Biocomputing

Notions of Genetics and Biochemistry

Introduction

- DNA plays the central role in this course
  - Learn how to cut it, lengthen it, measure it, sequence it, measure gentic distances based on it, etc.
- Why is DNA so important ?
  - Responsible for inheritance: encodes all genetic information of an organism
  - Responsible for life processes: encodes all “instructions” needed for the functioning of the organism
  - The most fundamental mechanisms of DNA manipulation are the same in all living organisms

Structure of the lecture

- Take a brief look at some notions of genetics and some famous experiments: *the story of how DNA was proved to be a central element of life*
- Structure of DNA, RNA, proteins
- Central dogma: how instructions in DNA are “executed”
- Genes and chromosomes

Genetics

- Children resemble their parents
  - Although they may look more like one than the other, most of them are a blend of the characteristics of both parents
- Useful traits may be accentuated by controlled mating
  - Speed in horses, strength in oxes, larger fruits, etc. were studied in centuries of breeding of domestic animals and plants
- **Question**: How are traits inherited from one generation to the other ?
  - Story begins with an Augustinian monk named Gregor Mendel, 1865.
  - Aristotle (384-322 BC): inheritance was a physical substance coming from both parents.
    - E.g., the giraffe must be a hybrid of leopard and camel.

Mendel - an abstract notion of gene

- Mendel did experiments of cross-fertilization with peas (1856-1863)
  - Instead of allowing them to self-fertilize, he crossed fertilized different plants among themselves
- Mendel is considered “Father of Genetics”
  - He deduced the mathematical patterns of inheritance, exposed in his three Laws of Heredity (1865)

Laws of Heredity

1. Law of Segregation
   - Each inherited trait is defined by a gene pair. Parental genes are randomly separated to the sex cells so that sex cells contain only one gene of the pair. Offsprings inherit one genetic allele from each parent
2. Law of Independent Assortment
   - The inheritance of one trait is not dependant on the inheritance of another
3. Law of Dominance
   - An organism with alternate forms of a gene will express the form that is dominant

Genes become real things

- Mendel's laws were dismissed until 1900
  - Biologists used to work only with direct observations, rather than abstract notions (such as gene)
  - Technological advances showed that the cells were the basic units of life
  - Intra-cellular structures were observed, including thread-like chromosomes
  - Different organisms may have different numbers of chromosomes – thus, they may carry information specific for each form of life
  - Thus, the abstract notion of gene started having the physical context it lacked before

Chromosomes carry genes

- T.H.Morgan, 1904: studies on the physical basis of heredity
  - Organism of choice: fruit fly
  - Established that genes are physically located on chromosomes
  - E.g., the gene for eye color in fruit flies is located on the X chromosome

Chromosomes carry genes

- Other early examples (for humans)
  - Sex-linked genetic disorders: muscular dystrophy (1913), red-green color blindness (1914), hemophilia (1916)
  - Disorderss related to recessive inheritance: alkaptonuria (1902), albinism (1903)
  - Disorders related to dominant inheritance: brachydactyly (short fingers, 1905), congenital cataracts (1906), Huntington's disease (1913)
  - Eye color inheritance (brown dominant, blue recessive) – 1907
    - Modern view: more genes are involved
  - 1941: one gene makes one protein

DNA is the genetic material

- **Question**: Genes are somewhere on the chromosomes, but where are they really placed ?
- Chromosomes are complex structures consisting of DNA and many kind of proteins

DNA is the genetic material

- DNA was discovered about the same time Mendel and Darwin published their work
  - DNA was known to be a large molecule, but it seemed likely that its four chemical components were assembled in a monotonous pattern
  - The hypothesis was that genes were made of chromosomal proteins and DNA was a structural polymer that was holding the chromosome together
- 1941-1952: DNA is indeed the genetic material

