Foundation of Bio-Informatics
Lecture 1

Lecture by: Ahmad R. Naghsh-Nilchi, PhD
Department of Artificial Intelligence
Faculty of Computer Engineering
University of Isfahan

What is Bioinformatics

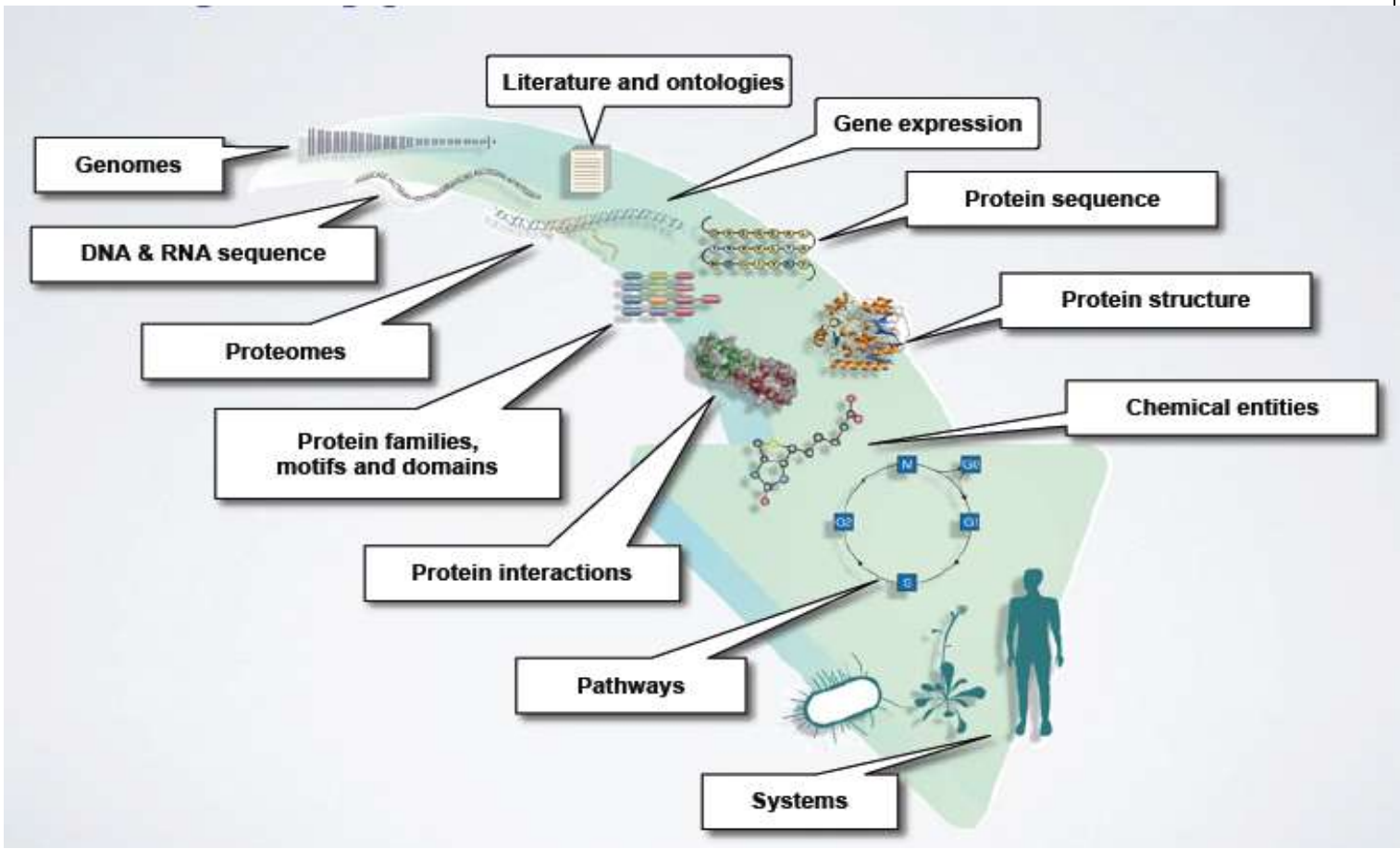
- *"Bioinformatics is conceptualizing biology in terms of **macromolecules** and then applying **"informatics" techniques** (derived from disciplines such as applied math, computer science, and statistics) to **understand** and **organize** the information associated with these molecules, on a **large-scale**".*
- "*Bioinformatics is the application of computers to the collection, archiving, organization, and analysis of biological data.*"

More Definitions …

- *"Bioinformatics is research, development, or application of* ***computational approaches*** *for expanding the use of* ***biological***, ***medical***, ***behavioral*** *or* ***health data***, *including those to* ***acquire***, ***store***, ***organize*** *and* ***analyze*** *such data." National Institutes of Health (NIH)*

- ***Bioinformatics is computer aided biology!***

Major Types of Bioinformatics Data



Where Bioinformatics Comes from?

