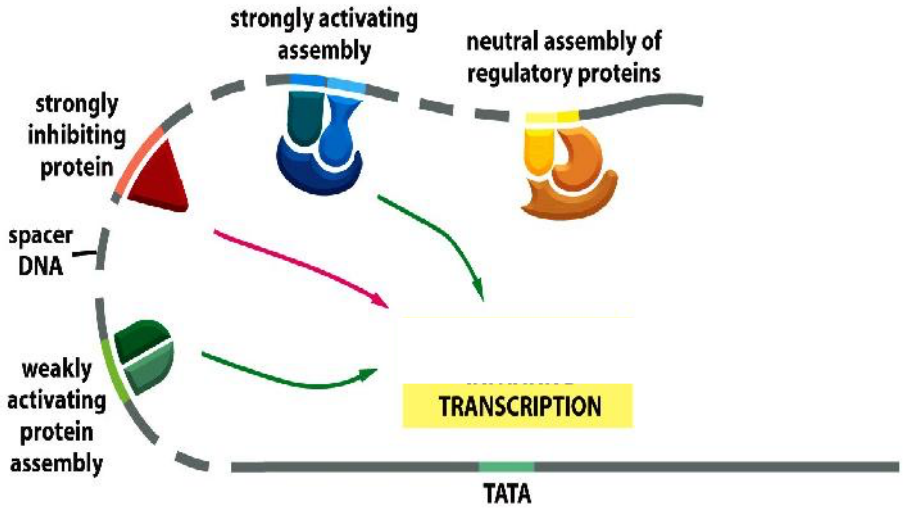


Molecular Biology Background

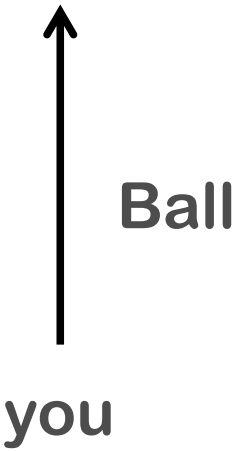
Genomics and Bioinformatics

Chapter 2

Collision in classical mechanics

What happens if you kick this ball?

The outcome of the experiment is determined by Newton's laws



Cause and effect are close in time and space

Collision in classical biology

Tim Newman, Paul Davies

What happens if you kick this sleeping tiger into the butt?

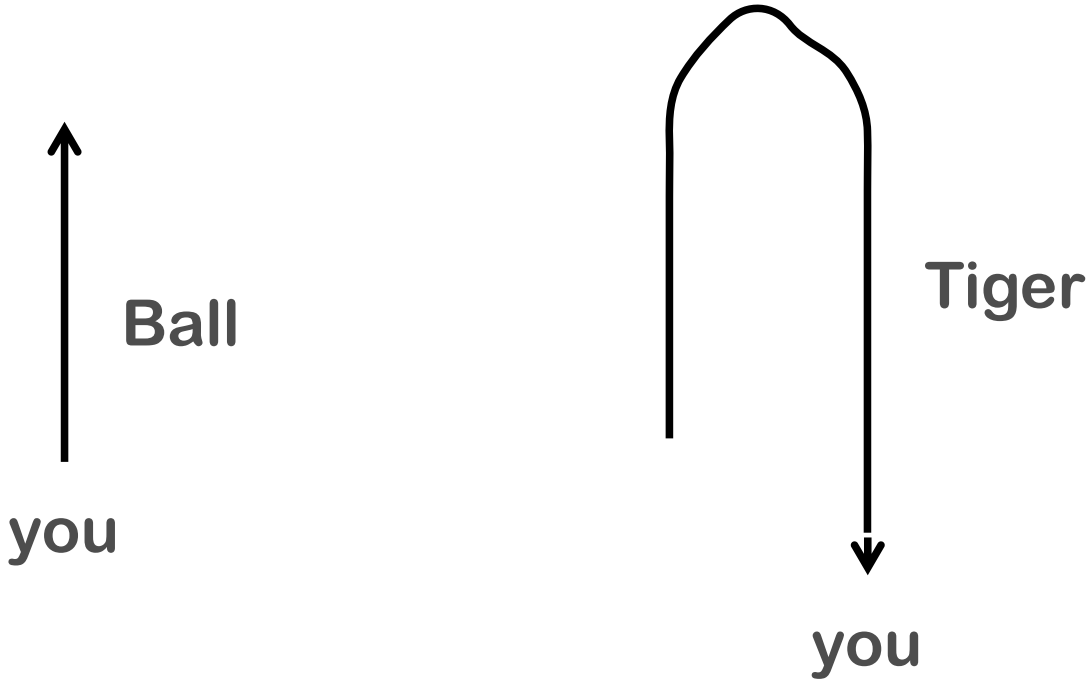
The first two seconds are physics

Tim Newman, Paul Davies

Newton's laws predict how the butt moves forward and how the fur slides along the ribs …

Cause and effect are only seconds apart

After two seconds the outcome is remarkably different to the collision experiment with the ball



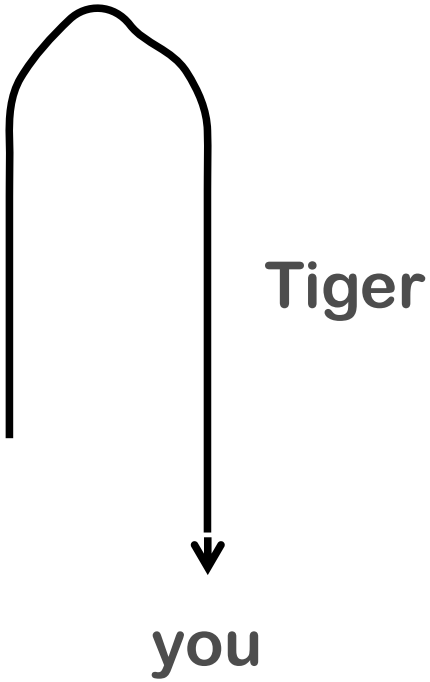
Once the tiger is awake it processes input signals into output signals

Tim Newman, Paul Davies

What laws determine these computations?

The output signal is determined by the tigers genome

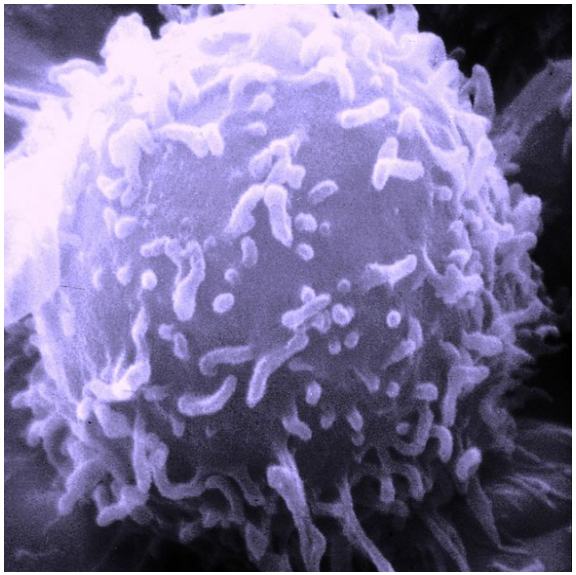
F(...attcaagcgc...) =



Cause (mutation and selection) and effect are millions of years and thousands of kilometers apart.

There is memory!

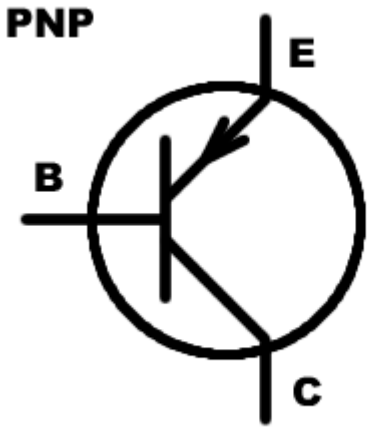
Kicking a tiger and treating a cancer cell are similar types of experiments

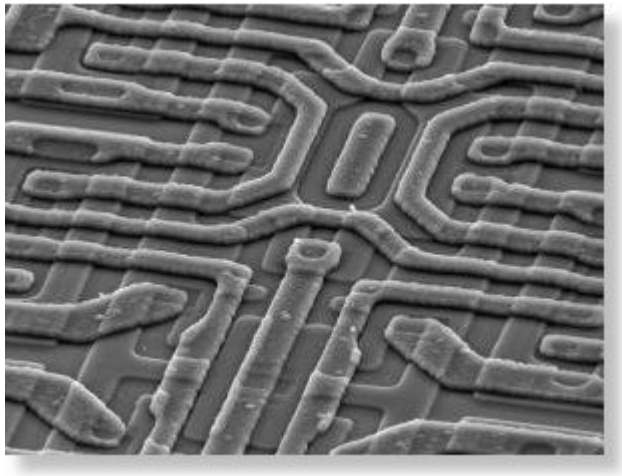


You kick the cancer cell with a drug and it processes input signals to output signals

The cell's genome (mostly) determines the outcome

Computers use transistors to compute





The most basic operations: and, or, plus,

The most basic operation of a cell is copying texts

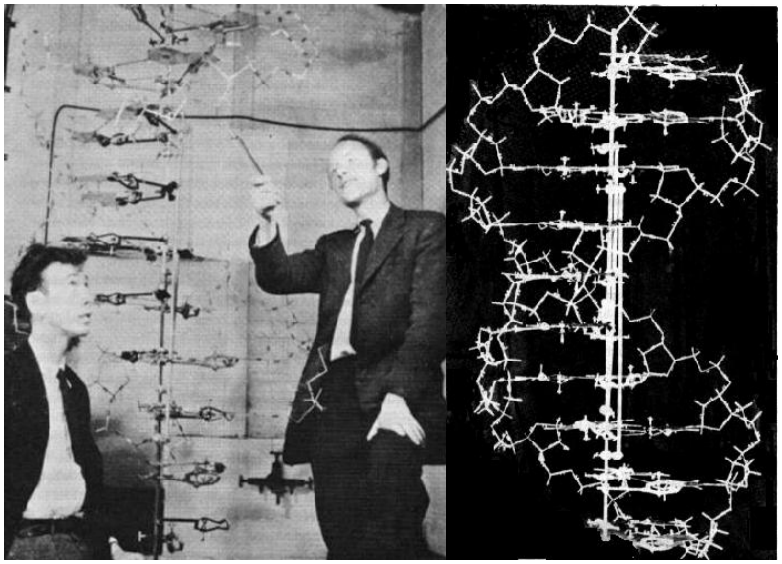
Template:

