Bioinformatics
BS (CS – 460)
Lecture 01

Dr. Hafeez Ur Rehman

National University of Computer & Emerging Sciences, Peshawar.

Course Outline

- Introduction to Bioinformatics
- Molecular Biology
- Biological Databases
- **Sequence Storage**, Retrieval and Analysis
- Sequence Similarity and Homology
- Creating alignments, Local and Global Alignment
- Pair wise and Multiple Sequence Alignments

National University of Computer & Emerging Sciences, Peshawar.

Course Outline (contd…)

- Phylogenetic Analysis

- Dot Matrix Plots

- Dynamic Programming Algorithms

- Word (k-tuple) methods

- Substitution Matrices PAM and BLOSUM

- Significance of scoring and the effect of gap penalties etc.

- Implementation of Research Papers plus Research Work

Consultation and Recommended Book

- Dr. Hafeez Ur Rehman, hafeez.urrehman@nu.edu.pk (please send an email to make an appointment - Room 31)

- Recommended Books:

1. **Bioinformatics Computing** by Bryan Bergeron, Harvard Medical School and MIT Massachusetts, USA.

2. **Molecular Biology of the Cell, 5th Edition:** Bruce Albert, Alexander Johnson and others..published by Garland Science.

3. **Introduction to Bioinformatics** by T K Attwood, D J Parry-Smith, Samiron Phukan, Pearson Education (Latest edition)

What is Bioinformatics?

- Bioinformatics is an interdisciplinary scientific field that develops **methods** and **software tools** for **storing, retrieving, organizing** and **analyzing biological data**.

- "The mathematical, statistical and computing methods that aim to solve biological problems using DNA and amino acid sequences and related information."

National University of Computer & Emerging Sciences, Peshawar.

DNA Sequencing

- **Bioinformatics** is **based on the fact** that **DNA sequencing** is cheap, and becoming easier and cheaper very quickly.

  - The Human Genome Project cost roughly $3 billion and took 12 years (1991-2003).

  - Sequencing James Watson's genome in 2007 cost $2 million and took 2 months

  - Today, you could get your genome sequenced for about $1,000 and it would take few days.
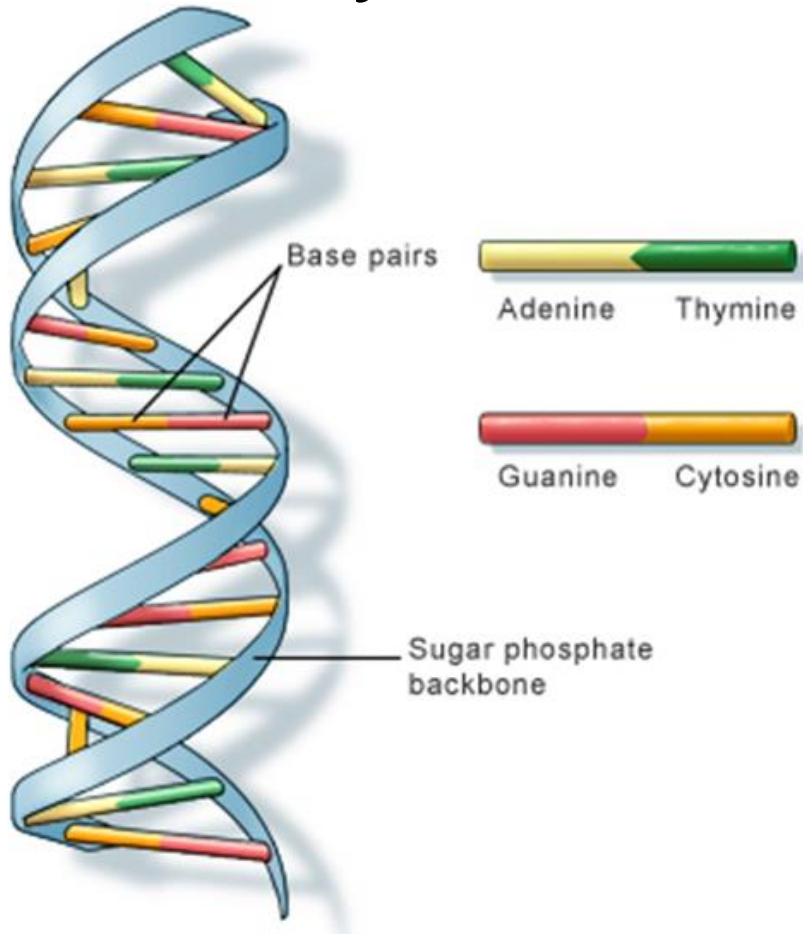
Background: What is DNA?
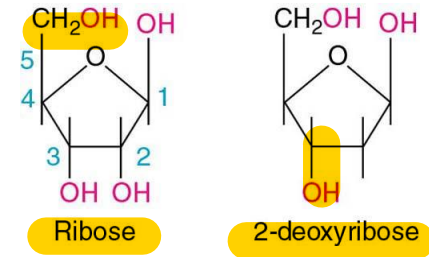
- Deoxyribonucleic acid
- **Blueprint** that carries **genetic information** from one generation to the next
- **Resides** in cell nucleus
- DNA contains *genes*
- Each gene is responsible for the production of a particular *protein*
- A strand of DNA is called a *chromosome.*
- Set of chromosomes carried by an organism is a *Genome.*

DNA Structure
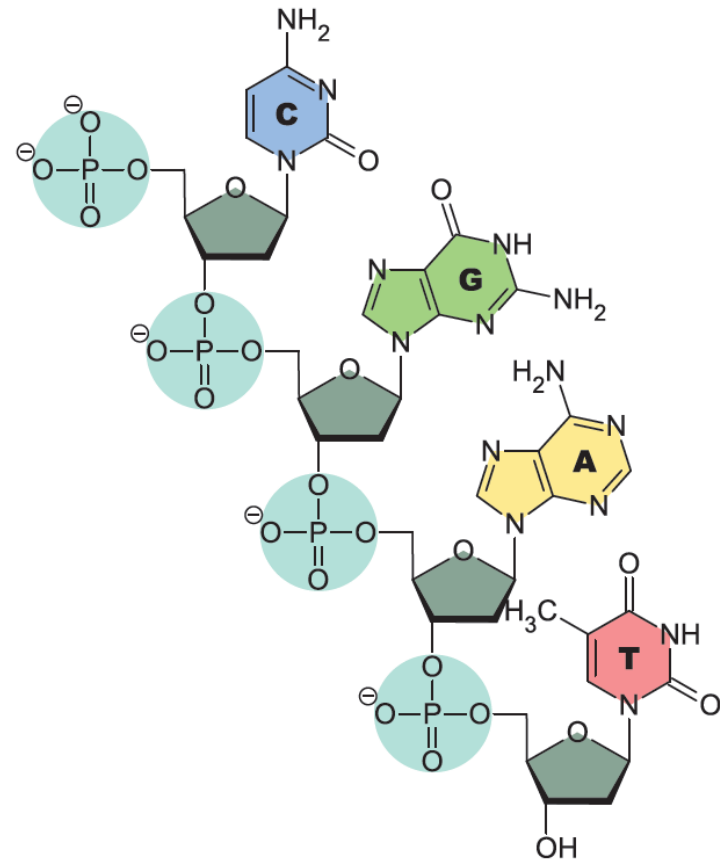
Physical

Chemical
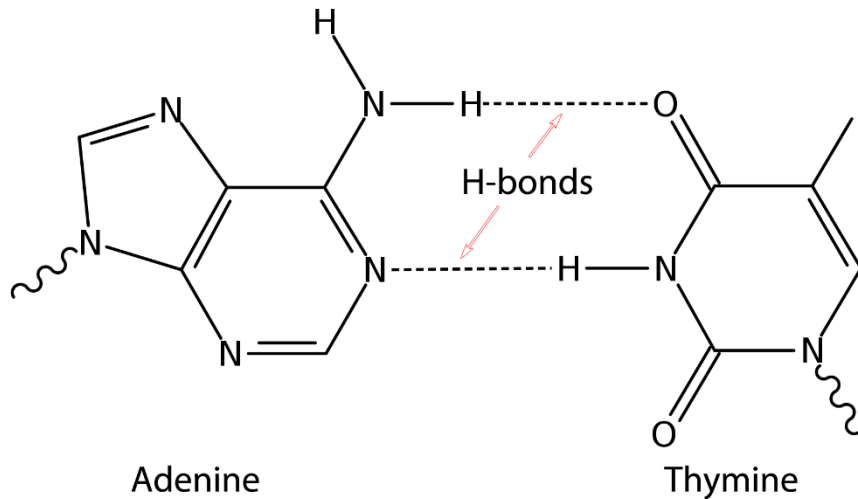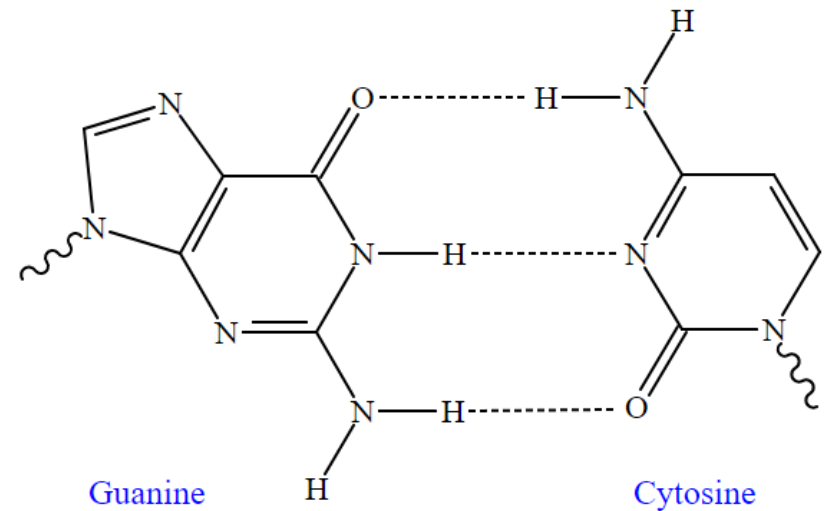
Base pairs

Adenine   Thymine

Guanine   Cytosine

Sugar phosphate backbone

National University of Computer &
Emerging Sciences, Peshawar.

Why specific base pairs bond?

A to T bonding

C to G bonding

Adenine   Thymine

Guanine   Cytosine

National University of Computer &
Emerging Sciences, Peshawar.

DNA Structure

- Double helix

- DNA consists of four nucleotides: Adenine, Cytosine, Guanine, Thymine (A, C, G, T)

- 5 prime (beginning), 3 prime (end)

- Base Pairs: A-T, C-G

- Ordering of these pairs is a *sequence*

- An example:

G A A T T C
C T T A A G

National University of Computer &
Emerging Sciences, Peshawar.

DNA Structure (Directionality)

- **Directionality**, in molecular biology and biochemistry, is the end-to-end chemical orientation of a single strand of nucleic acid.

- The 5′-end (pronounced "five prime end") designates the end of the DNA or RNA strand that has the fifth carbon in the sugar-ring of the deoxyribose or ribose at its terminus.

- The 3′-end of a strand is so named due to it terminating at the hydroxyl group of the third carbon in the sugar-ring, and is known as the *tail end*.
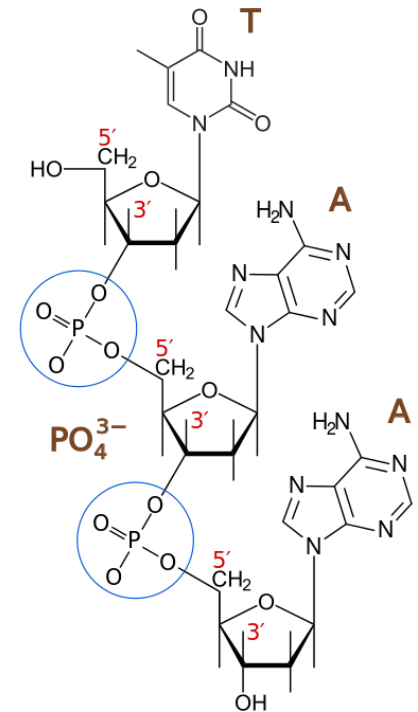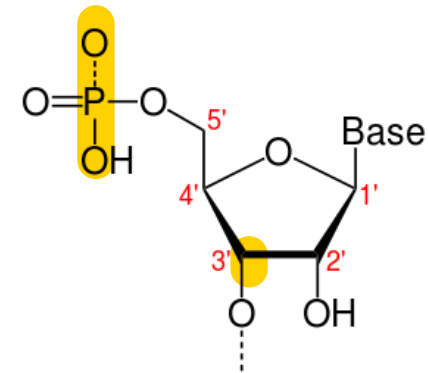
DNA replication

- In molecular biology, **DNA replication** is the biological process of producing two identical <mark>replicas of DNA from one original DNA molecule</mark>.

- This process occurs in all living organisms and is the basis for biological inheritance.

- The cell possesses the distinctive property of division, which makes replication of DNA essential.

- Why is DNA double stranded?

  - More stable

    - Replication consists of half-new, half old

  - Allows for error-correction

    - If a base is damaged, can correct

    - Like a RAID

Example Genome Sizes

Species	Number of base pairs	Number of genes
E. coli	4,600,000	3200
Fruit fly	180,000,000	13,600
Chicken	1,000,000,000	23,000
Mouse	2,500,000,000	30,000
Corn	2,500,000,000	59,000
Human	3,000,000,000	25,000-30,000
Grasshopper	180,000,000,000	?
Amoeba	670,000,000,000	?

National University of Computer &
Emerging Sciences, Peshawar.

Why so many base pairs?

- "Junk" DNA (or non-coding DNA)
- Portions of DNA sequence for which no function identified
- 98.5% of human genome
- May serve functions that are not yet understood

National University of Computer &
Emerging Sciences, Peshawar.

Why is DNA useful?

- Every living thing on earth uses DNA to store and transmit information

- Catalogs all the different functions performed in an organism

- Identify similarities among organisms

- Identify inherited traits

- Solving crimes

Why is DNA useful? (contd…)

- All of the information needed to build an organism is contained in its DNA. If we could understand it, we would know how life works.

  - Preventing and curing diseases like cancer (which is caused by mutations in DNA) and inherited diseases.

  - Curing infectious diseases (everything from AIDS and malaria to the common cold). If we understand how a **microorganism** works, we can figure out how to block it.

  - Understanding genetic and evolutionary relationships between species.