- As molecular biology began to be transformed by of the emergence of molecular sequence and structural data, bioinformatics arises.

What is the Goal of Bioinformatics?

- Integrate sequence, 3D structure, expression pattern, interaction and function of bio-molecules to gain a deeper understanding of biological mechanism, process and systems.

- Aims to bridge the gap between data and knowledge

Why do we need Bioinformatics?

- Bioinformatics is necessary because of the rapid expansion in quantity and in complexity of bimolecular data.
- Bioinformatics provide methods for the efficient storage, annotation, search & retrieval, data integration, data mining, and data analysis.

Analyzing Protein Sequences

- All proteins are made up of the same basic building block called “amino aside”.
- Amino Aside are complex organic molecules, made of carbon, hydrogen, oxygen, nitrogen, and sulfur atoms.
- Proteins are huge molecules (*macromolecules*) made of a large number of amino asides (typically between 100 to 500) picked out of 20 flavors.

Table 1-1 The 20 Amino Acids and Their Official Codes			
#	*1-Letter Code*	*3-Letter Code*	*Name*
1	A	Ala	Alanine
2	R	Arg	Arginine
3	N	Asn	Asparagine
4	D	Asp	Aspartic acid
5	C	Cys	Cysteine
6	Q	Gln	Glutamine
7	E	Glu	Glutamic acid
8	G	Gly	Glycine
9	H	His	Histidine
10	I	Ile	Isoleucine
11	L	Leu	Leucine
12	K	Lys	Lysine
13	M	Met	Methionine
14	F	Phe	Phenylalanine
15	P	Pro	Proline
16	S	Ser	Serine
17	T	Thr	Threonine
18	W	Trp	Tryptophan
19	Y	Tyr	Tyrosine
20	V	Val	Valine

Protein Formula:

- Any given type of protein (such as insulin) always contains exactly the same number of total amino asides (*residues*) in the same proportion. Thus a good formula for protein looks like

Insulin =

(30 glycines + 44 alanines + 5 tyrosines + 14 glutamines + …)

- These amino asides are linked together as a chain.

Protein Formula continued

- That is, true identity of a protein is derived not only from its composition, but also from precise order of its amino asides.
- The first amino aside sequence of a protein – insulin – was determined in 1951, following a chain of 110 residues:

insulin = MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERG
FFYTPKTRREAEDLQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLY
QLENYCN

Historical Steps:

- Alfred Sanger won his first Nobel Prize by identifying the human insulin sequence.
- This is the start of the modern era of molecular and structural biology.
- In 50s and 60s (Pre computer era), sequences were assembled, analyzed and compared by writing them on piece of a paper (manually).
- After introducing x8086-based PCs, these sequences were logged into the computer memory banks.

Historical steps continued

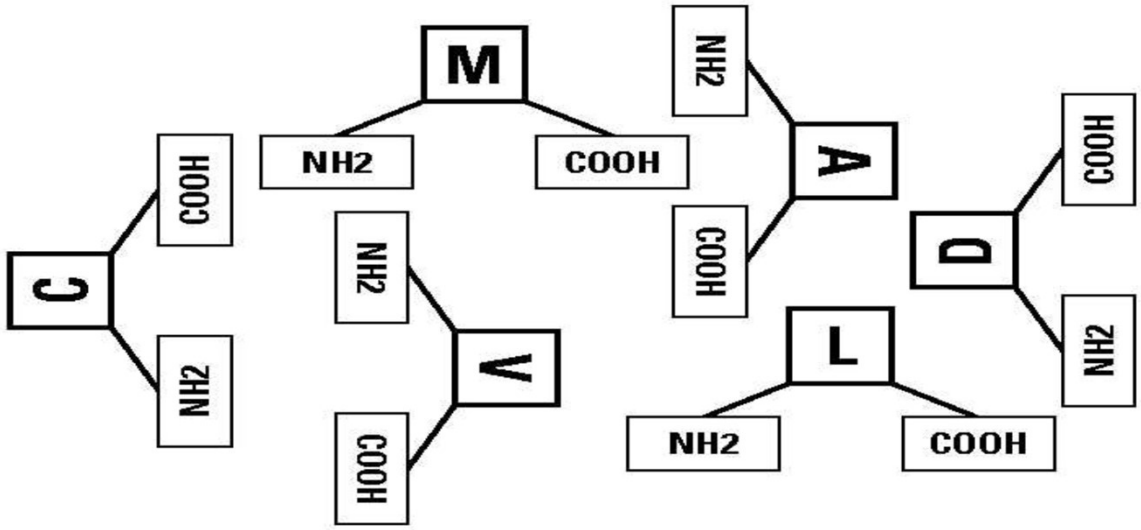
- There after, all the old manual techniques of analyzing protein had to be changed.
- The analysis of protein using computer born.
- This was the genesis of bioinformatics.