Proving that DNA is the genetic material

- 1928 - the Pneumococcus bacterium causes pneumonia in mice
  - Some strains of Pneumococcus are highly virulent (S-bacteria) and some are non-virulent (R-bacteria)
  - A virulent bacteria has a polysaccharide coat, while the non-virulent one lacks the coat
  - Injected in mice, the coat protects the bacteria from attack by white blood cells, they multiply, infect, and eventually kill the mouse
  - The presence of the coat is an inherited genetic trait

Proving that DNA is the genetic material

- Experiments in 1944 show that genes are made of DNA
  - Living S-bacteria injected in mouse: mouse dies
  - Living R-bacteria: mouse lives
  - Heat-killed S-bacteria: mouse lives
  - Heat-killed S-bacteria + living R-bacteria: mouse dies !!!
    - Living S-bacteria isolated in the blood of the dead mouse
    - These bacteria were fully functional, they could replaicate and infect other mice

The transforming principle

- Conclusion: some chemical called *transforming principle* was responsible for conferring new genetic trait on non-virulent bacteria
- Further experiments: that chemical turned out to be DNA
- The concept of DNA as genetic material received confirmation in another series of experiments in 1952

Structure of DNA

- DNA is composed of building blocks called nucleotides
  - A nucleotide consists of a deoxiribose sugar, a phosphate group and one of the four possible bases
  - The bases are: adenine (A), thymine (T), guanine (G), cytosine (C)
  - Sugar is a pentose (5 carbon atoms): the phosphate is attached to the 5'-carbon, the base to the 1'-carbon
  - Phosphates and sugars of adjacent nucleotides link (through strong phosphodiester bonds) to form a long polymer
  - The ratio of A-to-T and G-to-C constant in all living organisms
  - Final clue provided by X-ray crystalography: DNA is a double helix, shaped like a twisted ladder