A C C G T A C **_ _ _ _ _ _ _**

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

Cells use enzymes

The clue to the copy operation lies in the chemical structure of DNA

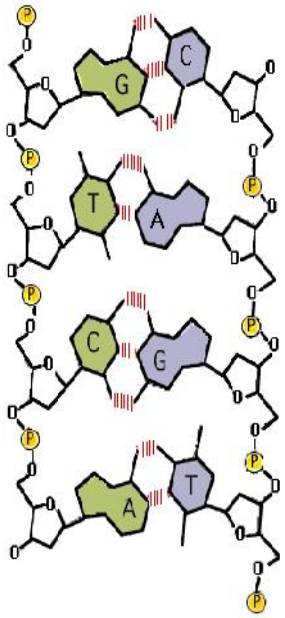


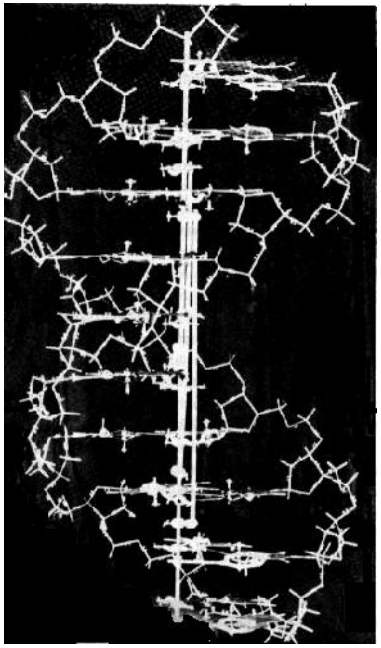
James Watson & Francis Crick

Nobel prize laureates 1962 for the discovery of the molecular structure of DNA

Why is the discovery of the double helix structure such an important finding?

The order of bases follows a rule





The double helix consists of two strands

The binding of the two strands to each other is weaker than the binding of nucleotides within a strand

Base Pairing Rule:
Opposite of a C is always a G and opposite of a T is always an A

"It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material."

Concluding remark in the paper by Watson and Crick announcing discovery of the structure of DNA

A double helix can be reconstructed from a single strand

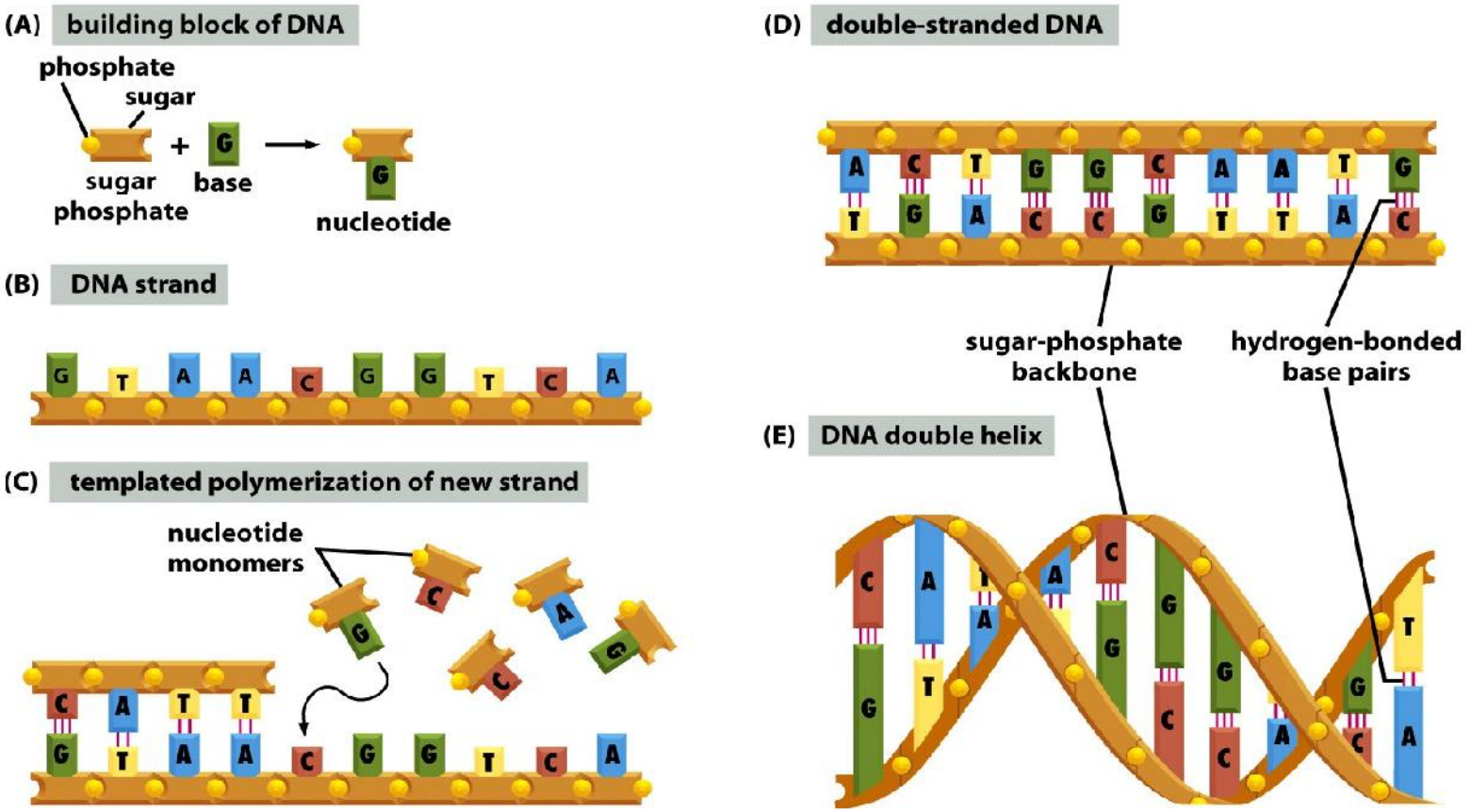


Figure 1-2 *Molecular Biology of the Cell*, Fifth Edition (© Garland Science 2008)

DNA replicates by templated polymerization

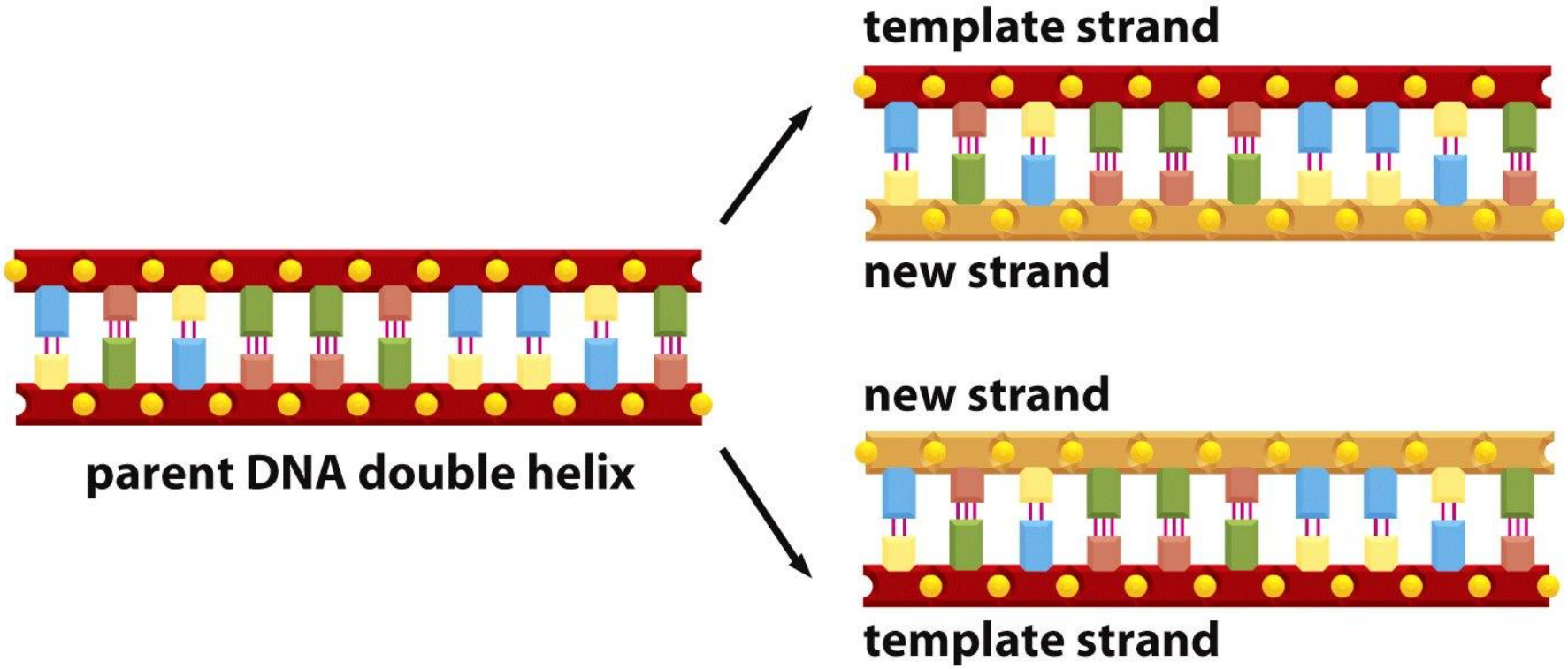
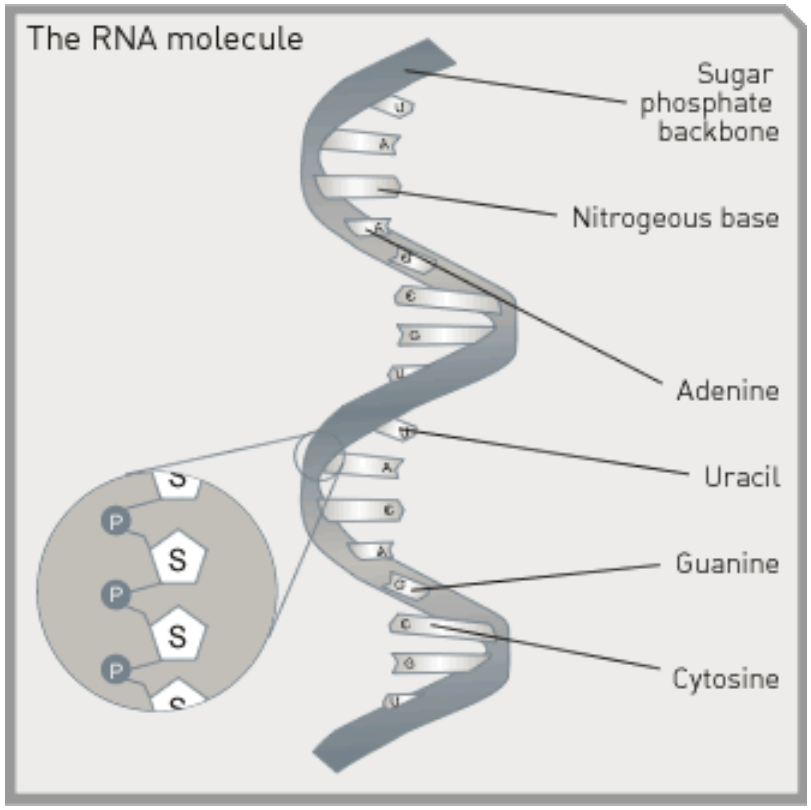