A Protein Sample Sequence

:"MKKIMLVFITLILVSLPIAQQTEAKDASAFNKENSISSMAPPAS
PPASPKTPIEKKHADEIDKYIQGLDYNKNNVLVYHGDAVTNVPPRKGYKDGNEYIVVE
KKKKSINQNNADIQVVNAISSLTYPGALVKANSELVENQPDVLPVKRDSLTLSIDLPG
MTNQDNKIVVKNATKSNVNNAVNTLVERWNEKYAQAYPNVSAKIDYDDEMAYSESQLI
AKFGTAFKAVNNSLNVNFGAISEGKMQEEVISFKQIYYNVNVNEPTRPSRFFGKAVTK
EQLQALGVNAENPPAYISSVAYGRQVYLKLSTNSHSTKVKAAFDAAVSGKSVSGDVEL
TNIIKNSSFKAVIYGGSAKDEVQIIDGNLGDLRDILKKGATFNRETPGVPIAYTTNFL
KDNELAVIKNNSEYIETTSKAYTDGKINIDHSGGYVAQFNISWDEVNYDPEGNEIVQH
KNWSENNKSKLAHFTSSIYLPGNARNINVYAKECTGLAWEWWRTVIDDRNLPLVKNRN
ISIWGTTLYPKYSNKVDNPIE"

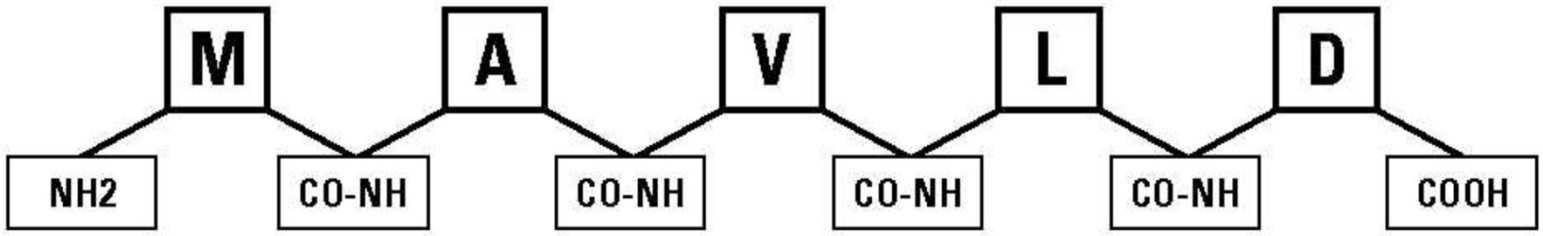
Protein Sequences from H to C

- All twenty amino aside molecules have the same hooks: NH2 – COOH
- These two atoms form the *peptide bonds* between the successive residues.



Protein Sequence or Fragment

- Example of forming a protein sequence petide bonds with N-C:



- The un-used NH2 and COOH ends are called *N-terminus* and *C-terminus* of the chain.
- The sequence: MAVLD = Met-Ala-Val-Leu-Asp

Protein 3-D Structure

- The previous slide sequence of amino aside is not give the protein's biological properties (i.e. sugar digestion or muscle fiber). Its 3-D structure could.
- The final 3-D shape of the protein molecule is uniquely dictated by its sequence.
- The first 3-D structure of a protein was determined in 1958 by Drs. Kendrew and Perutz using X-ray crystallography, winning Nobel Prize for their foundlings.