  - Understanding gene expression for personalized medicine.

National University of Computer & Emerging Sciences, Peshawar.

Chromosomes and Genes

- each chromosome is a long piece of DNA
    - Human chromosomes are 100-200 million bases long. We have 46 chromosomes (2 sets of 23, one set from each parent).
- genes are just regions on that DNA. It is not obvious where genes are if you look at a DNA sequence.
    - there is a lot of DNA that is not part of genes: in humans only 2% at most of the DNA is part of any gene.
    - Bacteria use more of their DNA: upto 80-90% of the chromosomes comprise of genes.
- Humans have about 25,000 genes.
    - We are far more complicated than bacteria: regulation of the genes is very complicated in humans
    - We use the same gene in different ways in different tissues

cell
nucleus
chromosome
gene
DNA
Adapted from National Human Genome Research Institute
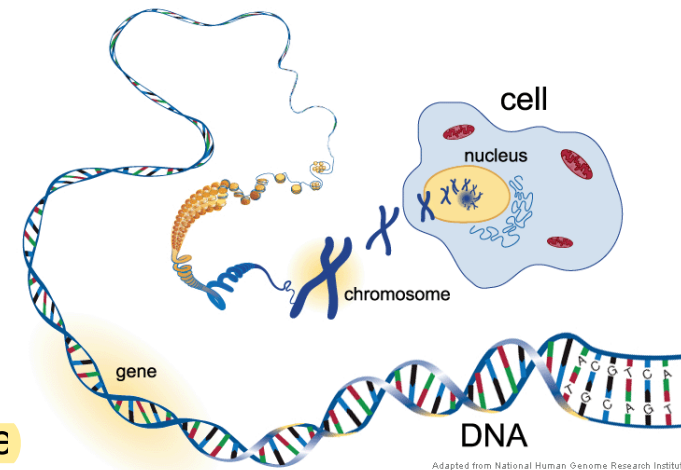
RNA vs DNA

Five differences between RNA and DNA

1. Sugar in RNA is ribose instead of deoxyribose
2. RNA is single stranded
3. RNA contains **Uracil base** instead of Thymine
4. RNA is disposable
5. RNA can be outside the nucleus, DNA can't

If you are going from DNA to RNA – what nitrogenous base would pair with following?

DNA: C A G T T A

RNA: _ _ _ _ _ _

ANSWER:

DNA: C A G T T A

RNA: G U C A A U

RNA

- DNA is blueprint, doesn't do much on its own
- Transcription – copy DNA into RNA
  - Only genes get transcribed
  - **Promoter** – DNA sequence that enables gene to be transcribed
  - **Exon** – protein coding sequence of gene
  - **Intron** – sections of DNA that are spliced out after transcription
- Messenger RNA (mRNA) moves out of cell nucleus to provide building plans for *proteins*
- RNA alphabet: Uracil (U) instead of Thymin (T)

National University of Computer &
Emerging Sciences, Peshawar.

How does DNA work?

The "Central Dogma"

DNA ↓	

The "Central Dogma"

DNA ⬇	
RNA ⬇	
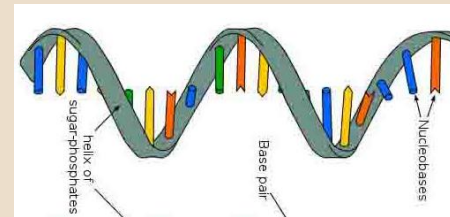
The "Central Dogma"
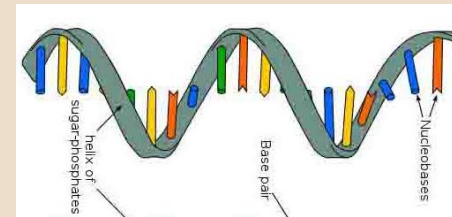
DNA ⬇	
RNA ⬇	
proteins ⬇	

The "Central Dogma"
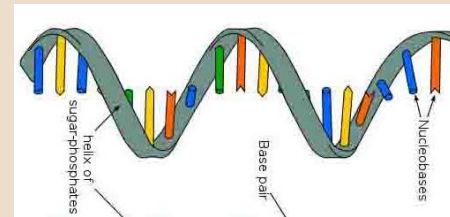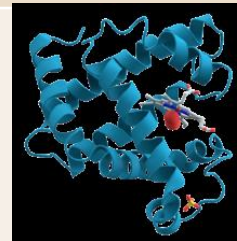
DNA ↓	
RNA ↓	
proteins ↓	
you	

The Key Players: Types of RNA

- Key players in Protein Synthesis -
  - <u>Messenger RNA </u>(mRNA): carry instruction copies
  - <u>Ribosomal RNA </u>(rRNA): makes up ribosome along with other proteins
  - <u>Transfer RNA </u>(tRNA): brings amino acids to the ribosome
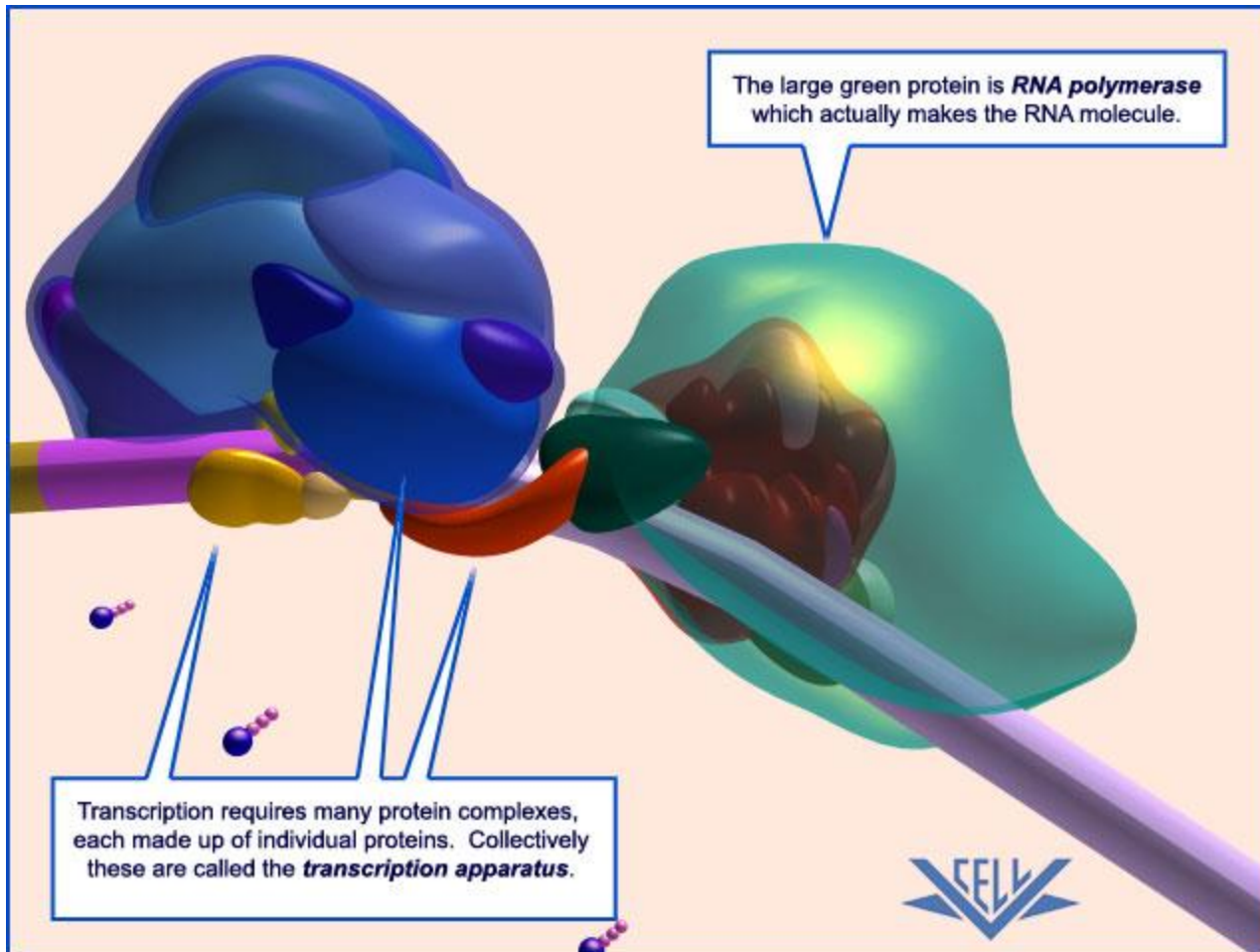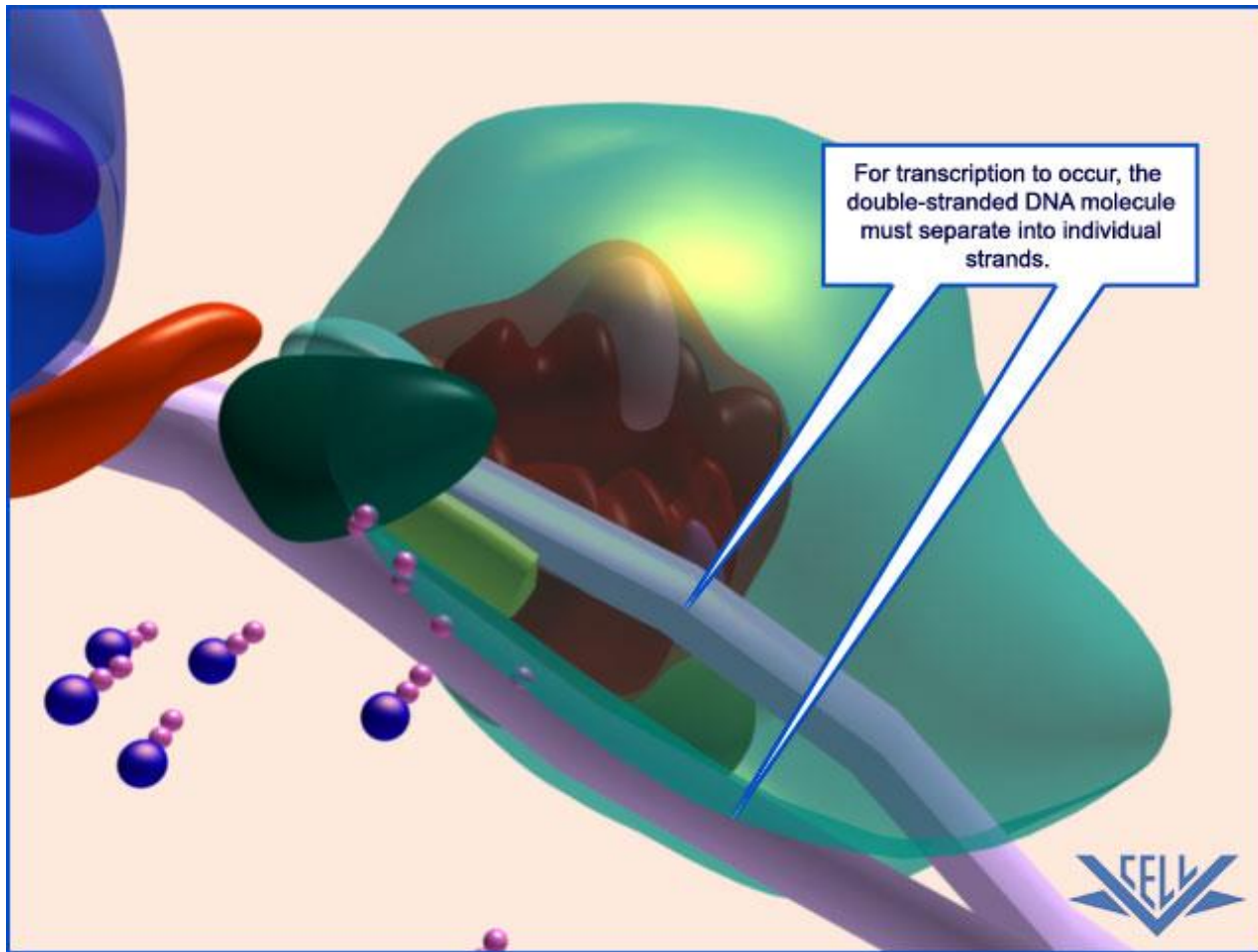
Big picture of protein synthesis

Antisense strand

RNA polymerase

CTGACGGATCAGCCGCAAGCGGAATTGGCGACATAA

GACUGCCUAGUCGGCGUU

RNA Transcript

GACTGCCTAGTCGGCGTTCGCCTTAACCGCTGTATT

Sense Strand
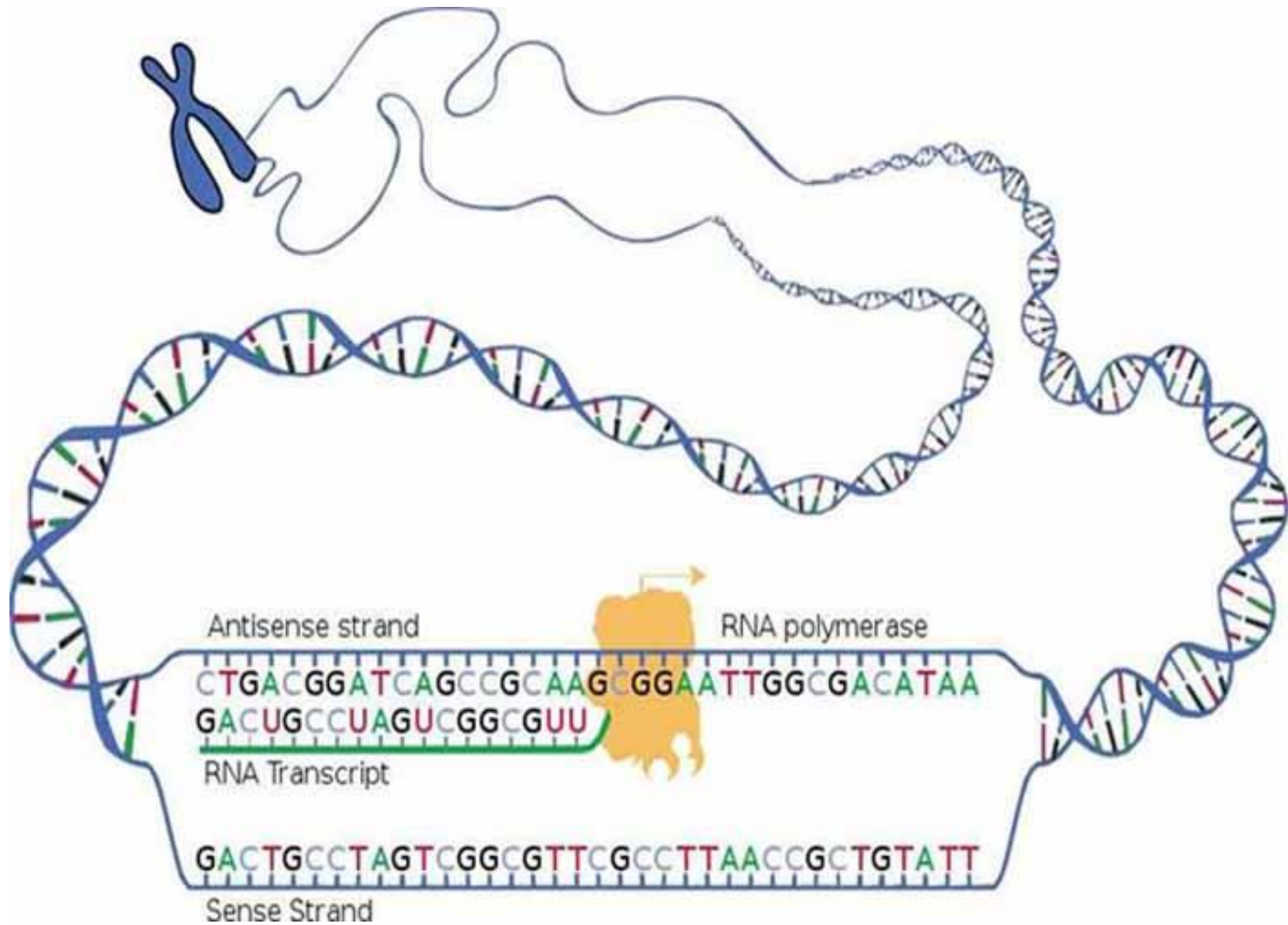
Transcription unit.