Figure 1-3 *Molecular Biology of the Cell*, Fifth Edition (© Garland Science 2008)

DNA is ROM! What about RAM?

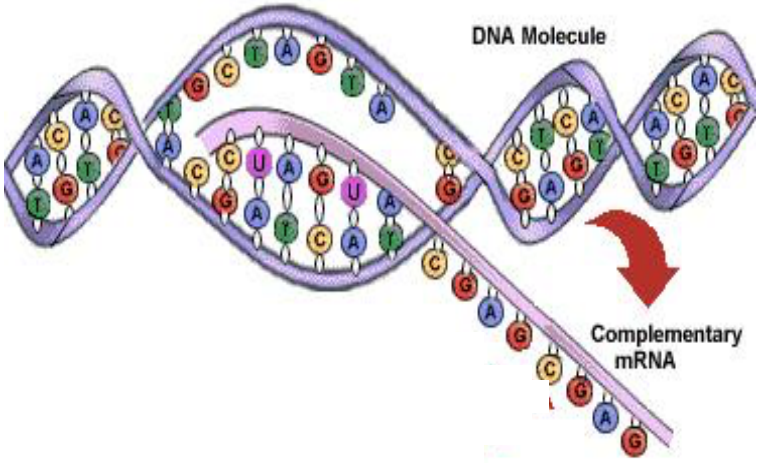
$$x \leftarrow x + 2$$

RNA is one form of transient memory



RNA is a single stranded nucleotide chain

RNA emerges by transcription of short segments of DNA



RNA molecules do not replicate them selves during cell division, but the offspring cells can produce new ones, using their DNA as template

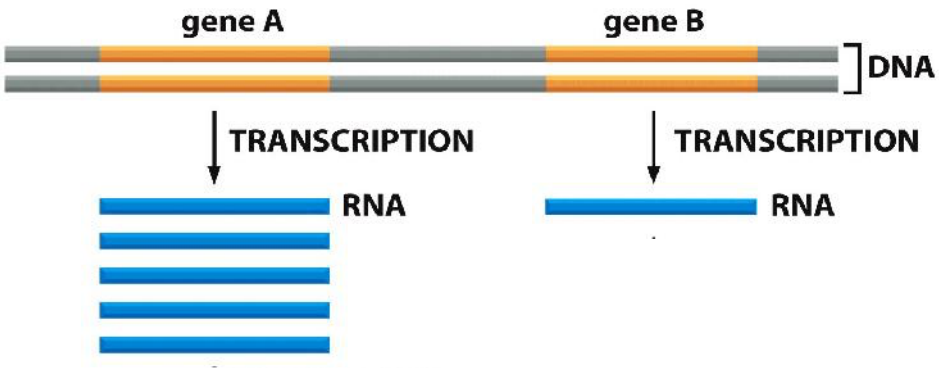
Again templated polymerization by base pairing is the guiding principal

Many different short segments of DNA are transcribed to RNA

Transcription is quantitative

A human cell contains two copies of each of its 23 different DNA molecules (chromosomes)

It contains thousands of different RNA molecules, whose abundance can be as little as 1-2 copies for one RNA sequence and reach several million copies for another RNA



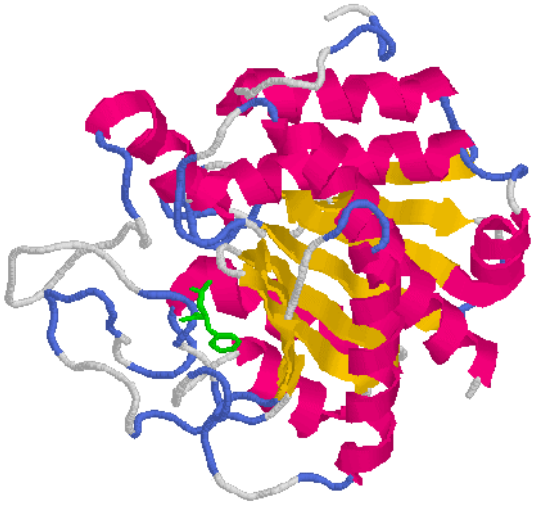
A cell can transiently store information in RNA copy numbers

Variable	Identifier	Value
RNA1:	atgagtccgcatcg...	56
RNA2:	ctgcgggcgttgct...	1.278.501
RNA3:	cggagtcgtcgccg...	0
RNA4:	gagtcgcggggatc...	9.444
...		
RNA72.314:	acggtgtccgcgat...	699

RNA: Thousands of integer variables (read and write)
DNA: 23 string variables (read only)

All values stored in RNA abundances can be combined in a high dimensional vector (the expression profile of the cell)

Proteins are another form of transient information



```
MGDVEKGKKIFVQKCAQCHT
VEKGGKHKTGXNLHGLFGRK
TGQAAGFSYTDANKNKGITW
GEDTLMEYLENPKKYIPGTK
MIFAGIKKKGERADLIAYLK
KATNE
```

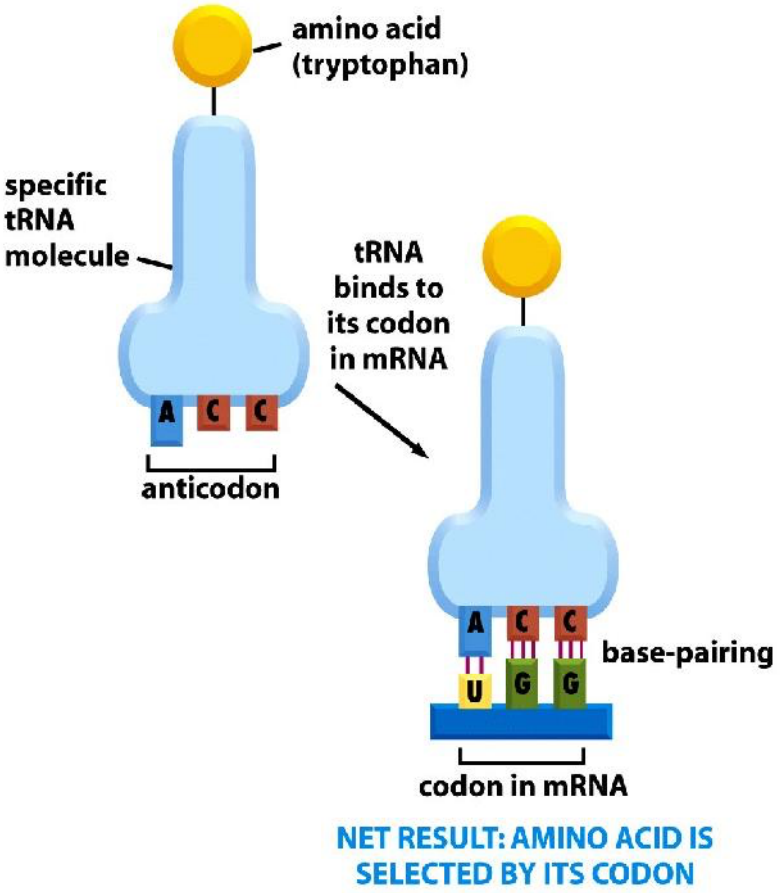
Proteins are chain molecules consisting of amino acids

There are 20 different amino acids found in proteins

Proteins are texts in a language that uses an alphabet of 20 letters

How does a cell make protein?