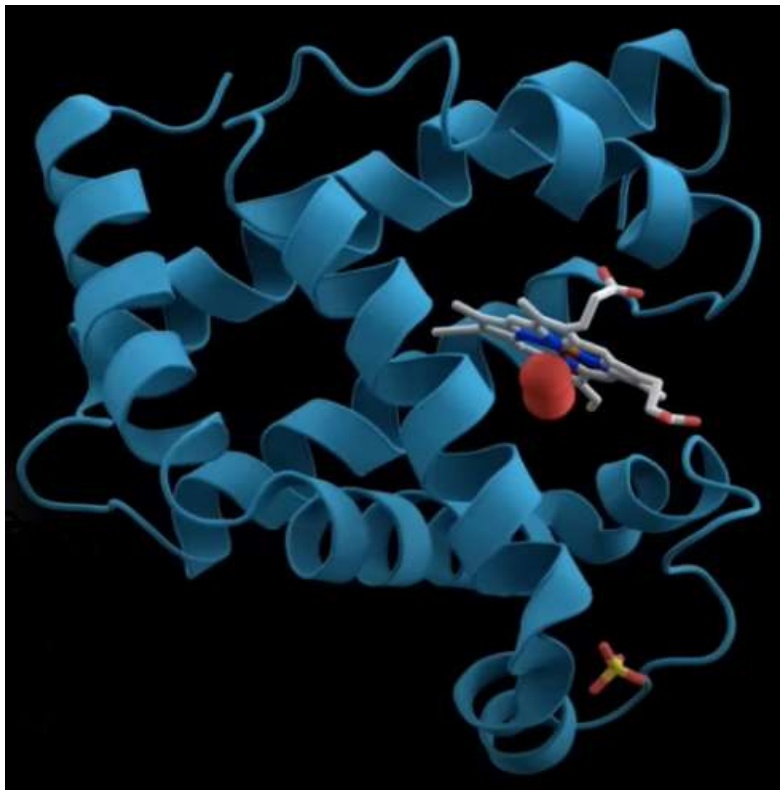
A 3-D shape sample

- This made us realize the proteins' precise and specific shapes.

- Myoglobin Protein

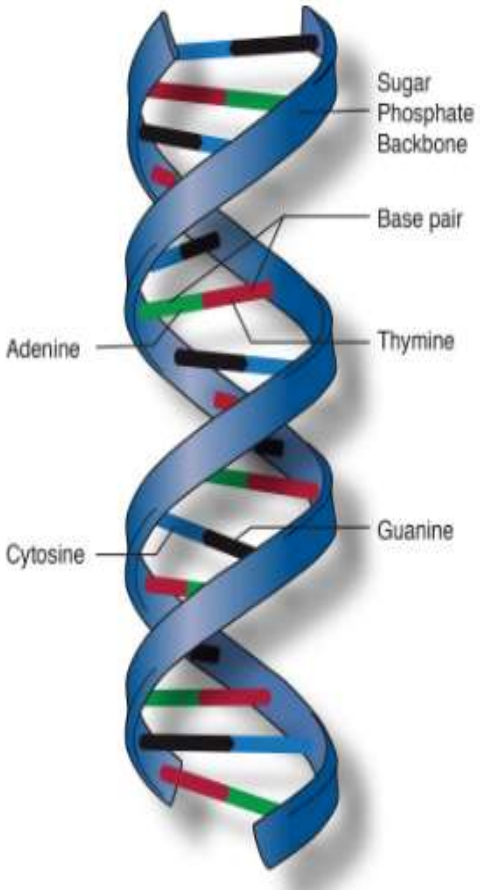
Computer Generated

3-D shape →



DNA Sequence

- In the 1950s, while Kendraw and Perutz were focused on the 3-D structure of proteins, concurrent research by biologists established that *deoxyribonucleic acid* (DNA) was a large macromolecule, characterized by its long, chain-like structure twisted into a double helix.



Analyzing DNA

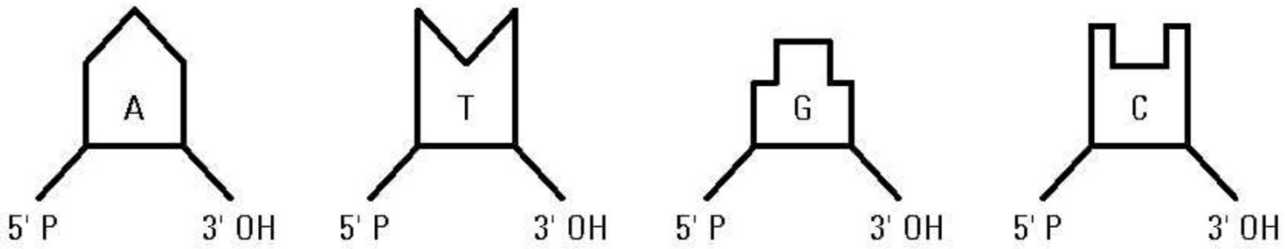
- Each link in the DNA chain pairs two out of 4 *nucleotides*.
- A nucleotide is made of one phosphate group linked to a pentose sugar, which is itself linked to one of 4 types of *nitrogenous* organic bases symbolized by four letters A, C, G, and T.
- Determination and getting direct access to DNA sequences happened until the 1970s by A. Sanger, earning 2nd Nobel Prize for that.
- This was a revolution because of the small DNA sequence (4 nucleotides, as compared to 20 amino aside), allowed a much simpler and faster reading.