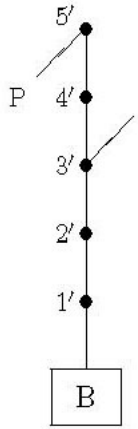


Figure 1.1: A schematic representation of a nucleotide

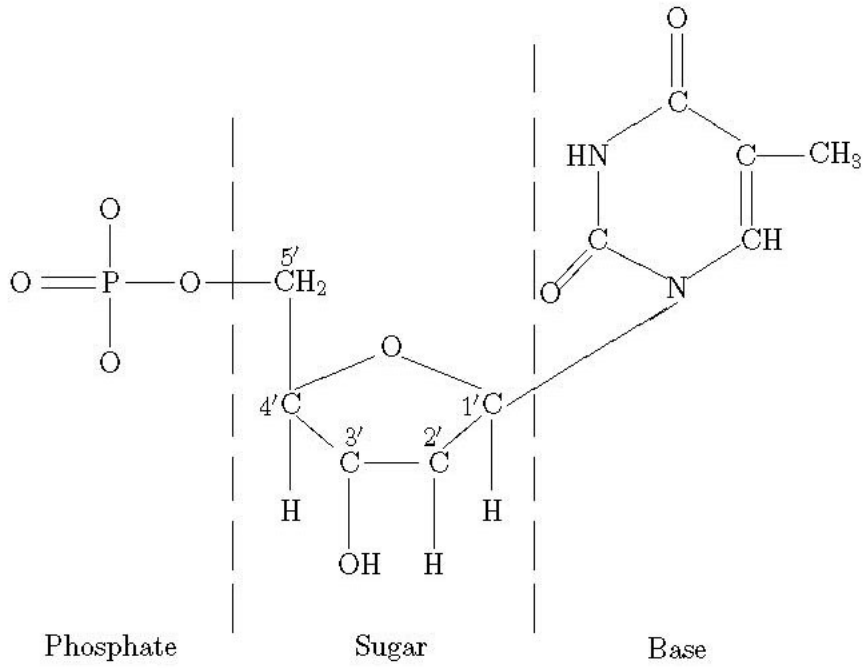


Figure 1.2: The chemical structure of a nucleotide with thymine base

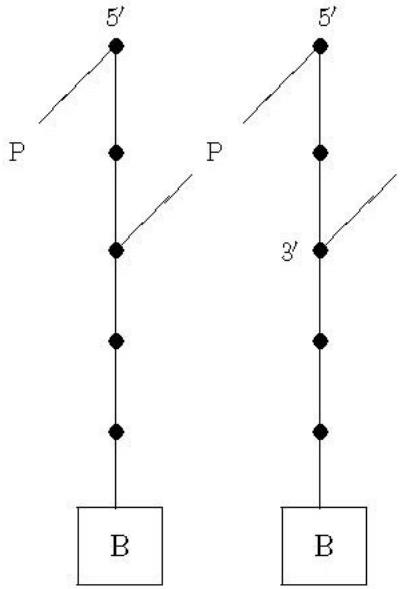


Figure 1.3: Phosphodiester bond

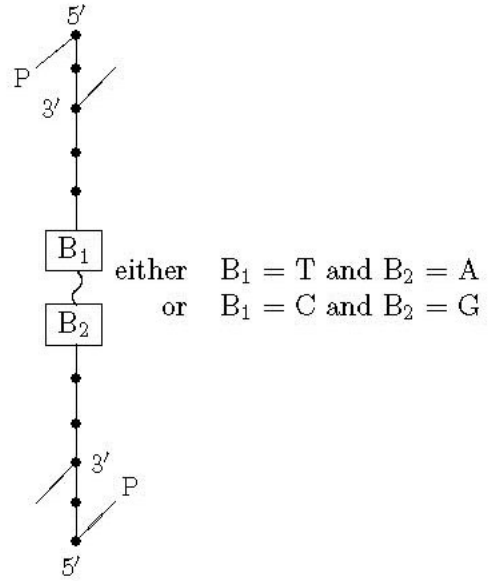


Figure 1.4: Hydrogen bond

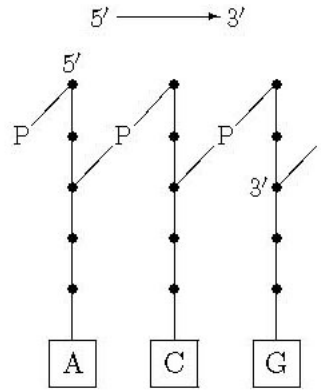


Figure 1.5: Single stranded DNA

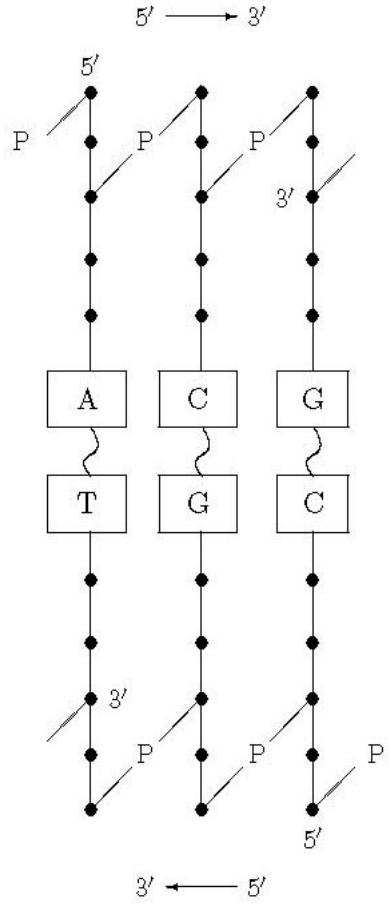


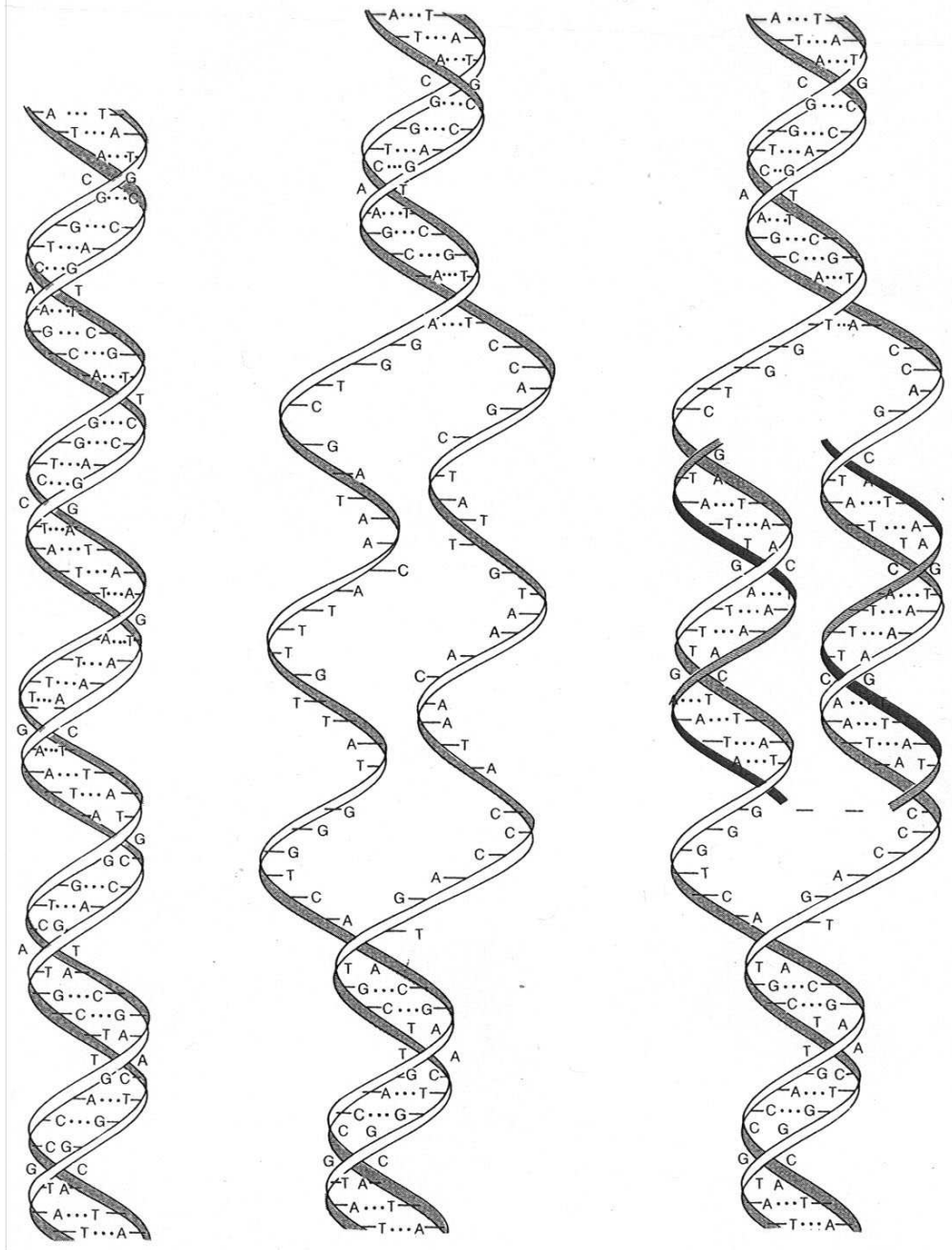
Figure 1.6: Forming double strands

DNA representation

- Single stranded DNA has polarity
  - It has phosphate (attached to the 5'-carbon of the sugar) available for binding at one end of the strand and the 3'-carbon of the sugar available at the other end: the 5'-end and the 3'-end
  - Four types of nucleotides (A,T,C,G)
  - One single stranded DNA molecule may be represented as a string over the alphabet {A,T,C,G}, **ALWAYS** written in the 5'-to-3' direction
  - Example: 5'-ATGCTAC-3' (most of the time omit 5' and 3')