Every amino acid is matched by a t-RNA molecule that can load it



The t-RNA can bind DNA that is complementary to a characteristic three base sequence (anticodon)

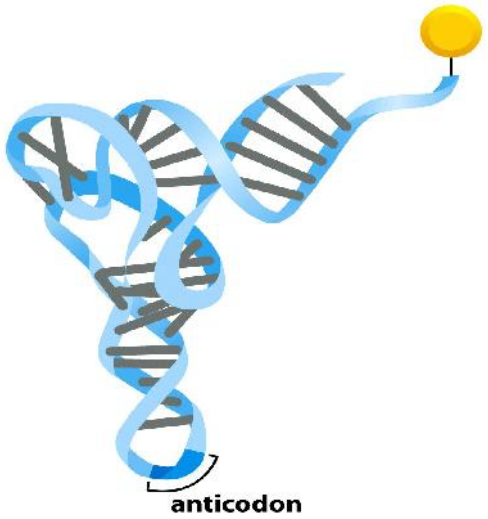


Figure 1-9 *Molecular Biology of the Cell*, Fifth Edition (© Garland Science 2008)

Ribosomes assemble proteins following the information stored in RNA molecules applying the genetic code

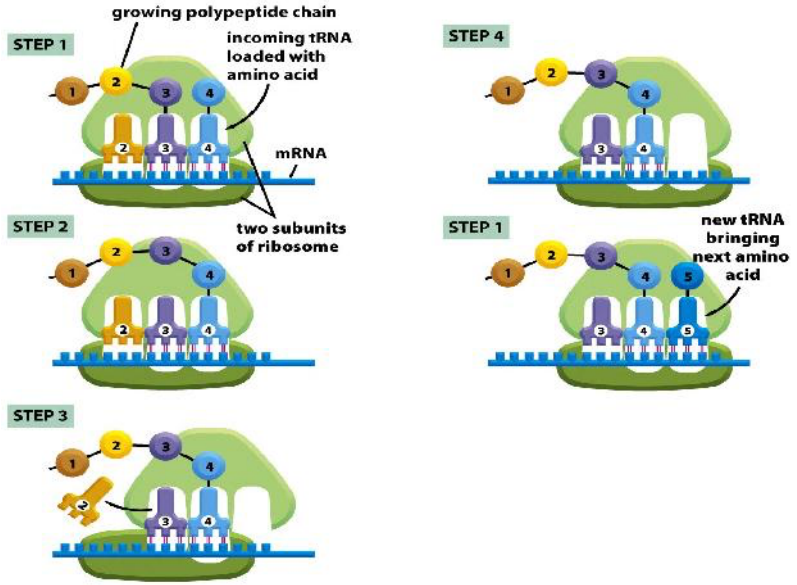


Figure 1-10a *Molecular Biology of the Cell*, Fifth Edition (© Garland Science 2008)

There is no new information on the protein the RNA information is just translated

The Genetic Code

AAA	Lys	ACA	Thr	AGA	Arg	AUA	Ile
AAC	Asn	ACC	Thr	AGC	Ser	AUC	Ile
AAG	Lys	ACG	Thr	AGG	Arg	AUG	Met
AAU	Asn	ACU	Thr	AGU	Ser	AUU	Ile
CAA	Gln	CCA	Pro	CGA	Arg	CUA	Leu
CAC	His	CCC	Pro	CGC	Arg	CUC	Leu
CAG	Gln	CCG	Pro	CGG	Arg	CUG	Leu
CAU	His	CCU	Pro	CGU	Arg	CUU	Leu
GAA	Glu	GCA	Ala	GGA	Gly	GUA	Val
GAC	Asp	GCC	Ala	GGC	Gly	GUC	Val
GAG	Glu	GCG	Ala	GGG	Gly	GUG	Val
GAU	Asp	GCU	Ala	GGU	Gly	GUU	Val
UAA	Stop	UCA	Ser	UGA	Stop	UUA	Leu
UAC	Tyr	UCC	Ser	UGC	Cys	UUC	Phe
UAG	Stop	UCG	Ser	UGG	Trp	UUG	Leu
UAU	Tyr	UCU	Ser	UGU	Cys	UUU	Phe

Some triplets (codons) do not encode for an amino acid but cause the termination of translation *stop codons*

All other triplet define a unique amino acid, but not vice versa

Proteins do not only store information but also use it

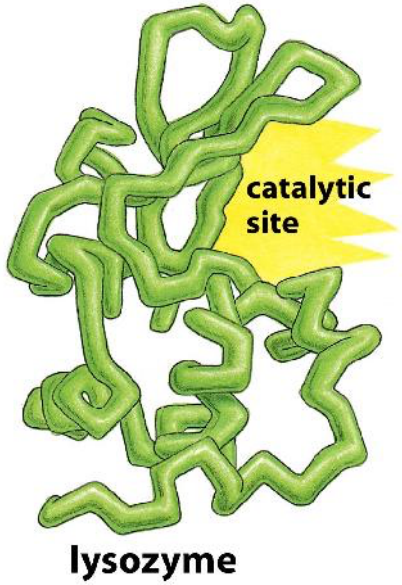
Amino acids interact differently than nucleotides leading to a characteristic 3 dimensional structure of the protein

The structure is self organizing and defined by the sequences

It determines:

- chemical activity

- binding specificity



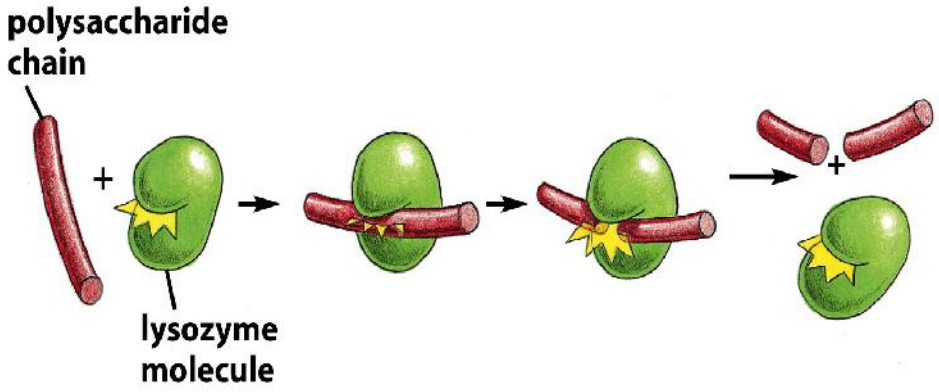


Figure 1-7 *Molecular Biology of the Cell*, Fifth Edition (© Garland Science 2008)

Proteins are executable

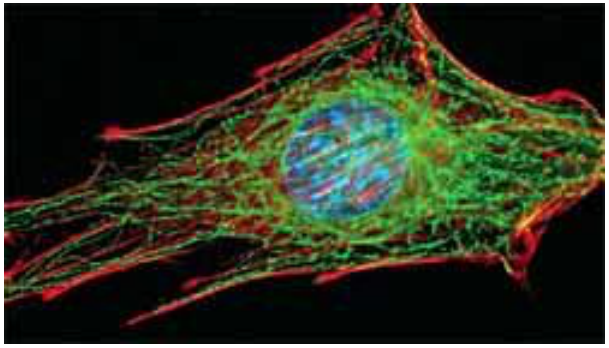
Maintain cellular structure

Digest nutrients

Catalyze enzymatic reactions

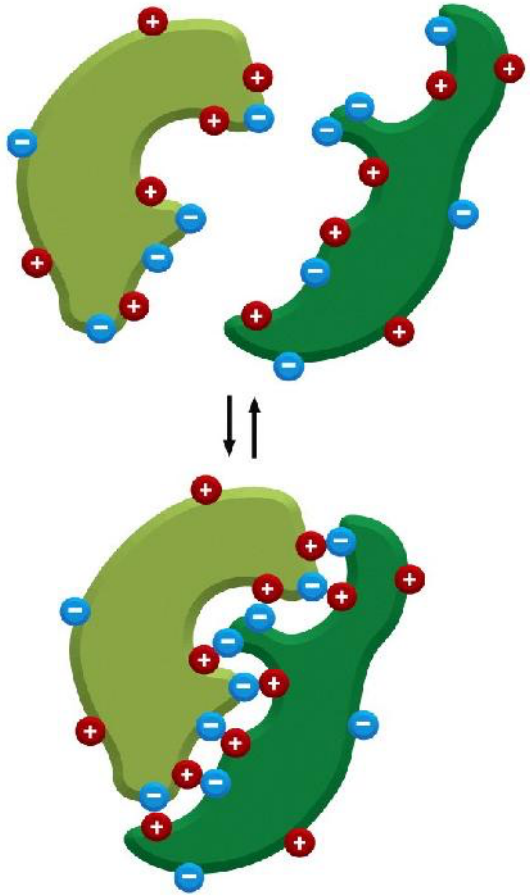
Synthesize Biomolecules

Process Information



Structural Proteins in the cell skeleton

Proteins interact with each other and with other molecules using a key-lock system



DNA information determines RNA information,

which determines protein sequences,

which determines the structure,

which determines the interaction partners,

which determines the function

Figure 2-16 *Molecular Biology of the Cell* (© Garland Science 2008)