A DNA Sequence Sample

```
TATTGAATTTTCAAAAATTCTTACTTTTTTTTTGGATGGACGCAAAGAAGTTTAATAATCATATTACATGGCATTACCACCATAT
ATCCATATCTAATCTTACTTATATGTTGTGGAAATGTAAAGAGCCCCATTATCTTAGCCTAAAAAAACCTTCTCTTTGGAACTTT
AATACGCTTAACTGCTCATTGCTATATTGAAGTACGGATTAGAAGCCGCCGAGCGGGCGACAGCCCTCCGACGGAAGACTCTCCT
GCGTCCTCGTCTTCACCGGTCGCGTTCCTGAAACGCAGATGTGCCTCGCGCCGCACTGCTCCGAACAATAAAGATTCTACAATAC
TTTTATGGTTATGAAGAGGAAAAATTGGCAGTAACCTGGCCCCACAAACCTTCAAATTAACGAATCAAATTAACAACCATAGGAT
ATGCGATTAGTTTTTTAGCCTTATTTCTGGGGTAATTAATCAGCGAAGCGATGATTTTTGATCTATTAACAGATATATAAATGGA
CTGCATAACCACTTTAACTAATACTTTCAACATTTTCAGTTTGTATTACTTCTTATTCAAATGTCATAAAAGTATCAACAAAAAA
TAATATACCTCTATACTTTAACGTCAAGGAGAAAAAACTATAATGACTAAATCTCATTCAGAAGAAGTGATTGTACCTGAGTTCA
TAGCGCAAAGGAATTACCAAGACCATTGGCCGAAAAGTGCCCGAGCATAATTAAGAAATTTATAAGCGCTTATGATGCTAAACCG
TTGTTGCTAGATCGCCTGGTAGAGTCAATCTAATTGGTGAACATATTGATTATTGTGACTTCTCGGTTTTACCTTTAGCTATTGA
GATATGCTTTGCGCCGTCAAAGTTTTGAACGAGAAAAATCCATCCATTACCTTAATAAATGCTGATCCCAAATTTGCTCAAAGGA
CGATTTGCCGTTGGACGGTTCTTATGTCACAATTGATCCTTCTGTGTCGGACTGGTCTAATTACTTTAAATGTGGTCTCCATGTT
ACTCTTTTCTAAAGAAACTTGCACCGGAAAGGTTTGCCAGTGCTCCTCTGGCCGGGCTGCAAGTCTTCTGTGAGGGTGATGTACC
GGCAGTGGATTGTCTTCTTCGGCCGCATTCATTTGTGCCGTTGCTTTAGCTGTTGTTAAAGCGAATATGGGCCCTGGTTATCATA
CAAGCAAAATTTAATGCGTATTACGGTCGTTGCAGAACATTATGTTGGTGTTAACAATGGCGGTATGGATCAGGCTGCCTCTGTT
GTGAGGAAGATCATGCTCTATACGTTGAGTTCAAACCGCAGTTGAAGGCTACTCCGTTTAAATTTCCGCAATTAAAAAACCATGA
AGCTTTGTTATTGCGAACACCCTTGTTGTATCTAACAAGTTTGAAACCGCCCCAACCAACTATAATTTAAGAGTGGTAGAAGTCA
AGCTGCAAATGTTTTAGCTGCCACGTACGGTGTTGTTTTACTTTCTGGAAAAGAAGGATCGAGCACGAATAAAGGTAATCTAAGA
TCATGAACGTTTATTATGCCAGATATCACAACATTTCCACACCCTGGAACGGCGATATTGAATCCGGCATCGAACGGTTAACAAA
CTAGTACTAGTTGAAGAGTCTCTCGCCAATAAGAAACAGGGCTTTAGTGTTGACGATGTCGCACAATCCTTGAATTGTTCTCGCG
ATTCACAAGAGACTACTTAACAACATCTCCAGTGAGATTTCAAGTCTTAAAGCTATATCAGAGGGCTAAGCATGTGTATTCTGAA
TAAGAGTCTTGAAGGCTGTGAAATTAATGACTACAGCGAGCTTTACTGCCGACGAAGACTTTTTCAAGCAATTTGGTGCCTTGAT
GAGTCTCAAGCTTCTTGCGATAAACTTTACGAATGTTCTTGTCCAGAGATTGACAAAATTTGTTCCATTGCTTTGTCAAATGGAT
TGGTTCCCGTTTGACCGGAGCTGGCTGGGGTGGTTGTACTGTTCACTTGGTTCCAGGGGGCCCAAATGGCAACATAGAAAAGGTA
AAGCCCTTGCCAATGAGTTCTACAAGGTCAAGTACCCTAAGATCACTGATGCTGAGCTAGAAAATGCTATCATCGTCTCTAAACC
TTGGGCAGCTGTCTATATGAATTATAAGTATACTTCTTTTTTTTACTTTGTTCAGAACAACTTCTCATTTTTTTCTACTCATAAC
GCATCACAAAATACGCAATAATAACGAGTAGTAACACTTTTATAGTTCATACATGCTTCAACTACTTAATAAATGATTGTATGAT
TTTTCAATGTAAGAGATTTCGATTATCCACAAACTTTAAAACACAGGGACAAAATTCTTGATATGCTTTCAACCGCTGCGTTTTG
CCTATTCTTGACATGATATGACTACCATTTTGTTATTGTACGTGGGGCAGTTGACGTCTTATCATATGTCAAAGTCATTTGCGAA
TTGGCAAGTTGCCAACTGACGAGATGCAGTAAAAAGAGATTGCCGTCTTGAAACTTTTTGTCCTTTTTTTTTTCCGGGGACTCTA
AACCCTTTGTCCTACTGATTAATTTTGTACTGAATTTGGACAATTCAGATTTTAGTAGACAAGCGCGAGGAGGAAAAGAAATGAC
AAATTCCGATGGACAAGAAGATAGGAAAAAAAAAAAGCTTTCACCGATTTCCTAGACCGGAAAAAAGTCGTATGACATCAGAATG
```