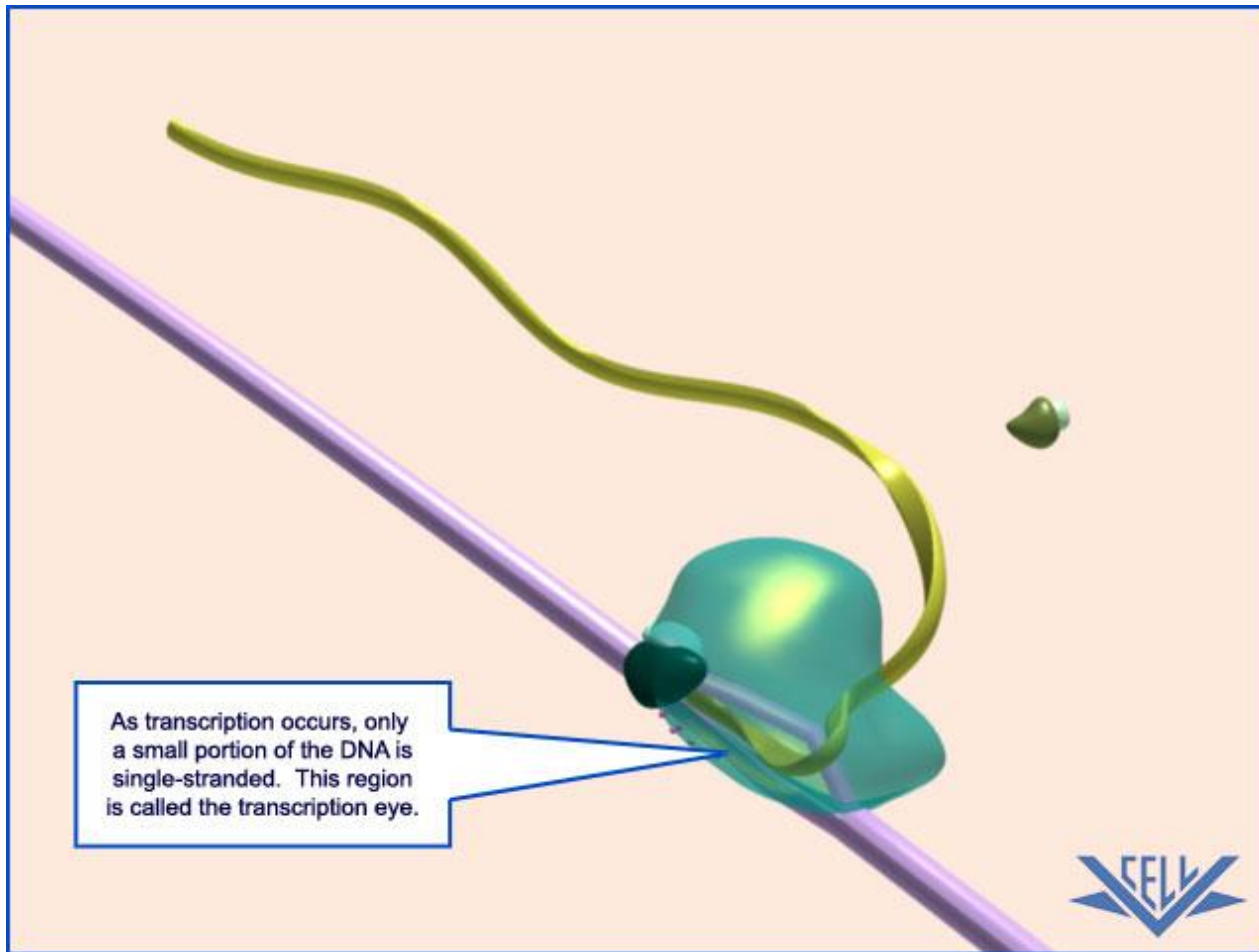# *Promoter* and the *TATA box.*

Transcription apparatus.

The large green protein is **RNA polymerase** which actually makes the RNA molecule.

Transcription requires many protein complexes, each made up of individual proteins. Collectively these are called the *transcription apparatus*.

Strand separation is necessary for *RNA Polymerase* to transcribe RNA.

For transcription to occur, the double-stranded DNA molecule must separate into individual strands.

RNA polymerase moves along the DNA.

The RNA polymerase complex actually moves from its initial site near the TATA box along the DNA during transcription.

Transcription eye.
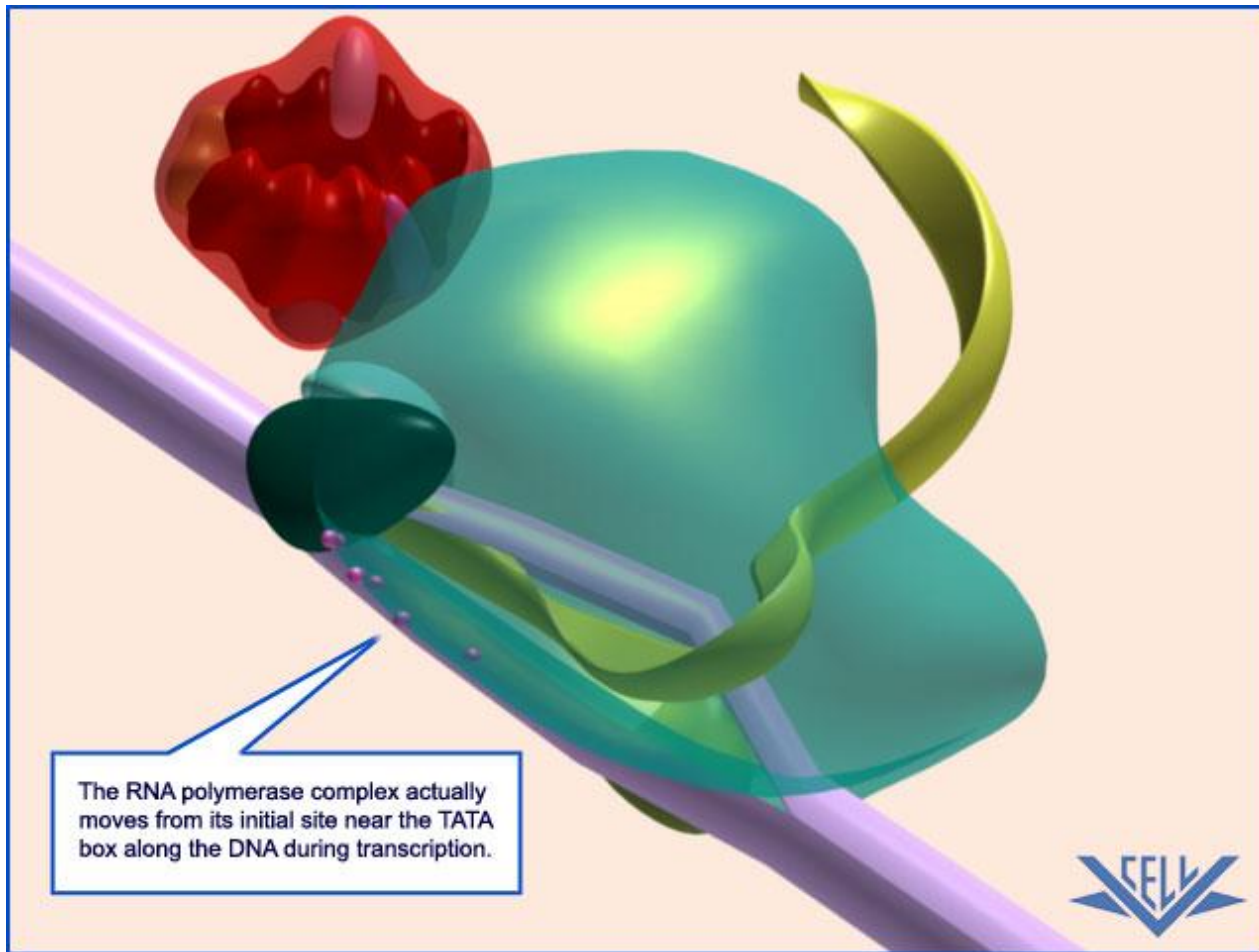
As transcription occurs, only a small portion of the DNA is single-stranded. This region is called the transcription eye.