Structure of DNA

- 1953: race to describe the 3D structure won by James Watson and Francis Crick
  - Alternate sugar and phosphate molecules form the twisted uprights of the DNA ladder
  - The rugs of the ladder are formed by complementary pairs of bases: A always paired with T, G always paired with C
  - One strands runs in the 5'-3' direction, the other one in the 3'-5' direction



DNA representation

- Representing double stranded DNA molecules: double strings over the alphabet {A,T,C,G}
- Example of one such double string:

5’-ATGTAC-3’

3’-TACATG-5’

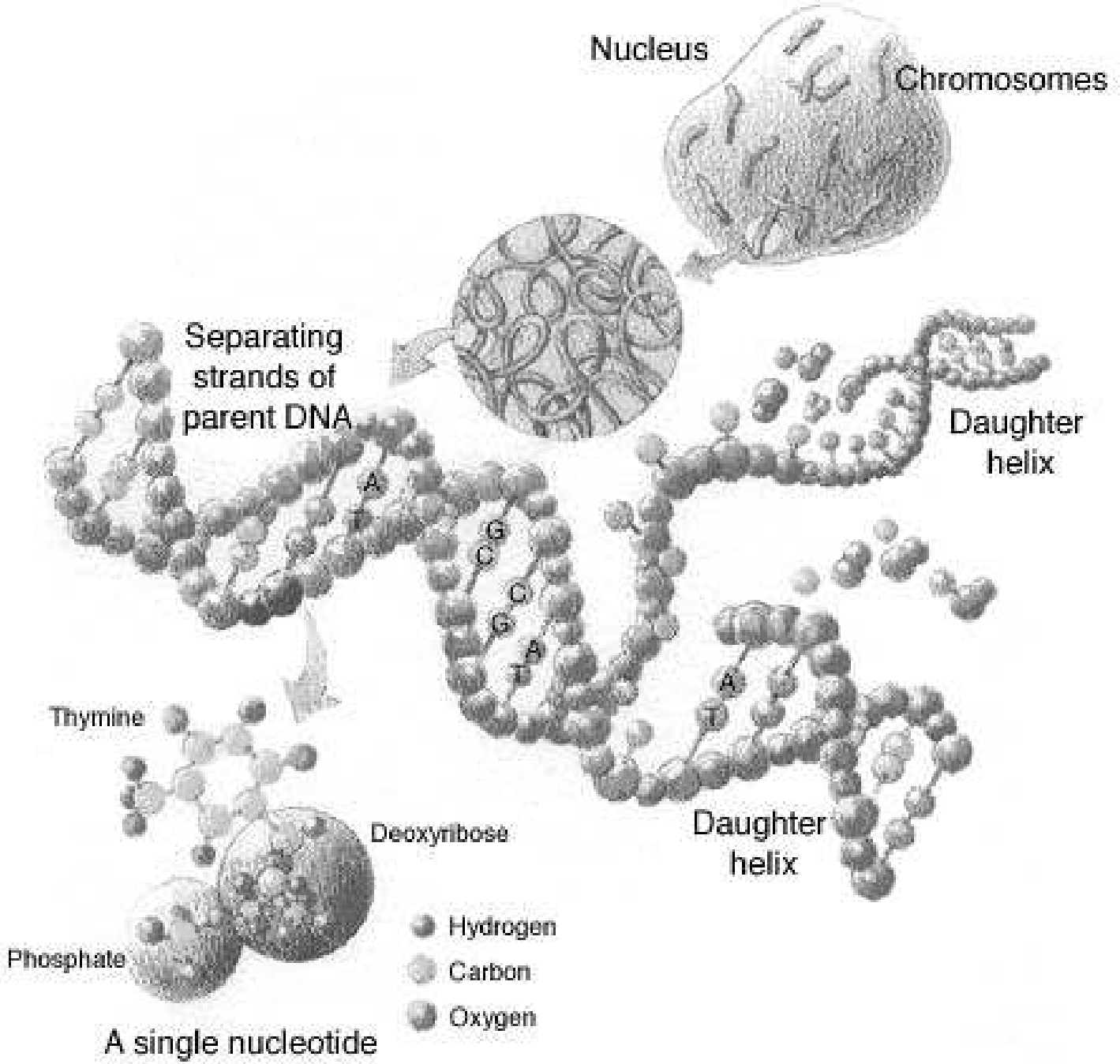
- Most often we omit the 5’ and 3’
- Worth noting that the same molecule is represented also by its “inverse”