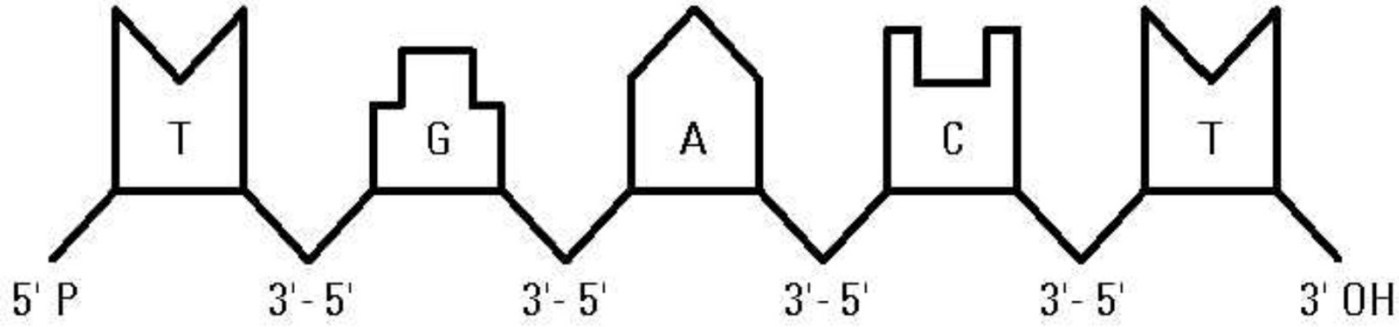
Reading DNA Sequences

- The rate of determining DNA sequence is mega faster than protein sequence.
- Like amino asides in proteins, the 4 nucleotides in DNAs have different bodies, but all have the same pair of hooks:
- 5’ phosphoryl and 3’ hydroxyle (‘ reads prime) in the sugar molecule. Free nucleotides look like:



DNA Molecule

- Forming a bound between the 5’ and 3’ positions of nucleotides makes the DNA molecule.
- A schematic representation of a DNA strand:

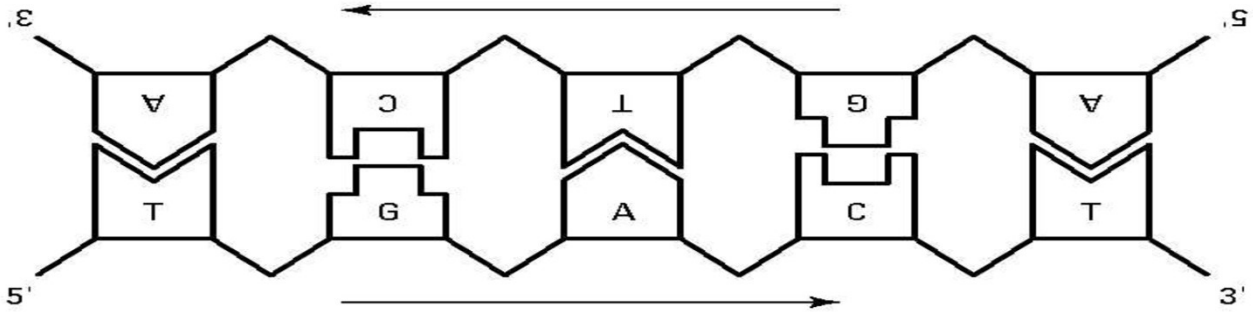


- A DNA sequence is always defined as the succession of its nucleotides listed from the 5’-to-3’ terminus. The alphabetic strand of the above DNA:

  TGACT = thymine-Guanine-Adenine-Cytosine-Thymine

The Two Sides of DNA Sequence

- The DNA molecule consists of two complementary strands, shown below:



- By complementarity, we mean that a thymine (T) on one strand is always facing an adenine (A), and guanine (G) is always facing cytosine (C), and vice versa.
- These couples (A-T) and (G-C), are not linked by a chemical bond but have a strict one-to-one reciprocal relationship.
- When you know the sequence of nucleotides along one strand, you can automatically deduce the sequence on the other one.