Transcription is complete.
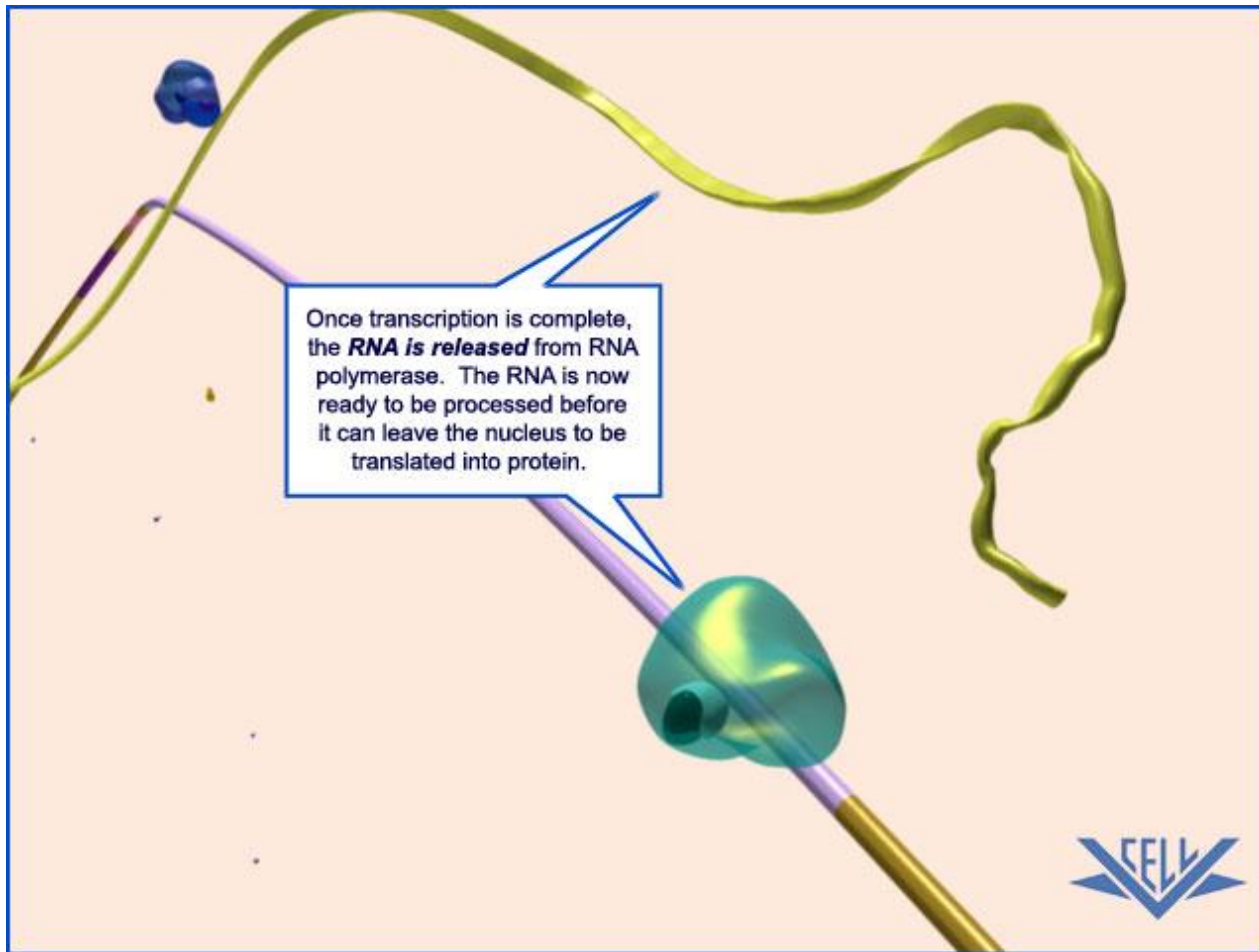
2. Translation
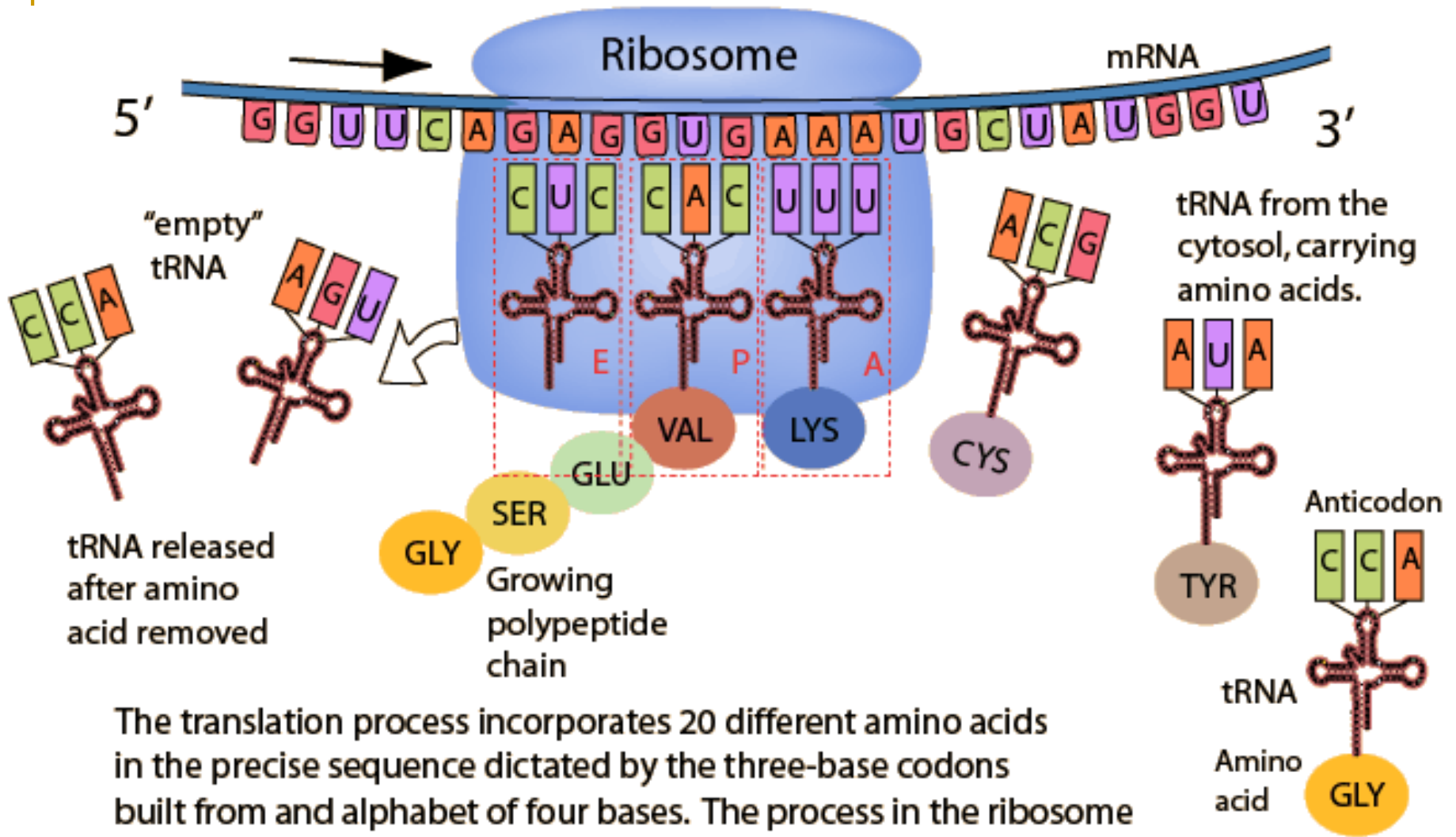
The translation process incorporates 20 different amino acids in the precise sequence dictated by the three-base codons built from and alphabet of four bases. The process in the ribosome builds the polypeptide chains tha will become proteins.
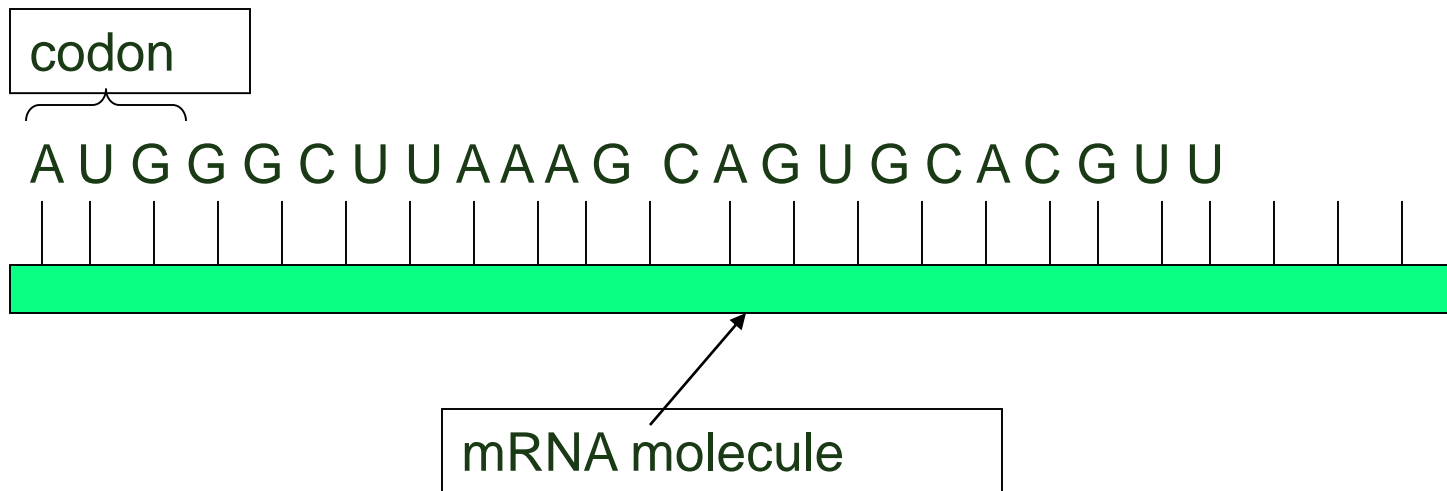
Just Like Construction…

- DNA is the master plan
- mRNA is the everyday blueprint
- rRNA is the builder
- tRNA is the provider
- Amino Acids are the wood
- Proteins are the building
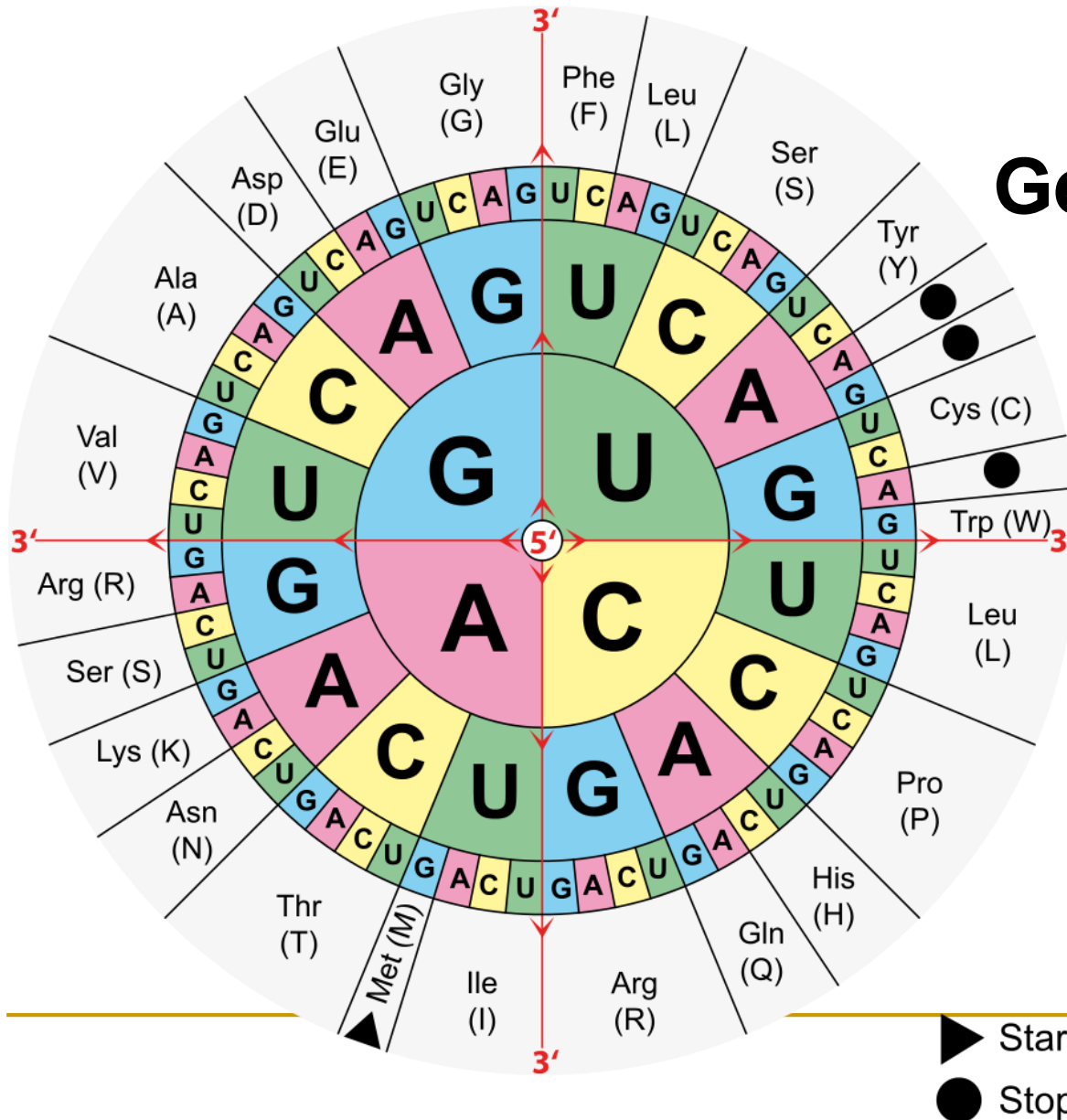
Translation of the Genetic Code -

Protein Synthesis

This is a molecule of messenger RNA.

mRNA is <u>transcribed</u> in the nucleus.
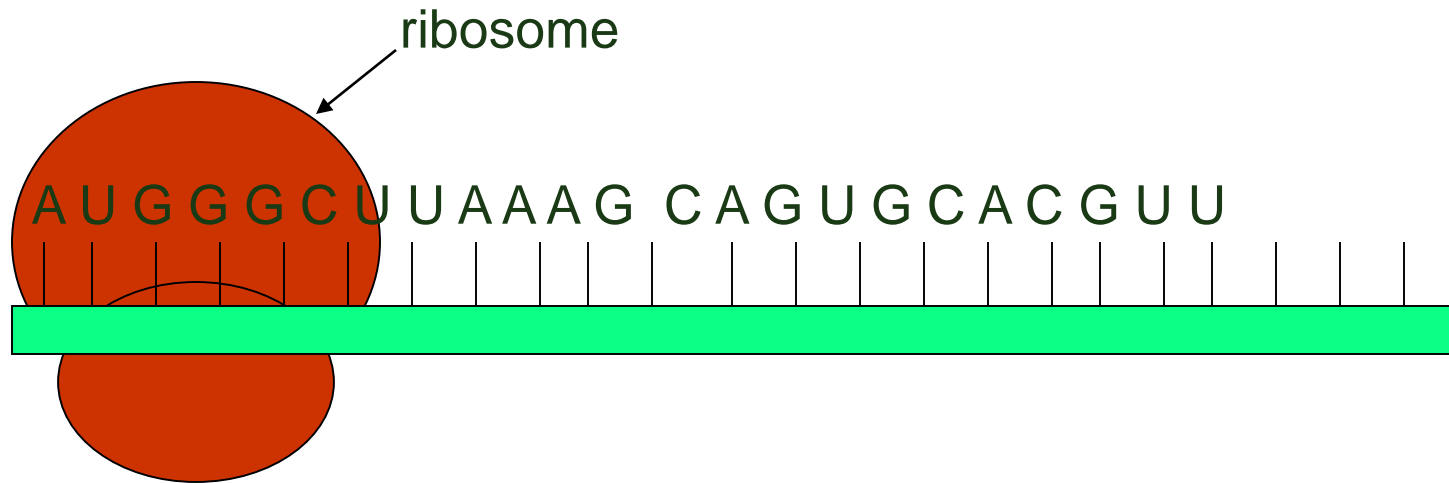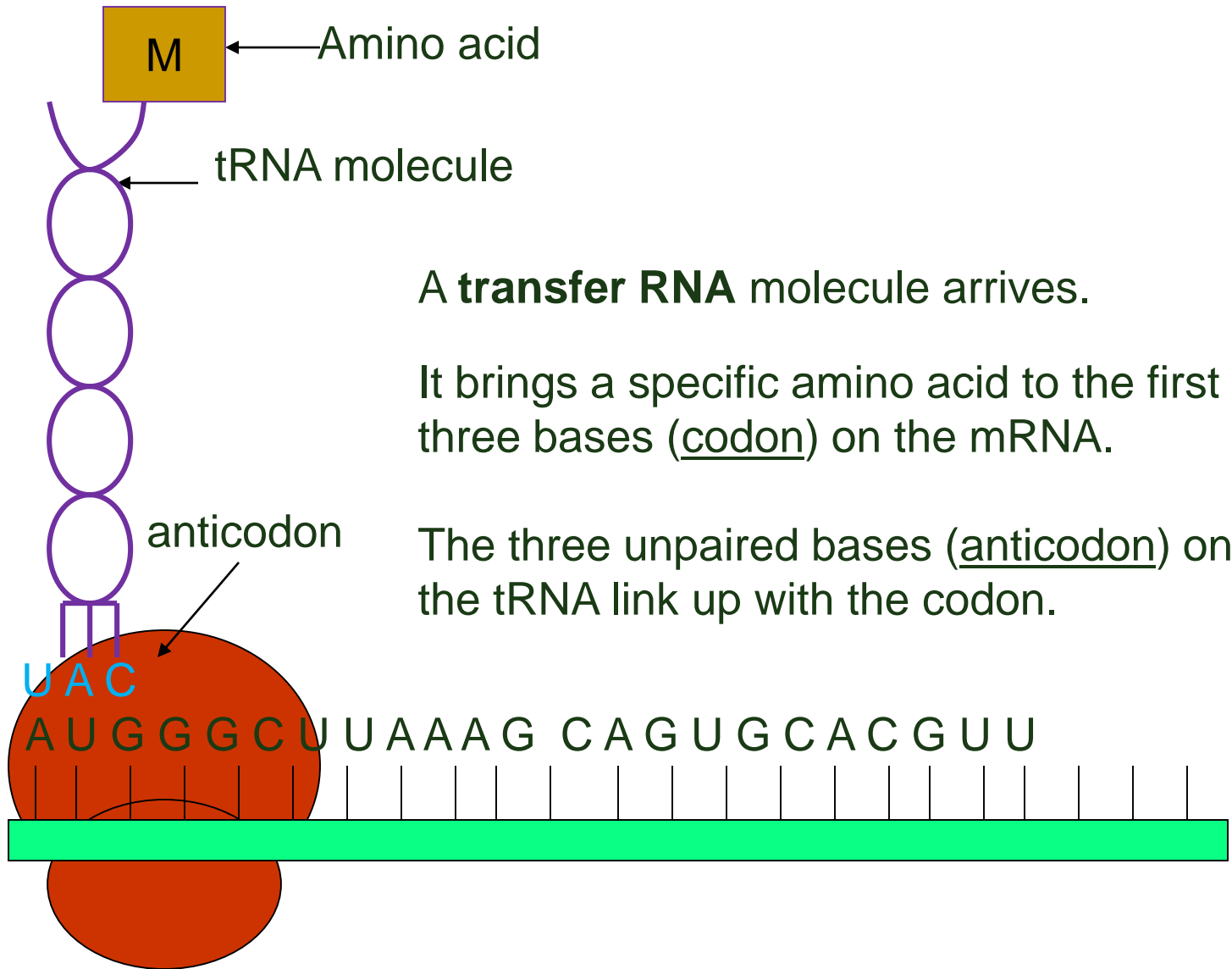
codon