Proteins can have several states like on and off

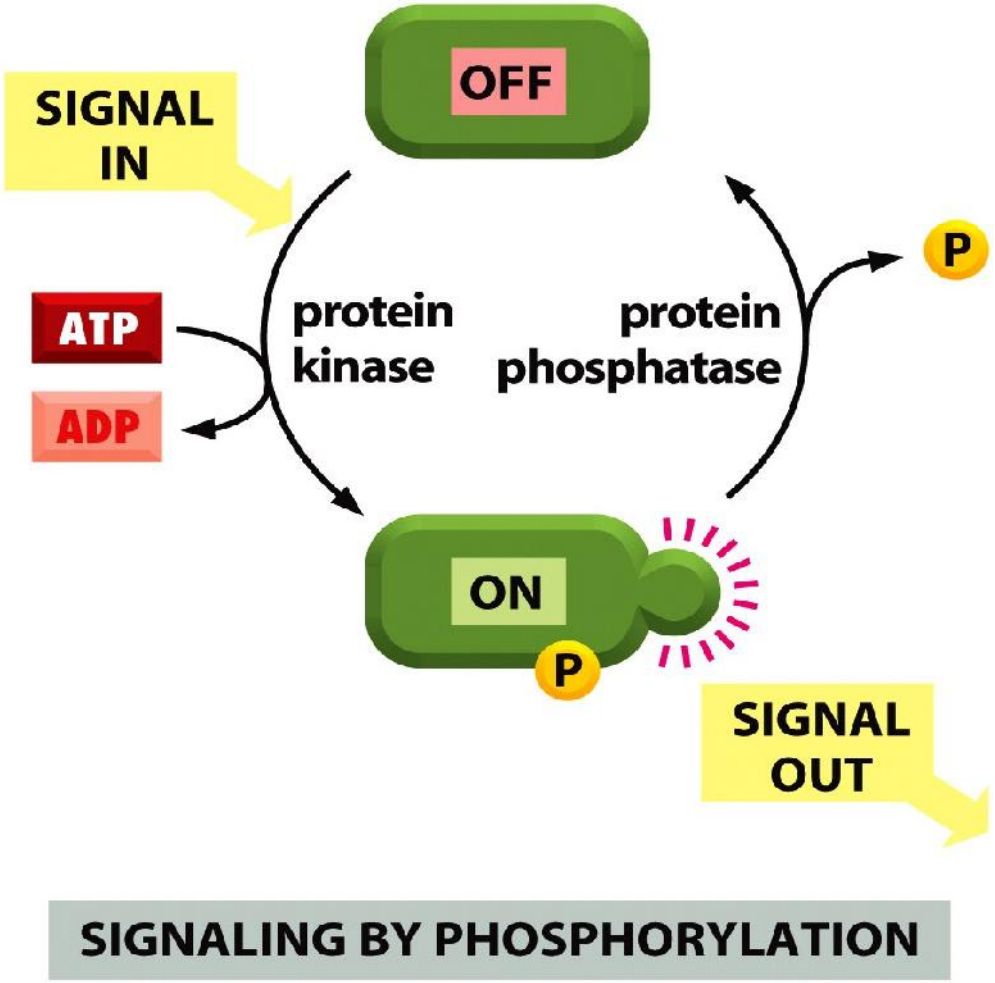


Figure 15-18a *Molecular Biology of the Cell* (© Garland Science 2008)

Proteins store information transiently in their sequences, in their abundances, and in their states

How does the cell control when to write on which RNA memory?

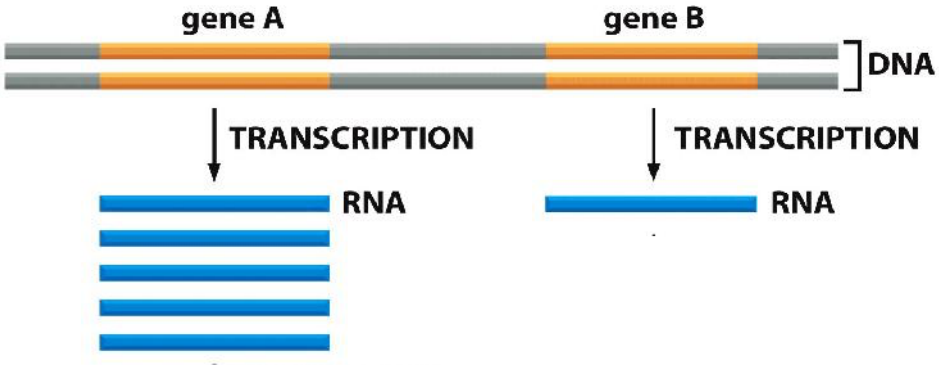
Variable	Identifier	Value
RNA1:	atgagtccgcatcg...	56
RNA2:	ctgcgggcgttgct...	1.278.501
RNA3:	cggagtcgtcgccg...	0
RNA4:	gagtcgcggggatc...	9.444
...		
RNA72.314:	acggtgtccgcgat...	699

```
If ( x== true)
  {
    RNA1=RNA1+17
  }
end
```

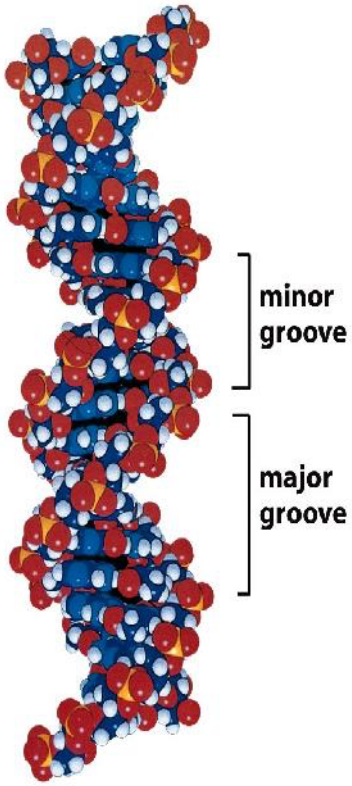
By Gene Regulation

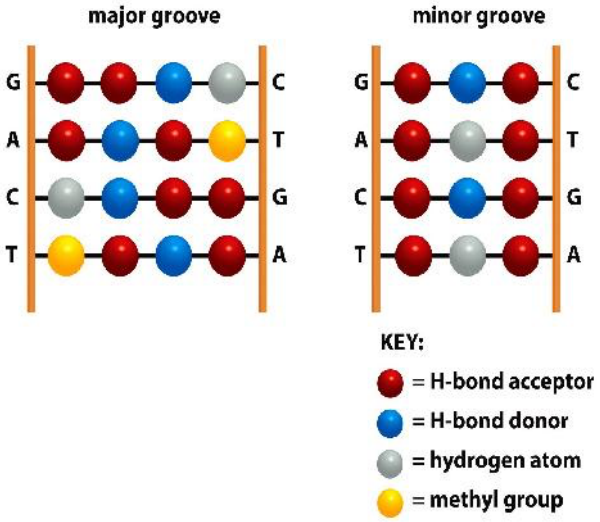
By controlling the amount and the sequence of transcribed RNA molecules

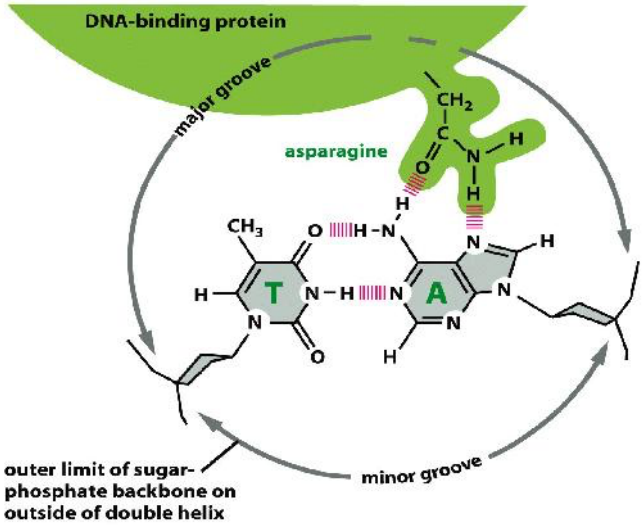
Variable	Identifier
RNA1:	atgagtccgcatcg...
RNA2:	ctgcgggcgttgct...
RNA3:	cggagtcgtcgccg...
RNA4:	gagtcgcggggatc...
...	
RNA72.314:	acggtgtccgcgat...



The outside of the DNA double helix can be read by proteins called transcription factors







Figures 7-6,7-8,7-9 *Molecular Biology of the Cell* (© Garland Science 2008)

Transcription factors bind to specific short regulatory sequences that they can recognize on the DNA

Table 7–1 Some Gene Regulatory Proteins and the DNA Sequences That They Recognize

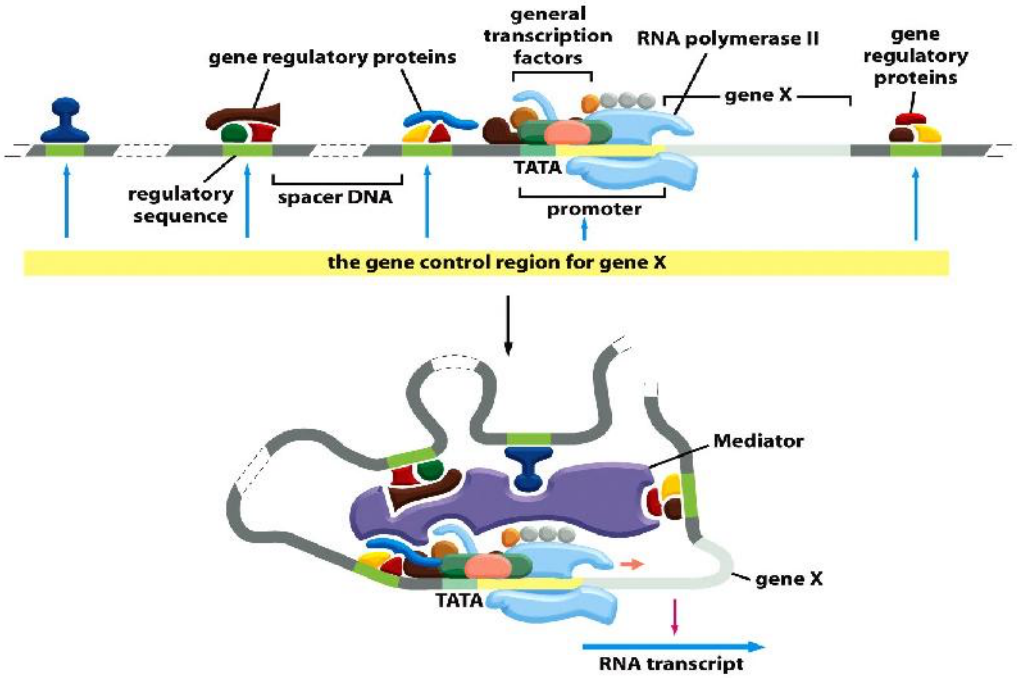
	NAME	DNA SEQUENCE RECOGNIZED*
Bacteria	Lac repressor	5′ AATTGTGAGCGGATAACAATT 3′ TTAACACTCGCCTATTGTTAA
	CAP	TGTGAGTTAGCTCACT ACACTCAATCGAGTGA
	Lambda repressor	TATCACCGCCAGAGGT ATAGTGGCGGTCTCCAT
Yeast	Gal4	CGGAGGACTGTCCTCCG GCCTCCTGACAGGAGGC
	Matα2	CATGTAATT GTACATTAA
	Gcn4	ATGACTCAT TACTGAGTA
Drosophila	Kruppel	AACGGGTTAA TTGCCCAATT
	Bicoid	GGGATTAGA CCCTAATCT
Mammals	Sp1	GGGCGG CCCGCC
	Oct1 Pou domain	ATGCAAAT TACGTTTA
	GATA1	TGATAG ACTATC
	MyoD	CAAATG GTTTAC
	p53	GGGCAAGTCT CCCGTTCAGA

*For convenience, only one recognition sequence, rather than a consensus sequence (see Figure 6–12), is given for each protein.