Rosetta Stone Explanation
of DNA Sequences

- When living organisms reproduce, each of their genes must be duplicated.
- Nature does not do that like a photocopier!
- Nature separates the DNA strands and makes two complementary ones: Two-sided structures of DNA molecules.
- Example: ATGCAATGAC and GTCATTGCAT are strands of the same DNA molecule.
- Most databases take this property into account (like BLAST), but some do not!

DNA Characteristics

- DNA is the most dignified member of the nucleic acid family.
- DNA sole and only task is to ensure-forever-the conservation of the genetic information for its organism.
- DNA is very stable and resistant and lies well protected in the nucleus of each cell.

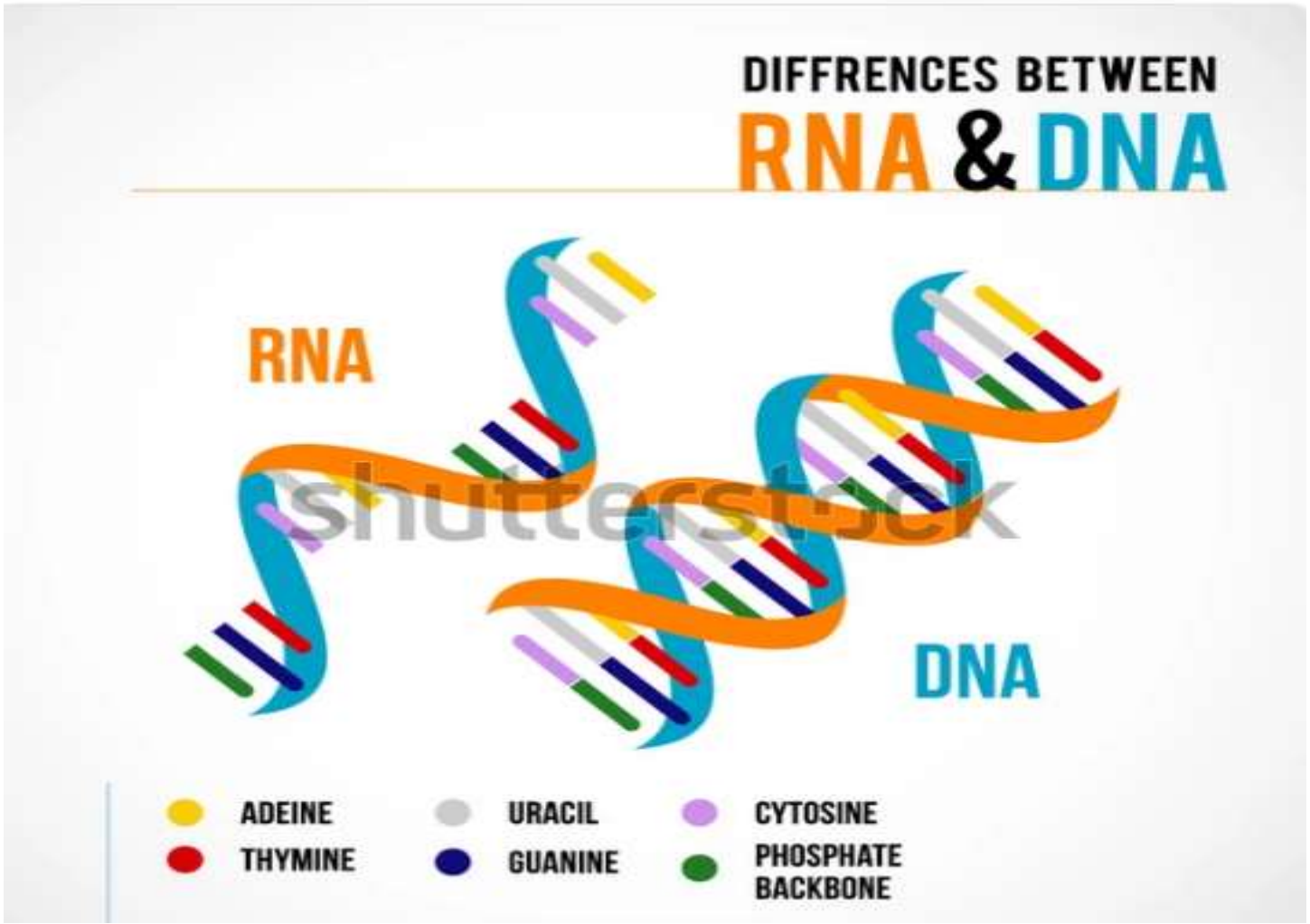
Ribonucleic Acid (RNA)