5’-GTACAT-3’

3’-CATGTA-5’

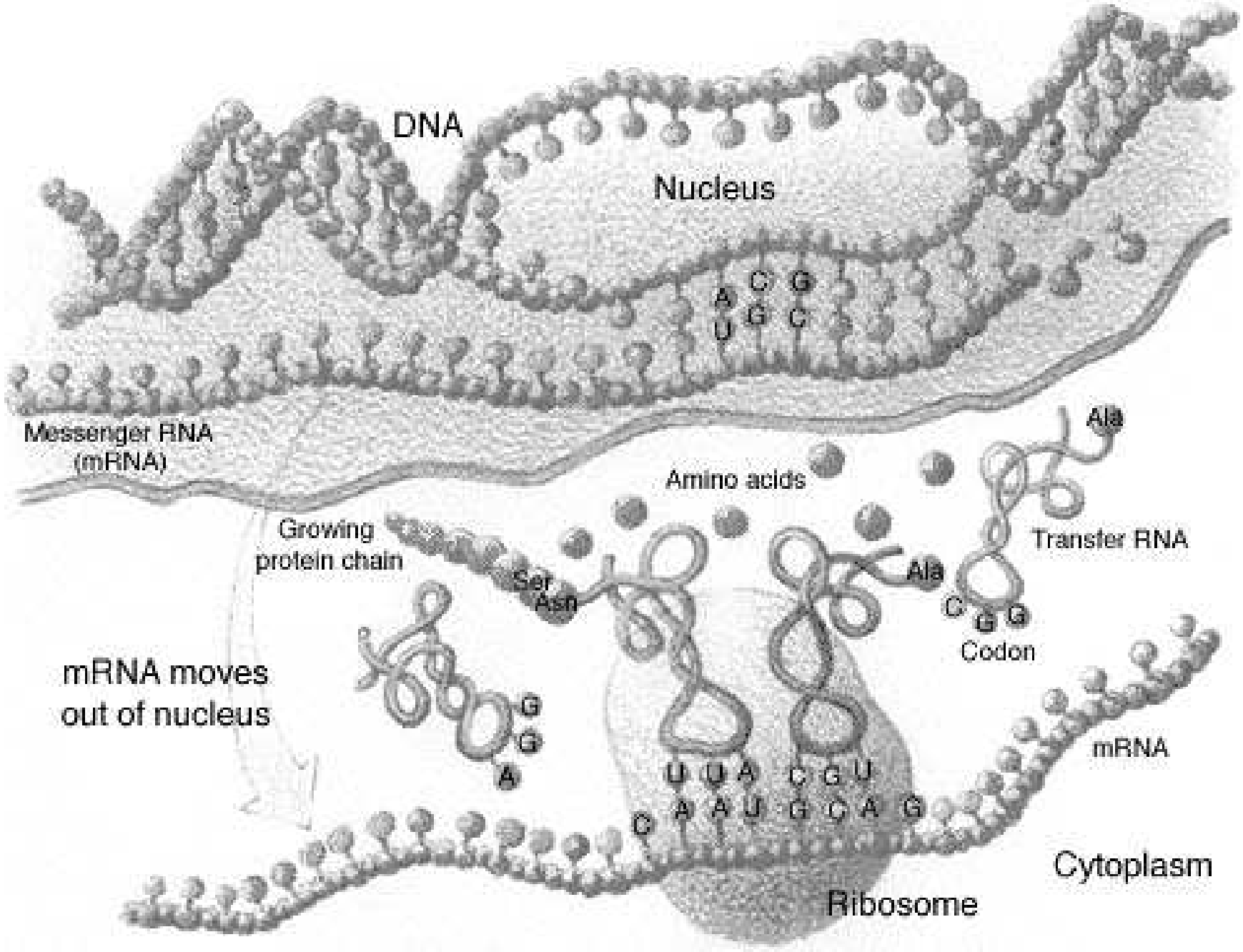
DNA replication

- One half of the DNA ladder serves as a template for recreating the other half during DNA replication
- Responsible for this: the enzyme *DNA polymerase* that adds complementary nucleotides to the template provided by a single stranded DNA molecule
- This was conjectured already by Watson and Crick but the enzyme was actually discovered later


Nucleus
Chromosomes
Separating strands of parent DNA
Daughter helix
Daughter helix
A
T
G
C
C
G
A