Binding of an amino acid in a protein to a nucleotide in DNA depends on the sequence context both in the protein and the DNA

Table 7-1 *Molecular Biology of the Cell* (© Garland Science 2008)

The transcription factors recruit the translation machinery to the start sites of genes



The transcription machinery is the reading head of the hard disk

The short regulatory sequences are the file system of the cell

They help the cell find programs on the genome

Figure 7-44 *Molecular Biology of the Cell* (© Garland Science 2008)

Some transcription factors do not activate but repress transcription

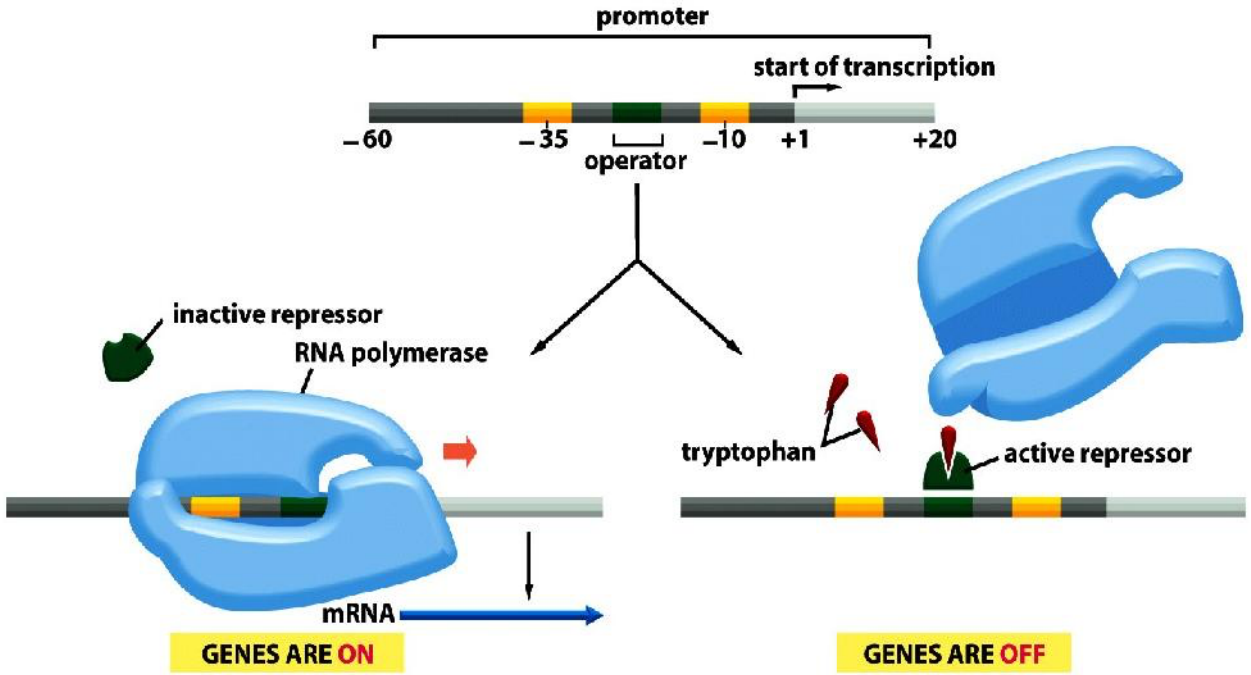
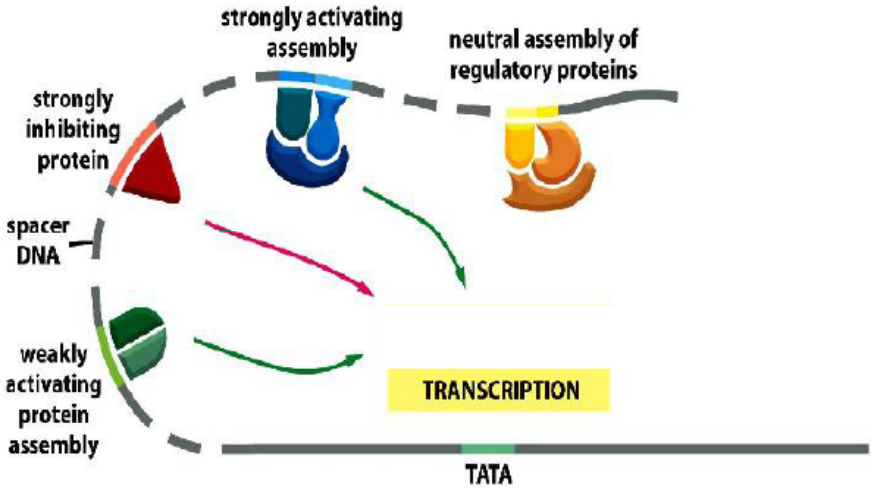


Figure 7-35 *Molecular Biology of the Cell* (© Garland Science 2008)

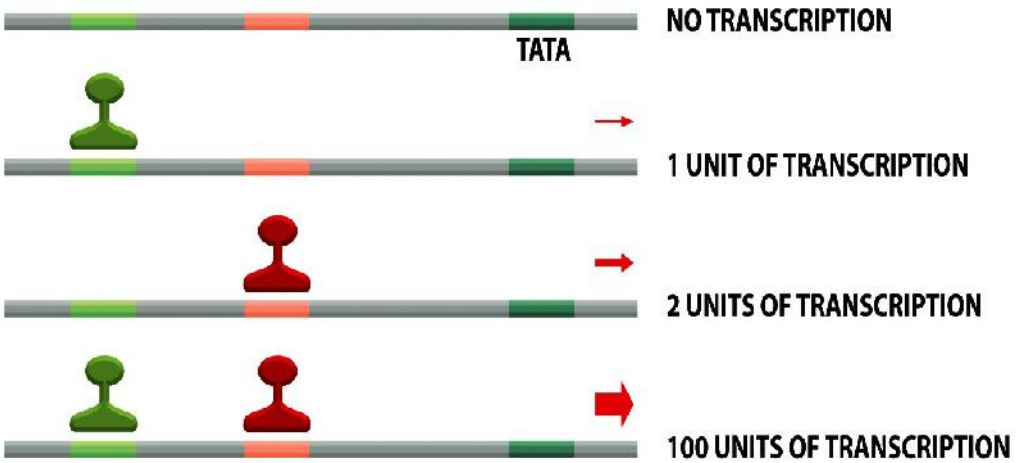
Every gene needs a certain combination of present activators and absent repressors for transcription




```
If (activators a,b,c = present
    and repressor d absent)
{
   begin(transcription)
}
endif
```


Figure 7-58 *Molecular Biology of the Cell* (© Garland Science 2008)

Transcription can be regulated quantitatively by different combinations of transcription factors




```
If (a=0 and b=1)
{
    RNA = 1
}
Elseif (a=1 and b=0)
{
    RNA = 2
}
Elseif (a=1 and b=1)
{
    RNA = 100
}
Else %No TF at all
{
    RNA = 0
}
endif
```


Figure 7-48 *Molecular Biology of the Cell* (© Garland Science 2008)

The same transcription factor is typically used for the regulation of many genes

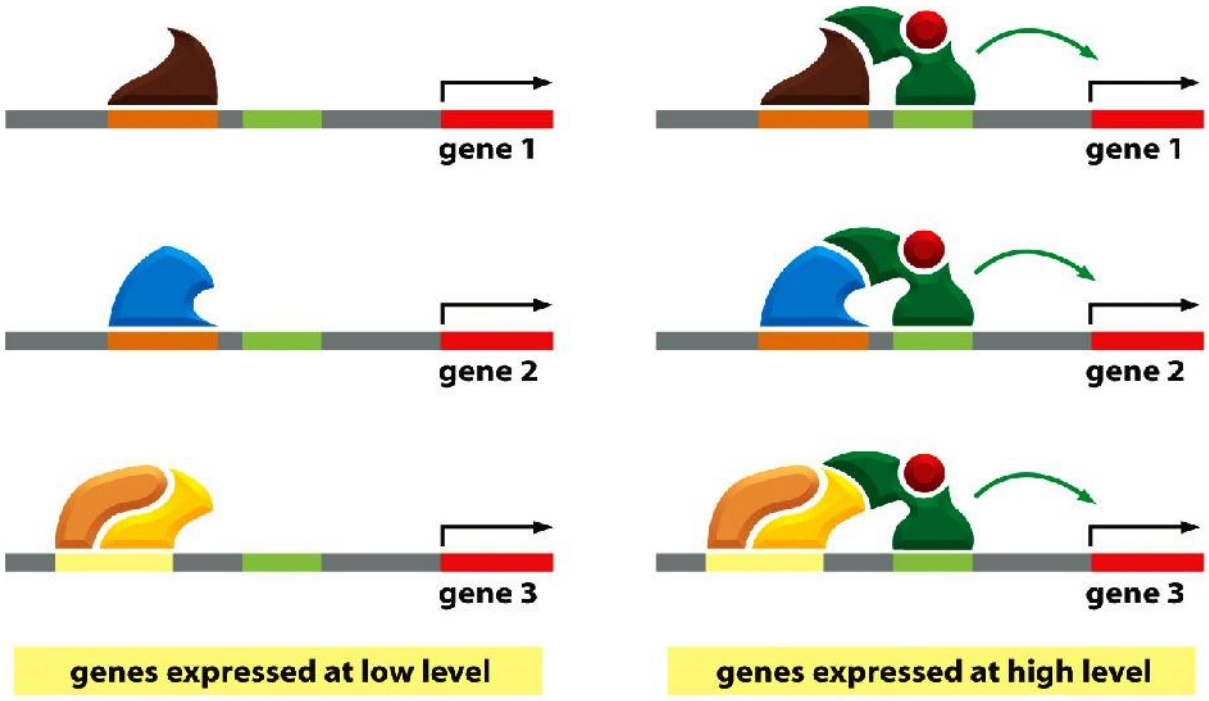


Figure 7-74 *Molecular Biology of the Cell* (© Garland Science 2008)

