currently living species
- The theory of evolution suggests all living things evolved from incremental change over millions of years

Mouse and Human overview

- Mouse has 2.1 x10^9 base pairs versus 2.9 x 10^9 in human.
- About 95% of genetic material is shared.
- 99% of genes shared of about 30,000 total.
- The 300 genes that have no homologue in either species deal largely with immunity, detoxification, smell and sex*

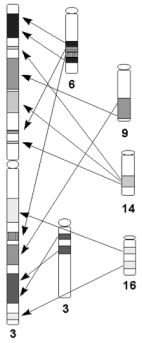
*Scientific American Dec. 5, 2002

Human and Mouse

Significant chromosomal rearranging occurred between the diverging point of humans and mice.

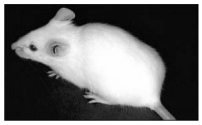
Here is a mapping of human chromosome 3.

It contains homologous sequences to at least 5 mouse chromosomes.



Comparative Genomics

- What can be done with the full Human and Mouse Genome? One possibility is to create "knockout" mice – mice lacking one or more genes. Studying the phenotypes of these mice gives predictions about the function of that gene in both mice and humans.



Causes of variation

- Gene duplications
- Gene insertions
- Mistakes in DNA replication
- Environment agents (e.g., radiation, chemicals)
- Horizontal transfer: genes from another organism
  - Basis of genetic engineering