T
A
T
Thymine
Deoxyribose
Phosphate
Hydrogen
Carbon
Oxygen
A single nucleotide

Central dogma

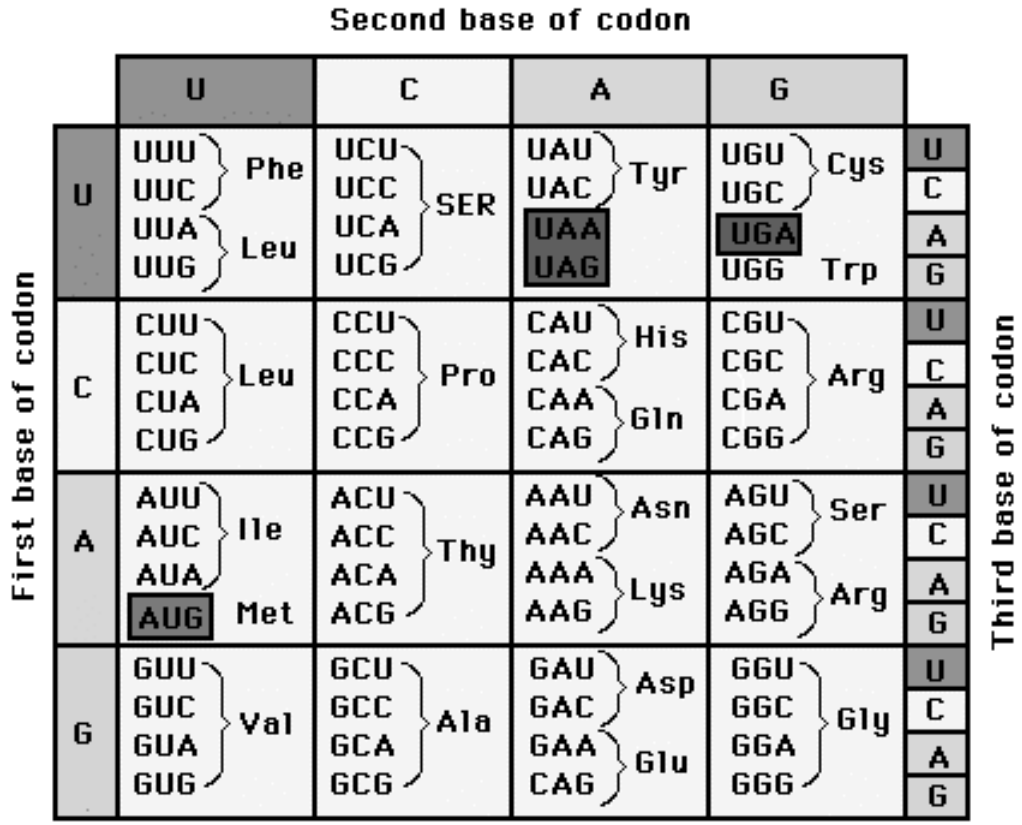
- DNA mostly found in the nucleus (in eukaryotes)
- Another type of nucleic acid commonly found in the cytoplasm: RNA
- RNA copies the DNA message in the nucleus and carries it out to the cytoplasm, where proteins are synthesized (in the rybosomes)
  - DNA transcribed in the nucleus into mRNA
  - mRNA translated in the cytoplasm (rybosomes) into amino acids – tRNA plays the role of "adaptor"