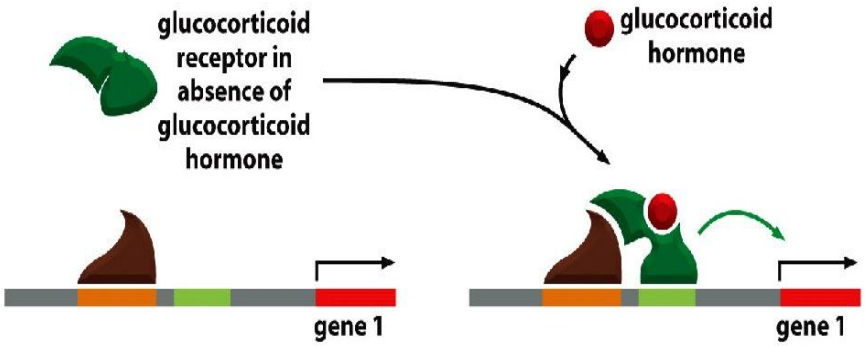
How does the cell control the logical expressions that control transcription?

Variable	Identifier	Value
RNA1:	atgagtccgcatcg...	56
RNA2:	ctgcgggcgttgct...	1.278.501
RNA3:	cggagtcgtcgccg...	0
RNA4:	gagtcgcggggatc...	9.444
...		
RNA72.314:	acggtgtccgcgat...	699

```
If ( x== true)
  {

RNA1=RNA1+17
  }
end.
```

By modifying the transcription factors
By modifying the DNA

Many transcription factors need to be switched on before they work



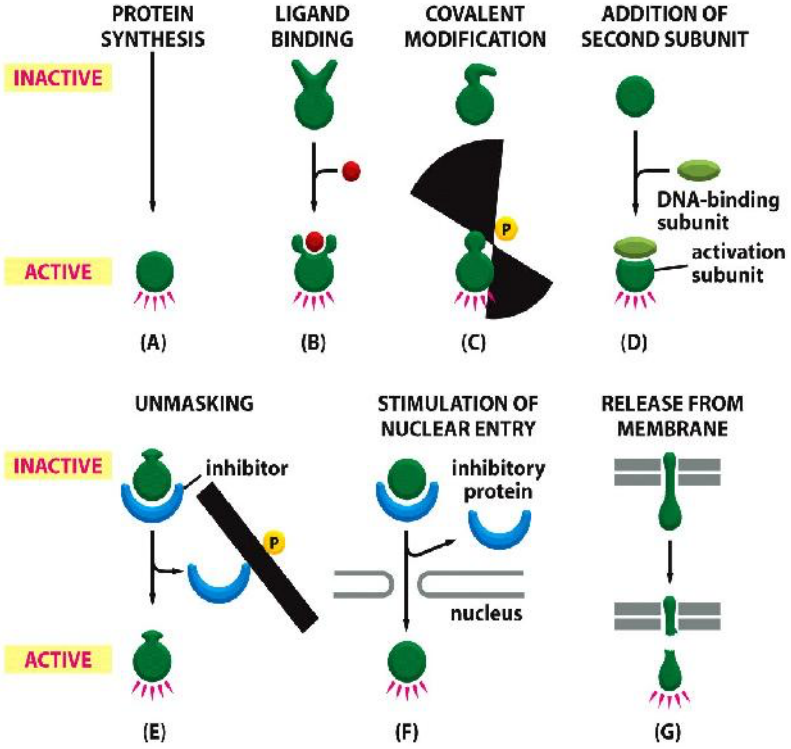
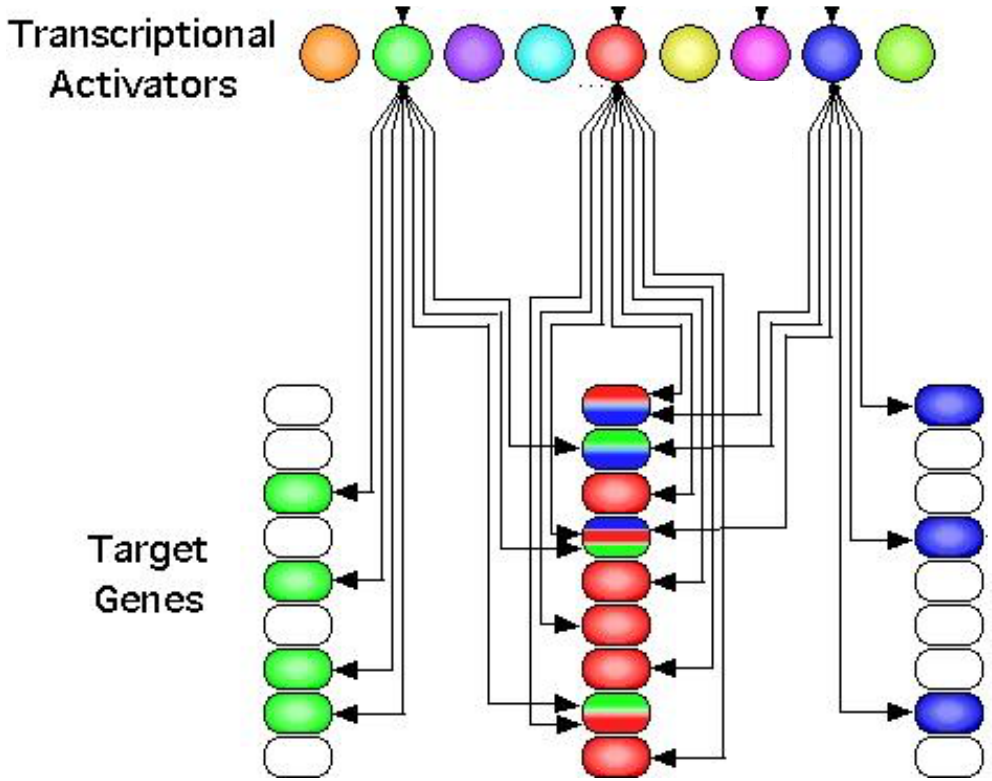


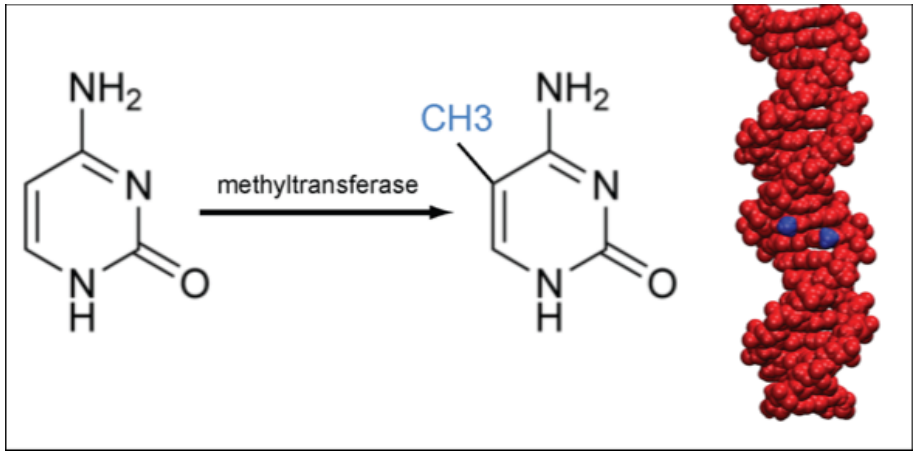
Figure 7-74 *Molecular Biology of the Cell* (© Garland Science 2008)

A characteristic expression profile is the consequence of characteristic constellation of transcription factor activities



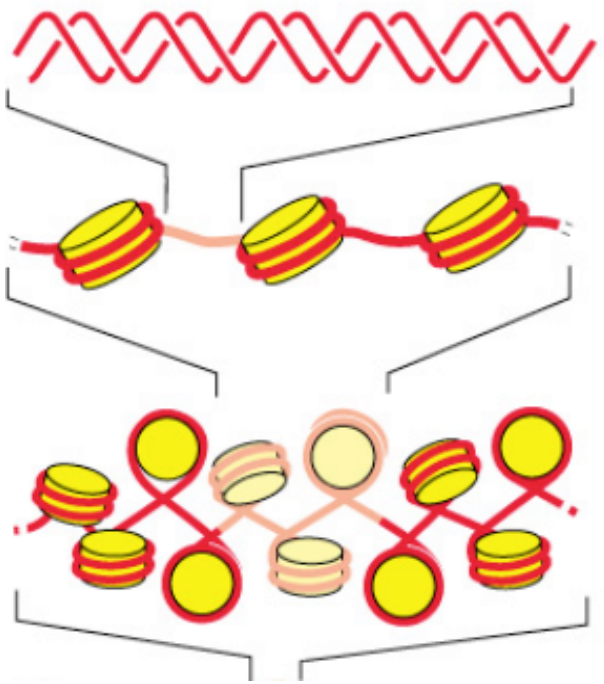
How are transcription factors activated?