A U G G G C U U A A A G C A G U G C A C G U U

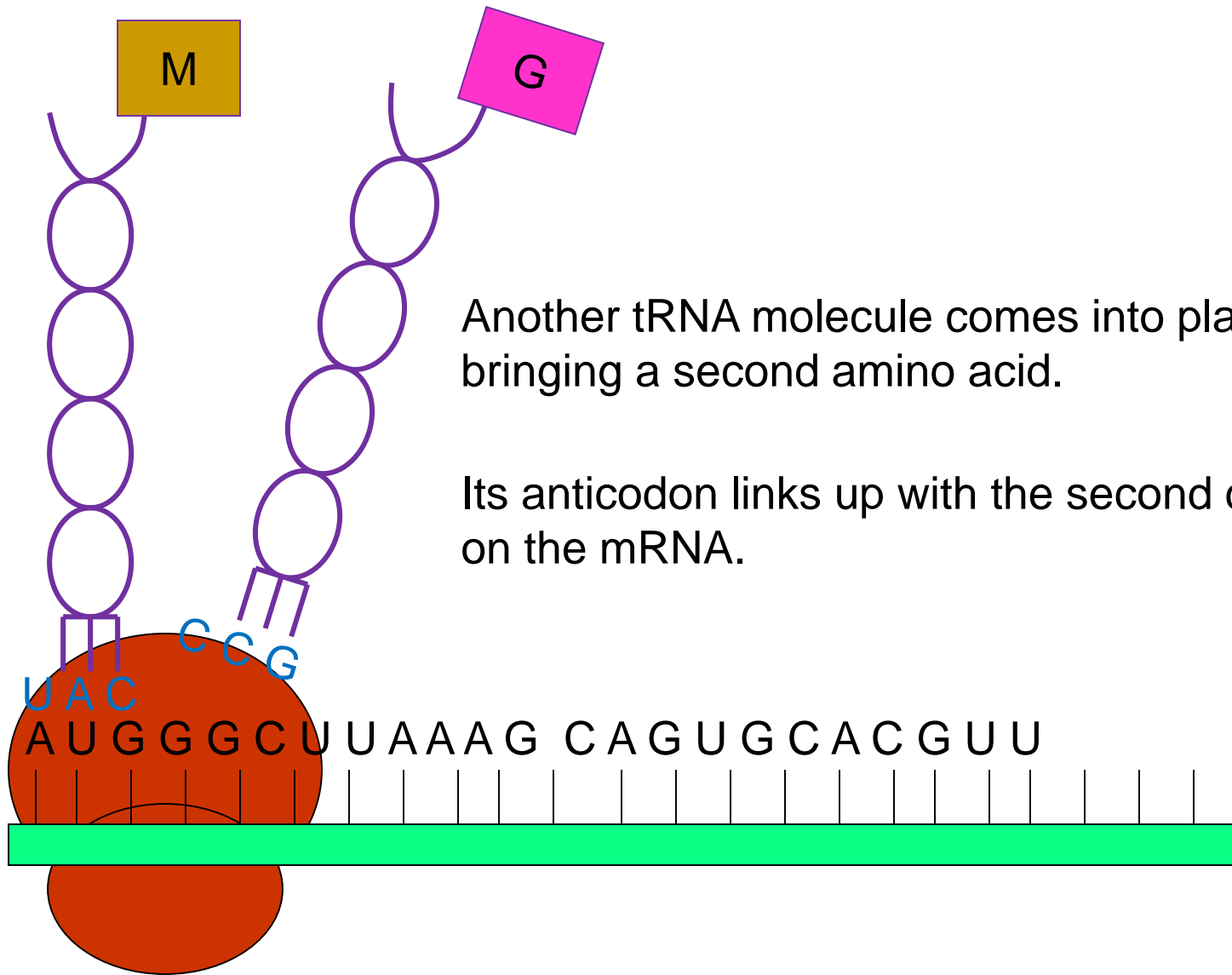mRNA molecule

The CODON TABLE

Reading Genetic Material

Start

Stop

A large molecule called ribosome, attaches to the mRNA molecule.

ribosome

A U G G G C U U A A A G C A G U G C A C G U U

M ← Amino acid

tRNA molecule

anticodon

A **transfer RNA** molecule arrives.

It brings a specific amino acid to the first three bases (codon) on the mRNA.

The three unpaired bases (anticodon) on the tRNA link up with the codon.

U A C

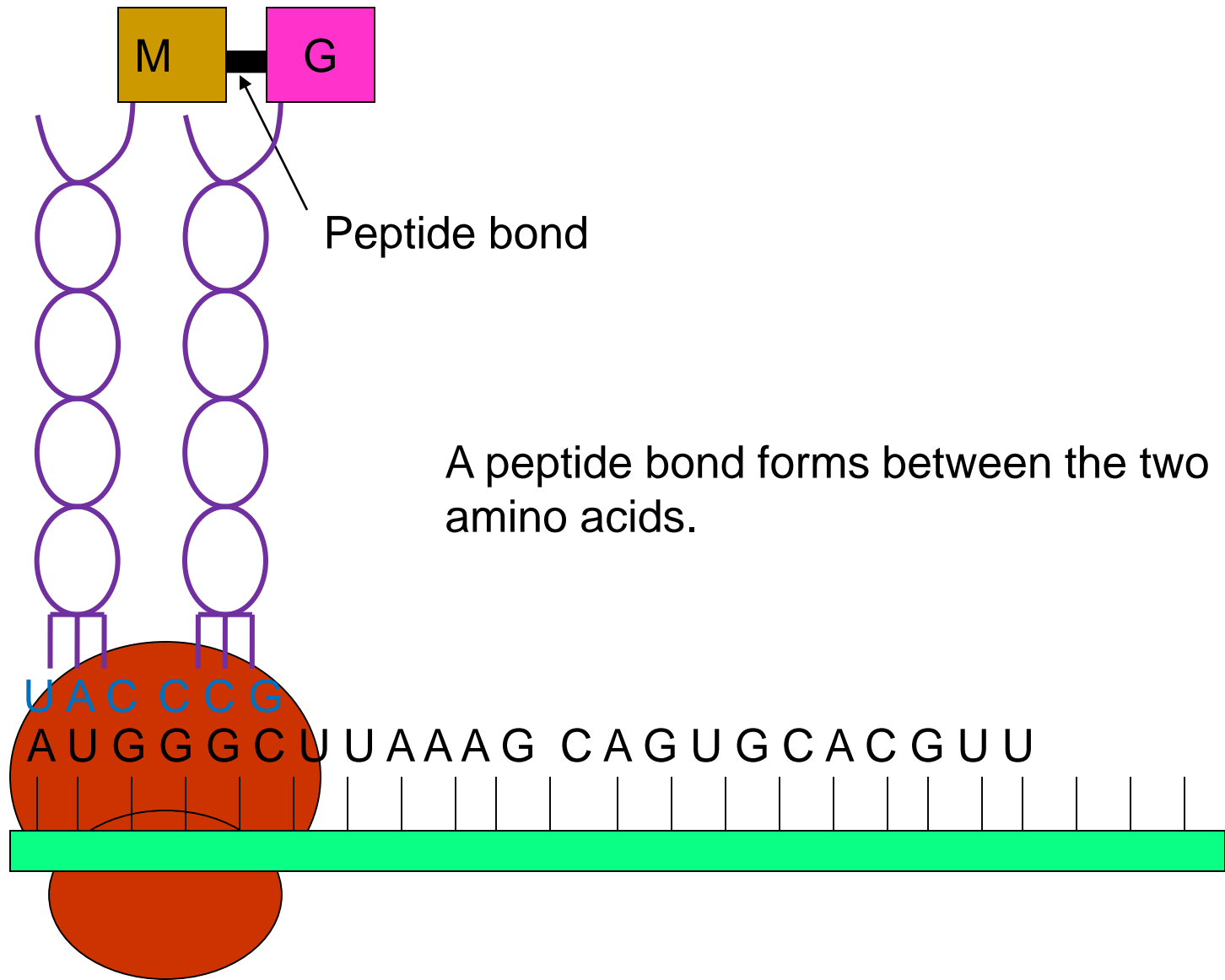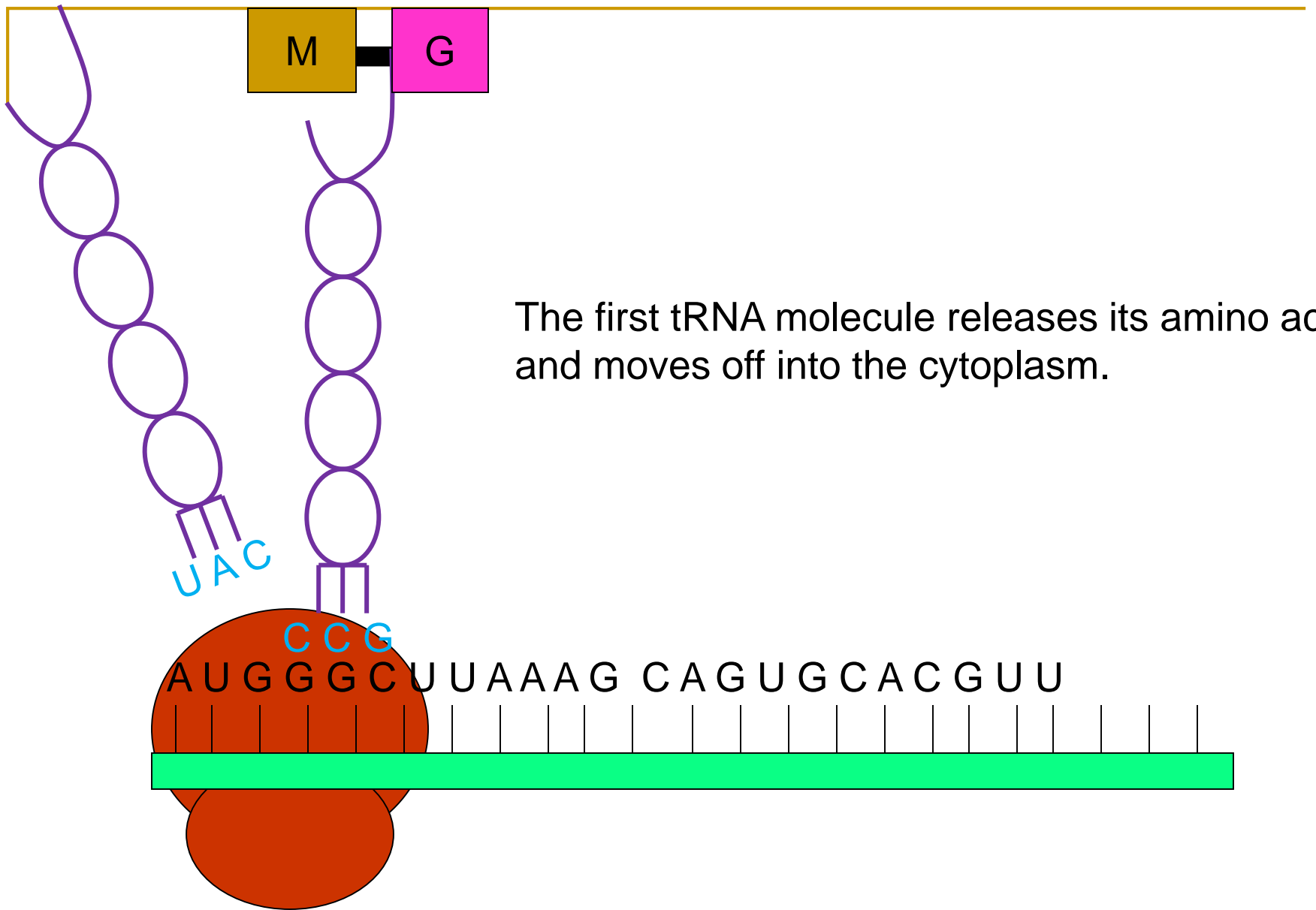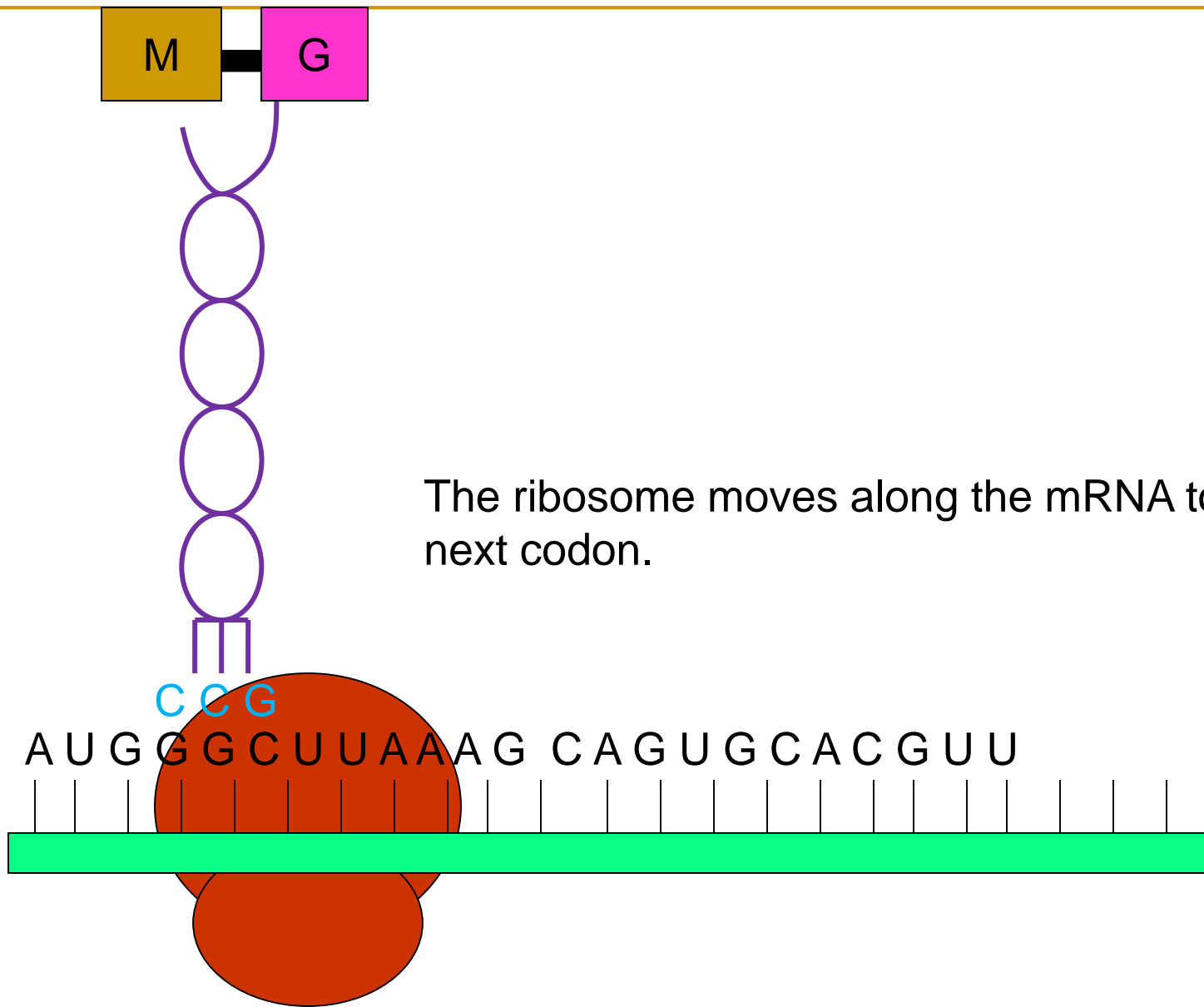A U G G G C U U A A A G C A G U G C A C G U U