DNA
Nucleus
A
U
C
G
G
C
Messenger RNA
(mRNA)
Amino acids
Ala
Transfer RNA
Growing
protein chain
Ser
Asn
Ala
C
G
G
Codon
mRNA moves
out of nucleus
G
G
A
U
U
A
C
G
U
C
A
A
U
G
C
A
G
mRNA
Cytoplasm
Ribosome

RNA → protein

- From mRNA to protein: the universal genetic code
- Each triplet of nucleotides (codon) specifies one amino acid
- One codon specify beginning of translation (AUG) and 3 codons specify the end of it.

The universal genetic code



The genetic code, written by convention in the form in which the Codons appear in mRNA. The three terminator codons, UAA, UAG, and UGA, are boxed in red; the AUG initiator codon is shown in green.

mRNA → protein

- Translation mRNA → protein: implemented through some "adaptors" that recognize both codons and amino acids: transfer RNA (tRNA)
- On one part the tRNA holds an anticodon and on the other side it holds the corresponding amino acid
- Ordering the tRNA molecules on mRNA is complex: using the ribosome – "large protein synthesizing machine"

Genes and chromosomes

- Each DNA molecule is packaged in a separate chromosome
- The total genetic information stored in the chromosomes: the *genome*
- Eukaryotes: (almost) every single cell contains a complete set of the genome
  - Difference in functionality: different expression of the corresponding genes

Coding and non-coding DNA

- Bacteria: virtually all DNA encodes proteins
- Eukaryotic DNA is composed of repeated sequences that do not encode proteins: non-coding sequences (junk DNA)
  - They separate relatively infrequent “islands” of genes
  - Many non-coding sequences (introns) are found also within the genes
  - Less than 5% of the human genome encodes proteins
  - Human genome: about 3 billion bp
    - Amoeba: 200 times more DNA than humans

Chromosomes

- Chromsomes are dynamic, changing structures
  - DNA may “jump” from one position on a chromosome to another (B.McClintock, Nobel prize)
  - Genes can be turned on and off
    - Regulatory genes produce inhibitors for other genes
    - There are thus both genetic “plans” for protein production and genetic regulatory programs for expressing those plans
  - Genes can be moved between species based on the universality of the genetic code
    - Different bacteria may get antibiotic-resistant genes by exchanging plasmids (small chromosomes)
    - Human insulin produced in E.coli bacteria (1980s)
    - Genetic engineering – exploding field nowadays

Living things share common genes

- Compelling evidence for shared ancestry of all living things
- Evolution of higher life forms require development of new genes, while retaining the old ones
- Genes may be “stolen” from other organisms
- Mutations may accumulate – history of the evolutionary life of a gene

Beyond DNA

- Passive role for DNA: encodes and transmits genetic info through time
- Active role for proteins encoded by DNA
- Post-genomic phase: what is the exact role of the proteins made by these specific genes?
- Animal of choice for test: mice
  - They share about 99% of the genes with humans

Acknowledgments

This lecture is based on materials developed by Cold Spring Harbor Laboratory