- RNA is a more active member of the nucleic acid family,
- RNA is synthesized and degraded constantly, making copies of genes available to the cell factory.
- The one-letter code of RNA sequences is:

1-Letter Code	*Nucleotide Base Name*	*Category*
A	Adenine	Purine
C	Cytosine	Pyrimidine
G	Guanine	Purine
U	Uracil	Pyrimidine
N	Any nucleotide	Purine or Pyrimidine
R	A or G	Purine
Y	C or U	Pyrimidine
--	-------	None (gap)

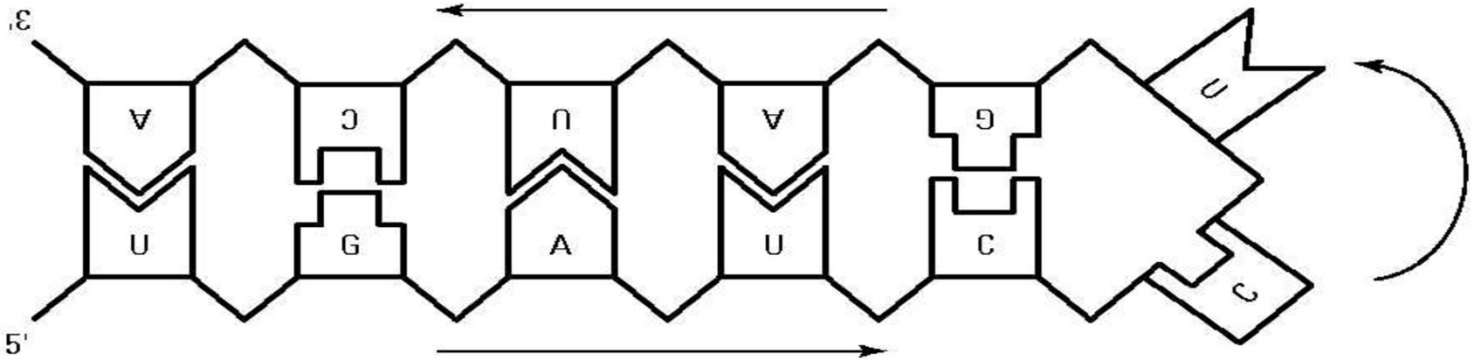
RNA vs. DNA

- In the context of bioinformatics, there are only two important differences between RNA & DNA:
  - RNA differs from DNA by one nucleotide
  - RNA comes as a single strand, not a helix
- Some programs automatically handle the U instead of T conversion, while others don't.
- So, some people work with the sequences of the RNA genes written in DNA rather than RNA sequences.


DIFFRENCES BETWEEN
RNA & DNA
RNA
DNA
shutterstock
ADEINE
THYMINE
URACIL
GUANINE
CYTOSINE
PHOSPHATE BACKBONE

RNA Structure

- Each single stranded RNA molecule - maybe seen as a free-floating piece – sticks with their complementary sequences.
- The following figure shows how RNA turns itself into a double-stranded structure:



Some Notes on RNA

- Basic elements of RNA secondary structure:

  loops (the unpaired C-U in the figure) and

  stems (the paired regions)

- Single-stranded RNA molecules pair different regions of their sequences to form stable double-helical structures, less regular than the double-helical structure of DNA.
- Once synthesized, each RNA molecule quickly adapts a compact fold, trying to pair as many nucleotides as possible, while keeping the chain flexible and true to its geometry.

3-D Shape of RNA

- As with proteins, the linear sequence of the building blocks of RNA dictates the final 3-D shape.
- The biological functions of RNA molecules derive from their 3-D shapes or their sequence complementarity with specific genes.
- Predicting the final fold of an RNA molecule from its sequence is a challenging problem that drove many developments in bioinformatics.

Different Words and Abbreviations?!

- Books, Courses, Articles, … use different words and abbreviations to designate the building blocks of nucleic acids:
- "base", "base pair", "nucleoside" and "nucleotide" are used for the term "*nucleotide*", abbreviated *nt (as in "400-nt-long sequence"),* which we used.
- Note also that we say the number of positions rather than the number of nucleotides.
- A 400-nt long DNA molecule has 400 positions for nucleotides, but it contains twice (800) since every position contains a pair of nucleotides.

Thank you

for your attention