M

G

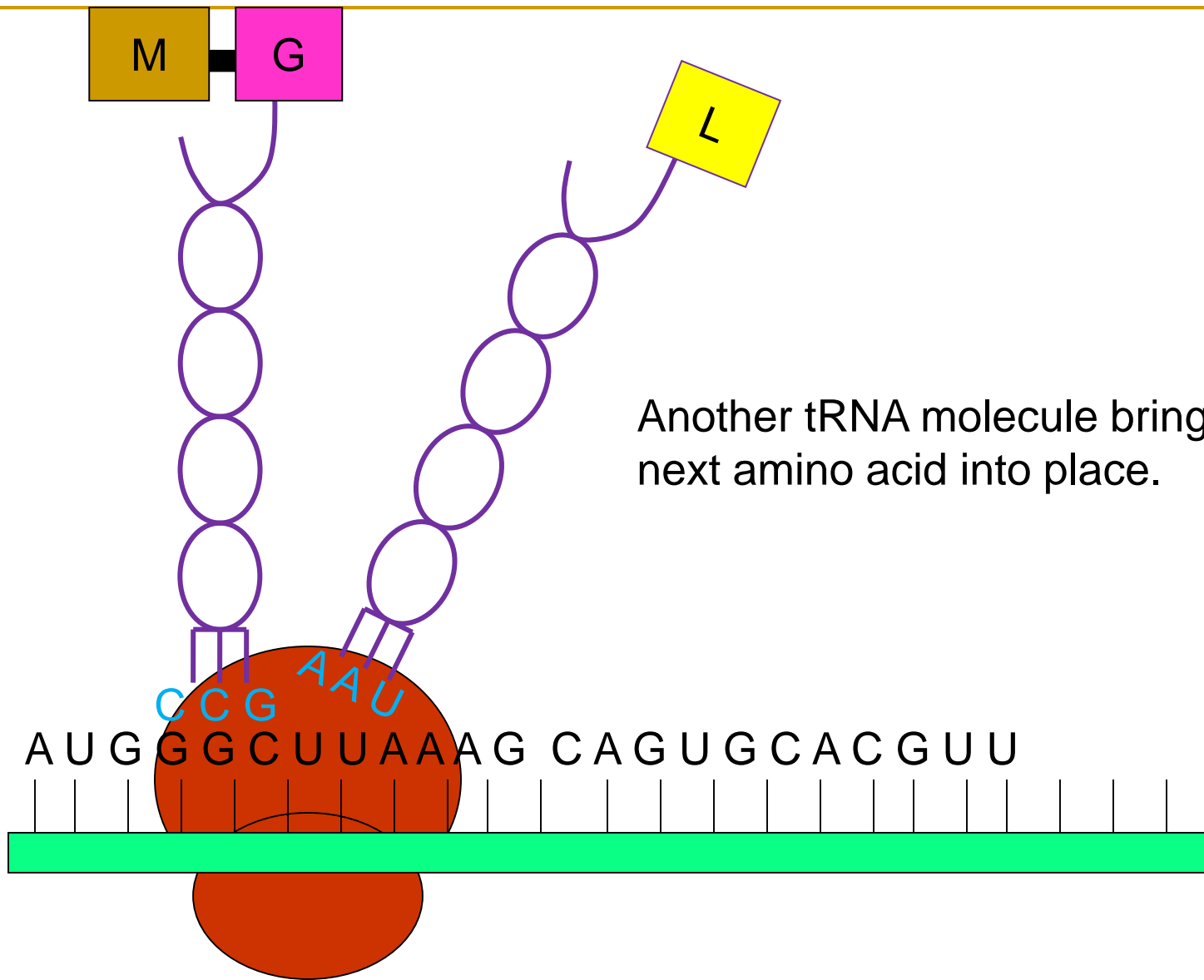Another tRNA molecule comes into place, bringing a second amino acid.

Its anticodon links up with the second codon on the mRNA.

C C G

U A C

A U G G G C U U A A A G C A G U G C A C G U U

Peptide bond

A peptide bond forms between the two amino acids.

M G

U A C   C C G

A U G G G C U U A A A G   C A G U G C A C G U U

The first tRNA molecule releases its amino acid and moves off into the cytoplasm.

M   G

U A C

C C G

A U G G G C U U A A A G C A G U G C A C G U U
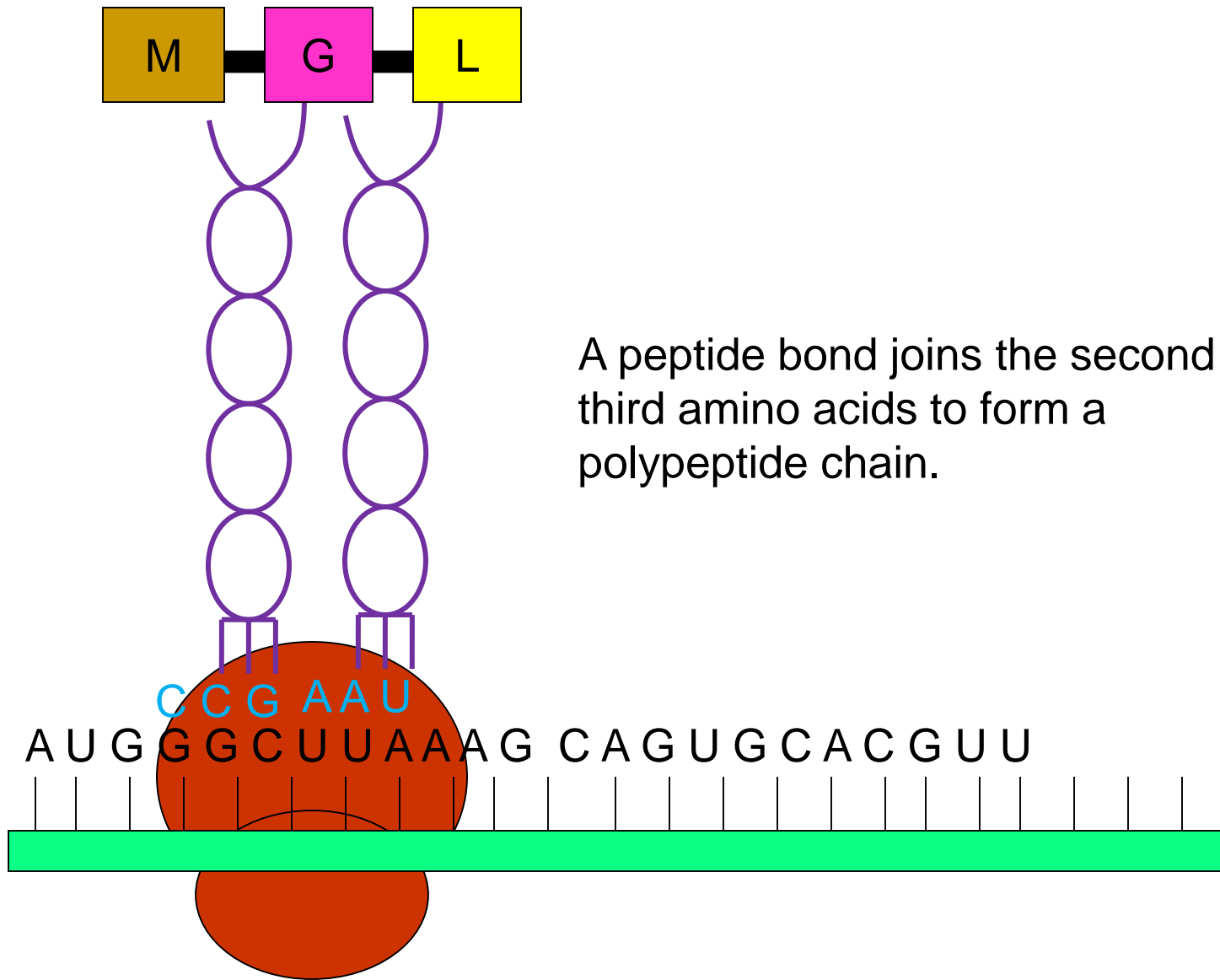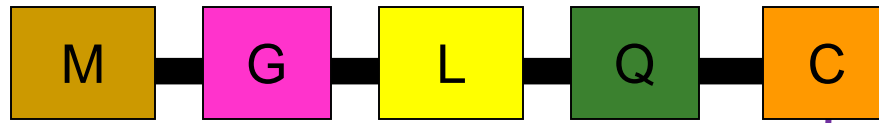
The ribosome moves along the mRNA to the next codon.

Another tRNA molecule brings the next amino acid into place.

A peptide bond joins the second and third amino acids to form a polypeptide chain.
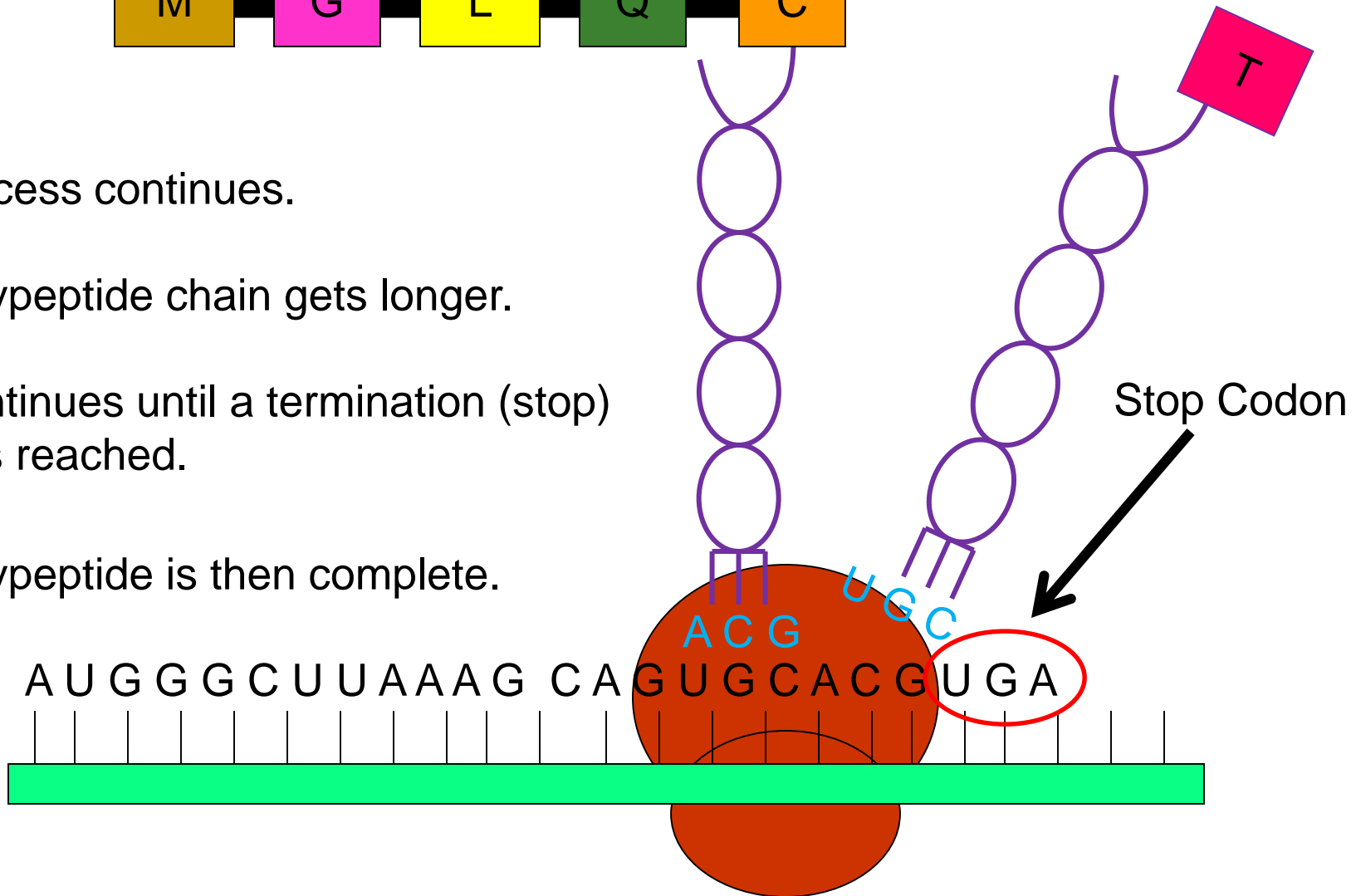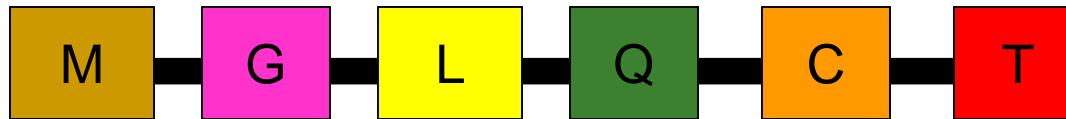
M — G — L — Q — C

The process continues.

The polypeptide chain gets longer.

This continues until a termination (stop) codon is reached.
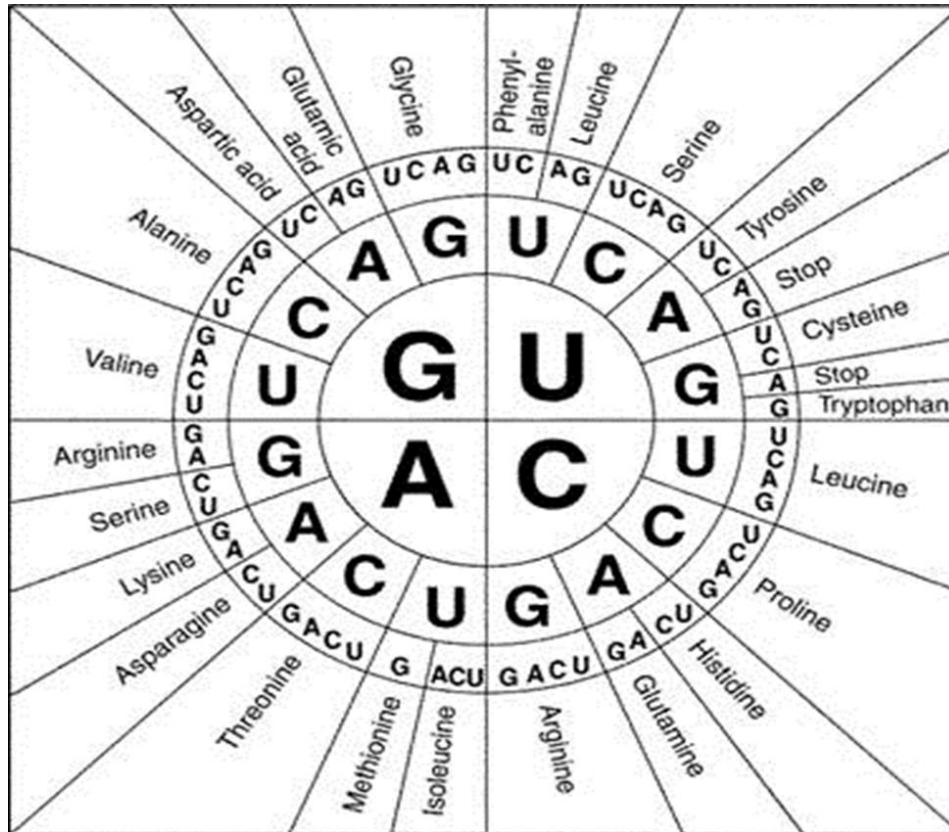
The polypeptide is then complete.

T

Stop Codon

U G C

A C G

A U G G G C U U A A A G C A G U G C A C G U G A

Newly formed protein

National University of Computer &
Emerging Sciences, Peshawar.

POLAR SIDE CHAINS		NON POLAR SIDE CHAINS	
SERINE	NH_2 $HOH_2C - C - COOH$ H	GLYCINE	NH_2 $H - C - COOH$ H
THREONINE	NH_2 $H_3C - CHOH - C - COOH$ H	ALANINE	NH_2 $H_3C - C - COOH$ H
TYROSINE	NH_2 $HO\bigcirc CH_2 - C - COOH$ H	CYSTEINE [1]	NH_2 $HS - CH_2 - C - COOH$ H
ASPARAGINE	NH_2 $H_2N - CO - CH_2 - C - COOH$ H	VALINE	H_3C NH_2 $CH - C - COOH$ H_3C H
GLUTAMINE	NH_2 $H_2N - CO - CH_2 - CH_2 - C - COOH$ H	LEUCINE	H_3C NH_2 $CH - CH_2 - C - COOH$ H_3C H
		ISOLEUCINE	CH_3 NH_2 $H_3C - CH_2 - CH - C - COOH$ H

Example:

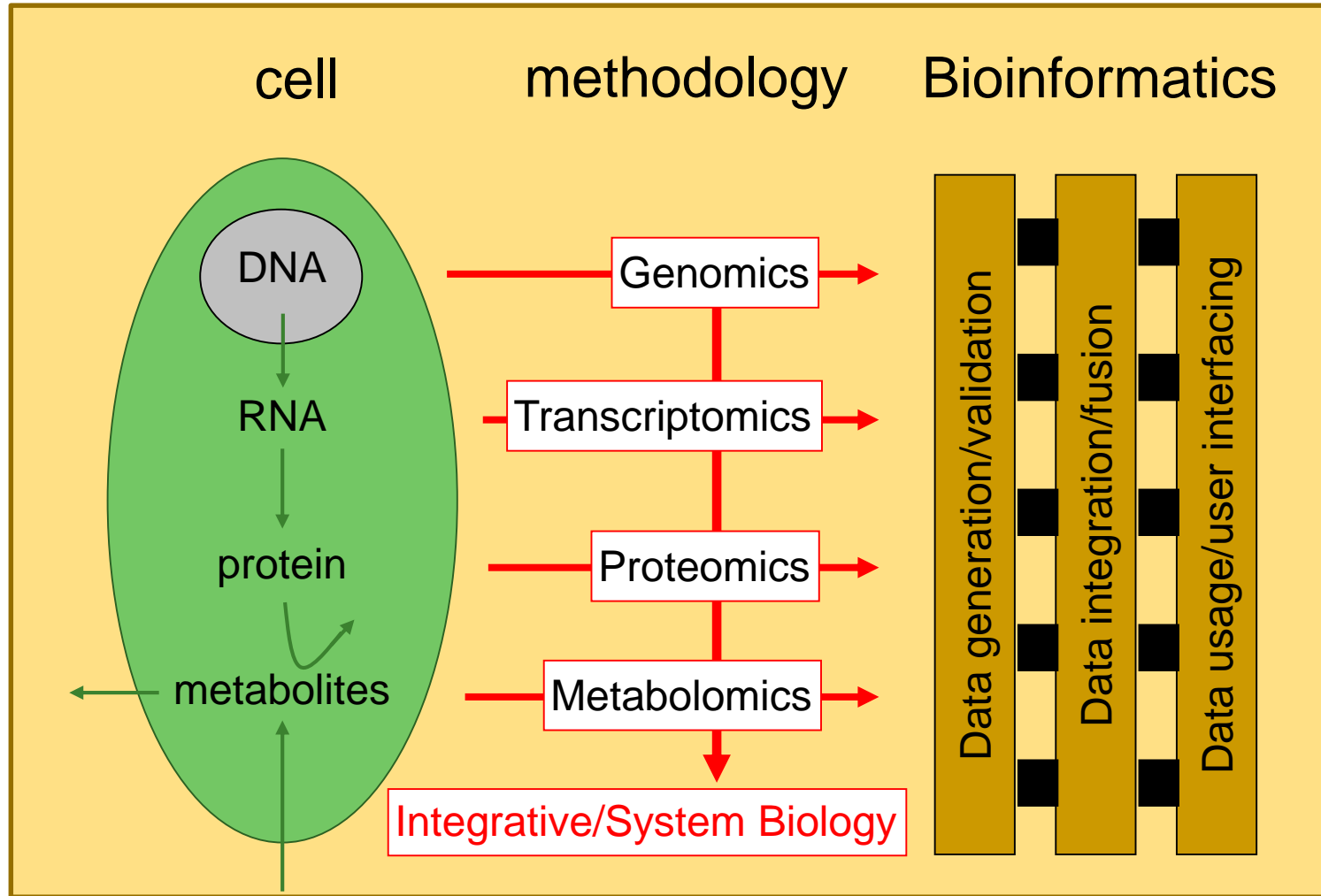1. What amino acids would be coded by the following DNA sequence?

TACCGGACCGCTATC

2. What would be the DNA sequence for a polypeptide chain (protein) that had the following amino acids?

Lysine-Cysteine-Proline

Role of Bioinformatics

Applications/Impact

Personalized Medicine

- An individual's genetic inheritance affects the body's response to drugs:

  - *At present*, some drugs fail to make it to the market because a small percentage of the clinical patient population show adverse affects to a drug due to sequence variants in their DNA. As a result, potentially life saving drugs never makes it to the marketplace.

  - *Today*, doctors have to use **trial and error** to find the best drug to treat a particular patient as those with the same clinical symptoms can show a wide range of responses to the same treatment.

  - *In future*, doctors will be able to analyze a patient's genetic profile and prescribe the best available drug therapy and dosage from the beginning.

Preventative Medicine

- With the specific details of the genetic mechanisms of diseases being unraveled, the development of diagnostic tests to measure a persons susceptibility to different diseases may become a distinct reality.

- Preventative actions such as change of lifestyle or having treatment at the earliest possible stages when they are more likely to be successful, could result in huge advances in our struggle to conquer disease.

# Biological complexity & Networks An Example of Gene Regulatory Networks

Complex Systems

- A **system** is a set of interacting or interdependent components forming a complex/intricate whole.

- A **complex system** is a system that exhibits some (and possibly all) of the following characteristics:

  - feedback loops;

  - some degree of spontaneous order/organization;

  - the order is robust;

  - emergent organization;

  - numerosity;

  - hierarchical organization.

- Examples of complex systems are, the **human brain**, **social organization (e.g. cities, towns etc)**, **an ecosystem**, **a living cell**, and ultimately the entire universe.

Why Networks?

- Networks are a useful model to study COMPLEX SYSTEMS, as most biological systems are.

- Complex systems cannot be described by a single rule and they exhibit properties that emerge from the interaction of their parts and which **cannot be predicted only from the properties of their parts**.

- Networks allow to easily model the interconnections between different components of a complex system and study its emerging properties.

National University of Computer &
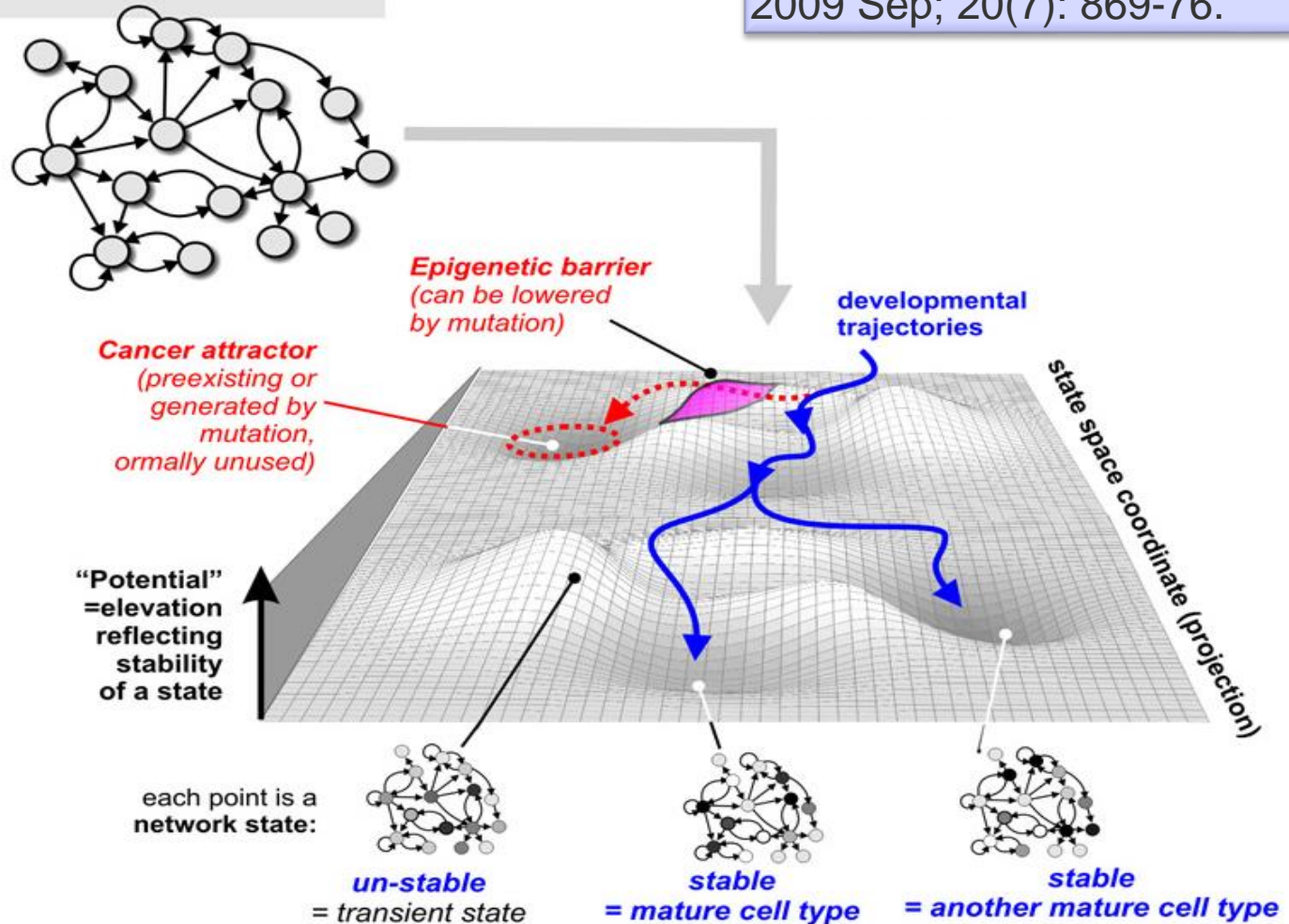Emerging Sciences, Peshawar.
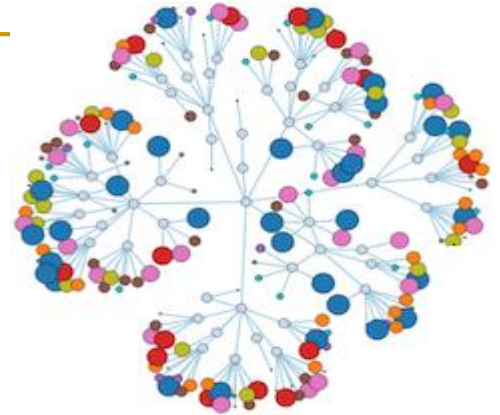
Gene Regulatory Networks (GRN)
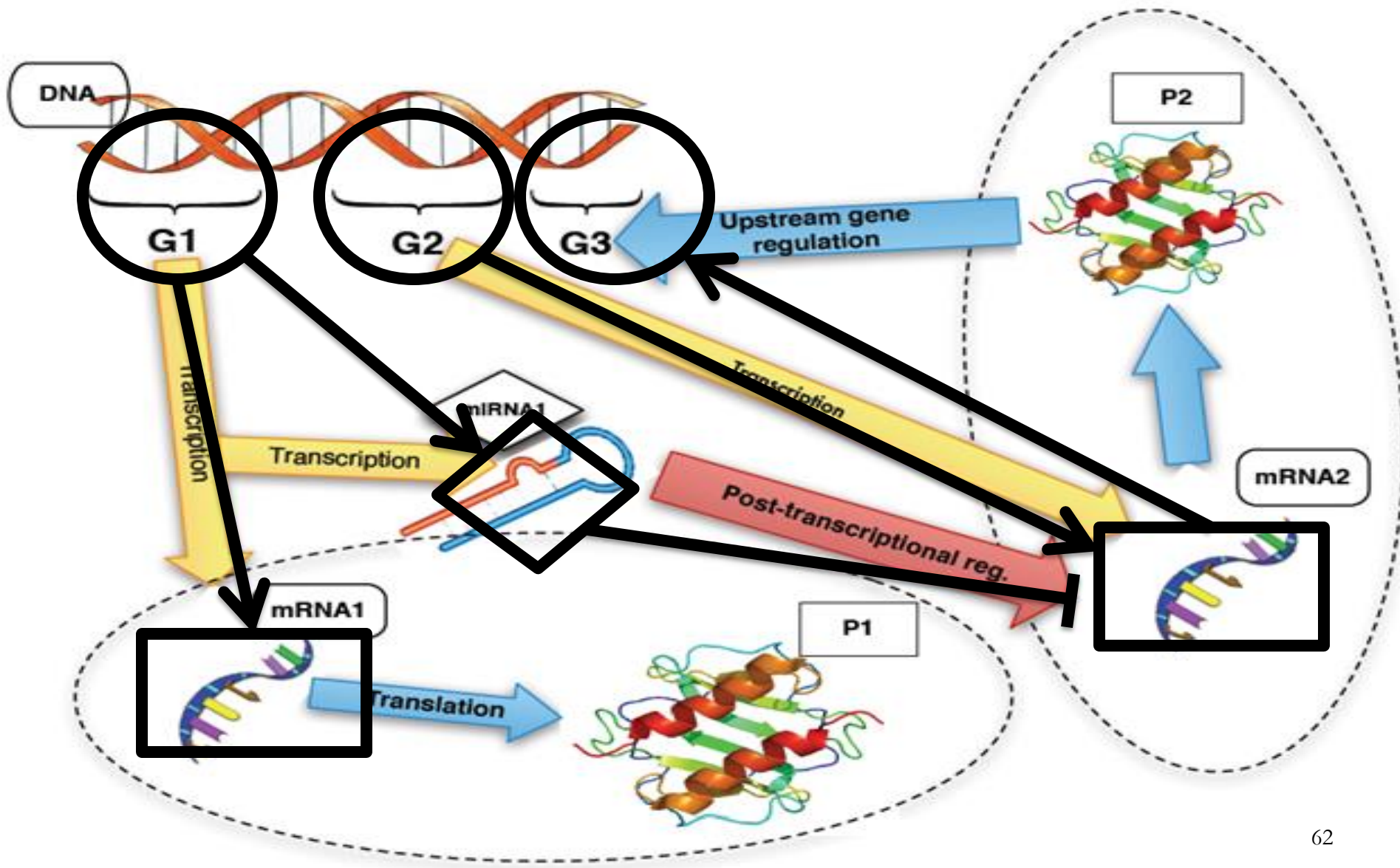
Impact: Epigenetic Landscape

Huang et al.,
Semin Cell Dev Biol.
2009 Sep; 20(7): 869-76.

COMPLEX NETWORK (*N* gene genome)

Epigenetic barrier
(can be lowered by mutation)

developmental
trajectories

Cancer attractor
(preexisting or generated by mutation, ormally unused)

"Potential"
=elevation
reflecting
stability
of a state

state space coordinate (projection)

each point is a
network state:

un-stable
= transient state

stable
= mature cell type

stable
= another mature cell type

Boolean Network Model

$$G3 = mRNA2$$

G1

G2

G3

The state of a node at time (t+1) is determined by the Boolean function of its k-inputs at time (t).

miRNA 1

mRNA2

$$miRNA1 = G1$$

$$mRNA2 = G2 \cdot \overline{miRNA1}$$

mRNA1

Biological constraints

$$C1 = mRNA1 \cdot \overline{G1}$$

$$mRNA1 = G1$$

$$C2 = miRNA1 \cdot \overline{G1}$$

The network is simulated to identify
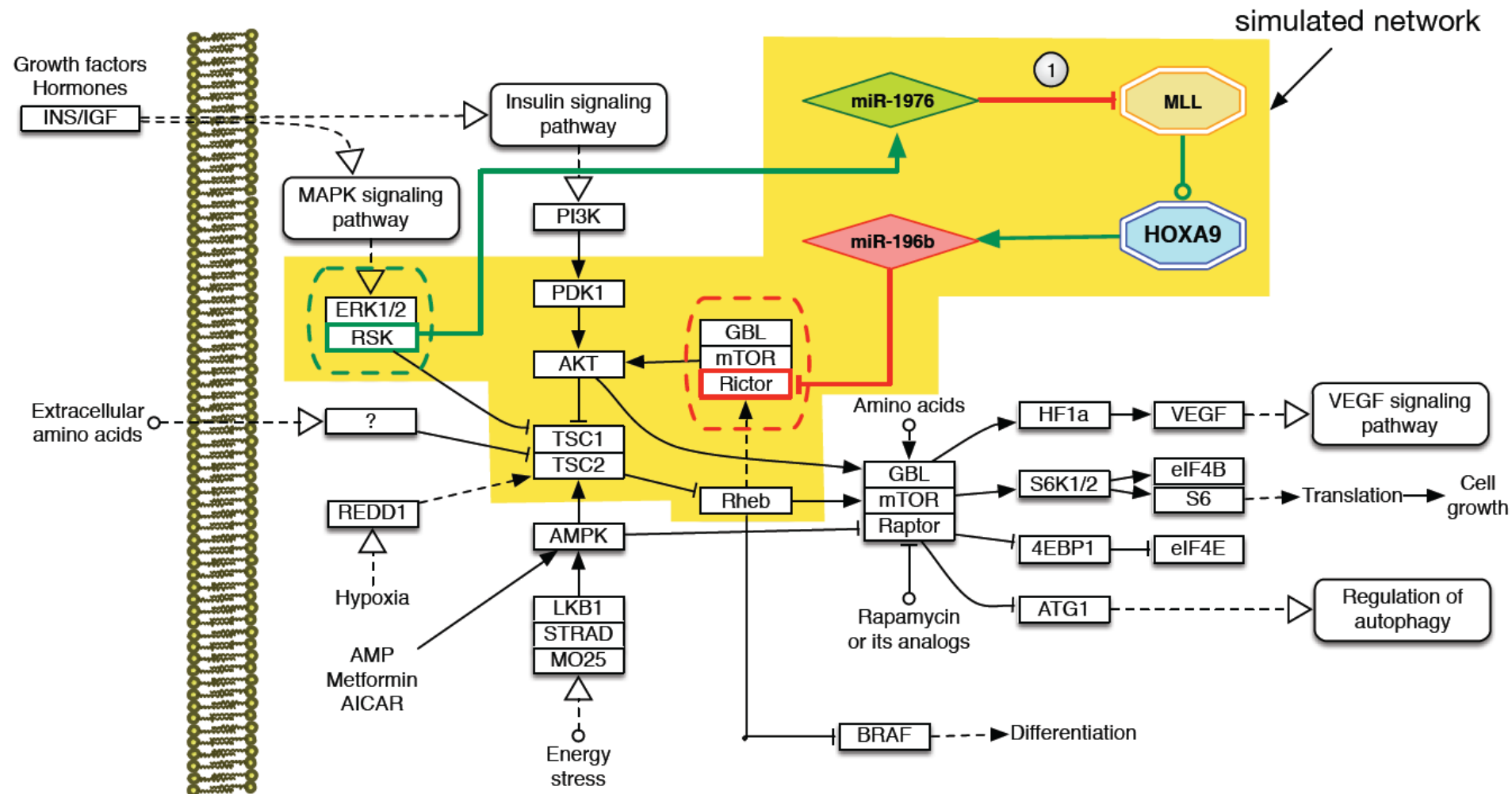
$$C3 = mRNA2 \cdot \overline{G2}$$

- stable states (**attractors**)

Must be false

- trajectories
- state spaces
- state potentials

Gene Regulatory Networks (GRNs)
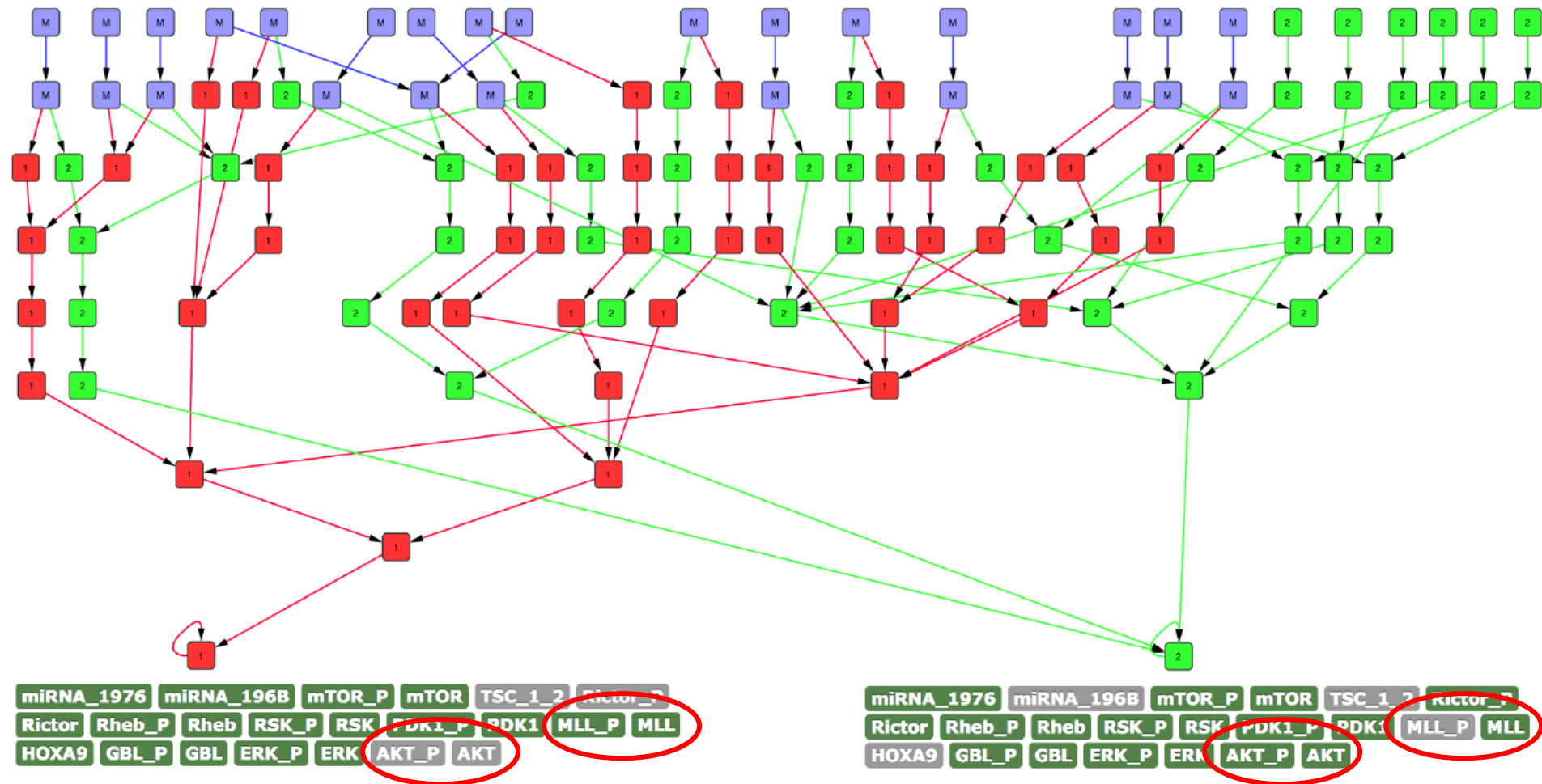
- **Modeling GRNs with Boolean Networks (BNs):**
    - Directed graphs
    - Nodes are Genes/Proteins/mRNAs/miRNAs
    - State of Every Node is 0/1.
- **Simulator identifies:**
    - Stable States (attractors)
    - Trajectories
    - State Space
- **Avoids:**
    - Biologically impossible states

Tool to Study Dynamics
(Example: mTOR Pathway)

National University of Computer &
Emerging Sciences, Peshawar.

Simulation Result

National University of Computer &
Emerging Sciences, Peshawar.

End of Lecture Set 01

A message!

"There are realities we all share, regardless of our nationality, language, or individual tastes. As we need food, so do we need emotional nourishment: love, kindness, appreciation, and support from others."

~ Swami Kriyananda