DNA methylation is another type of transient memory (more long term)



CG → ĆG

DNA is not naked



Chromatin = DNA + Proteins

Figure 7-74 *Molecular Biology of the Cell* (© Garland Science 2008)

Parts of the DNA are accessible to transcription factors others are not

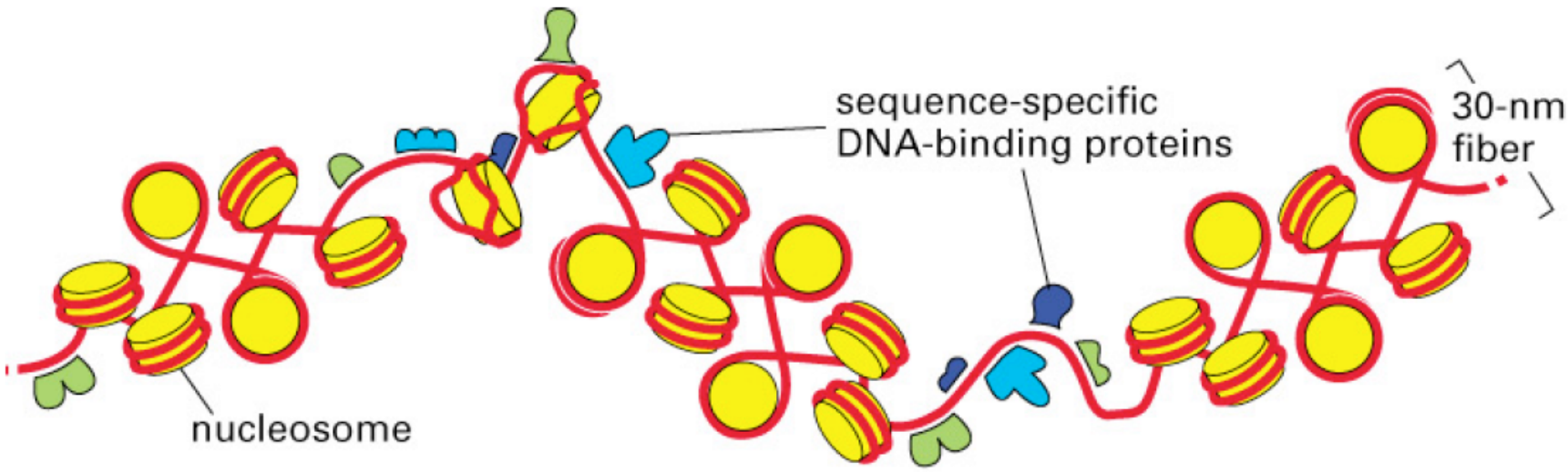
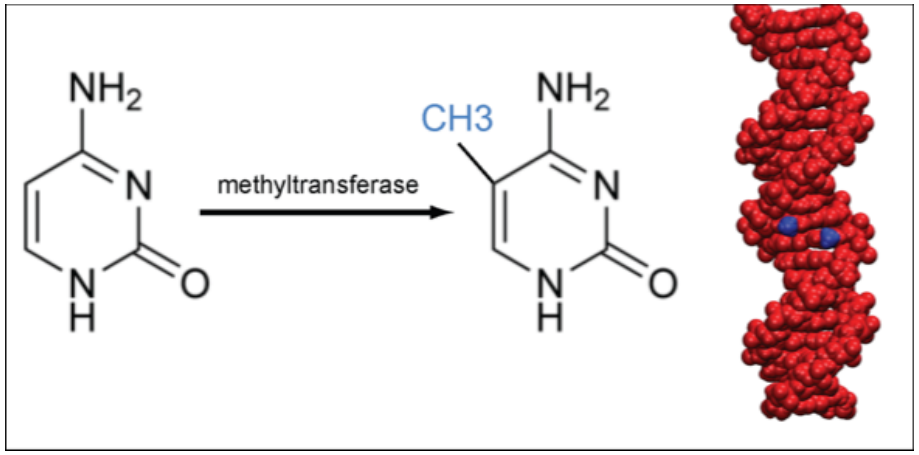


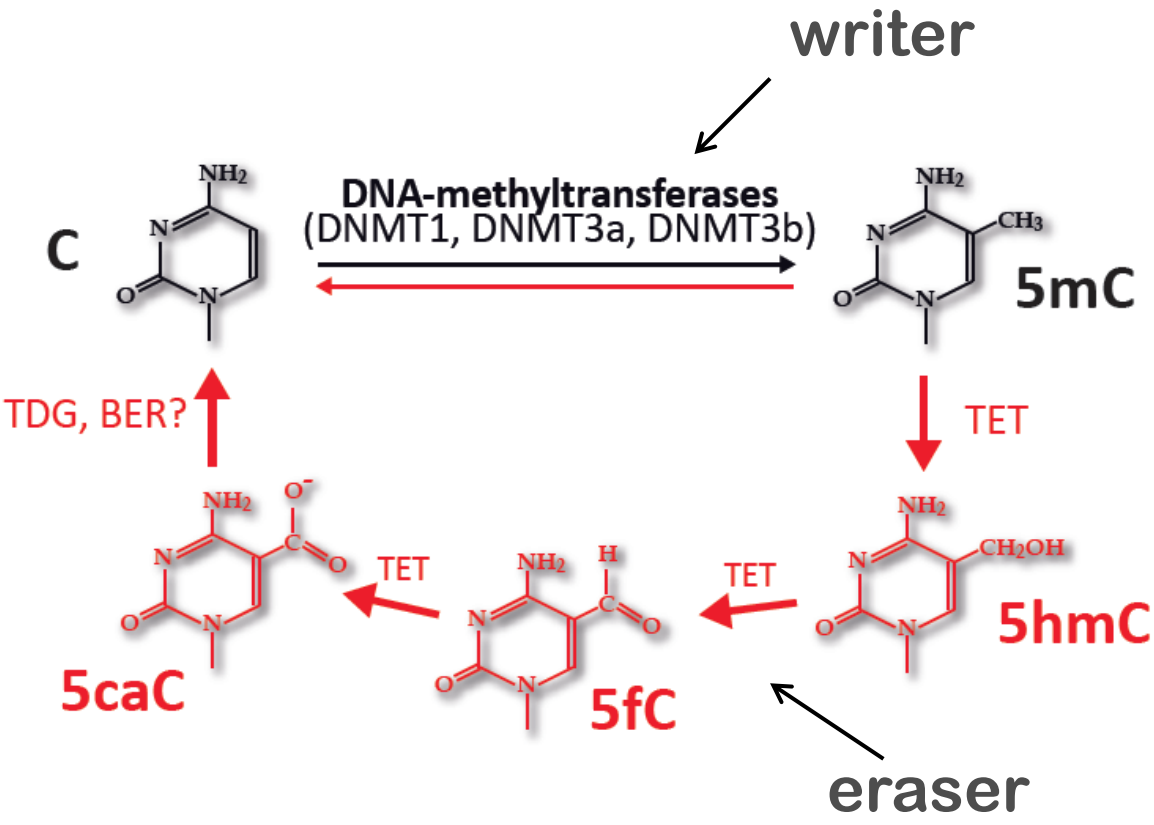
Figure 7-74 *Molecular Biology of the Cell* (© Garland Science 2008)

DNA methylation affects the accessibility of DNA



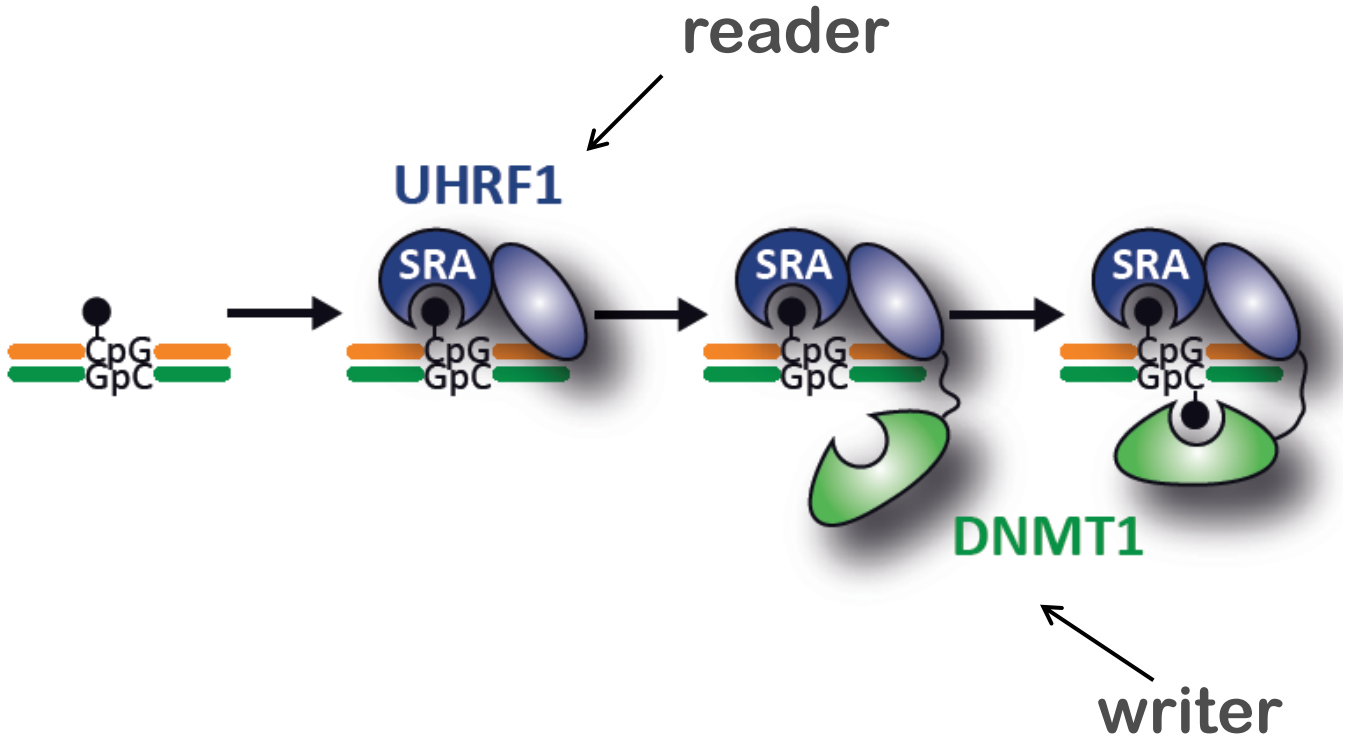
CG → ĆG

Our cells have readers, writers, and erasers of DNA methylation

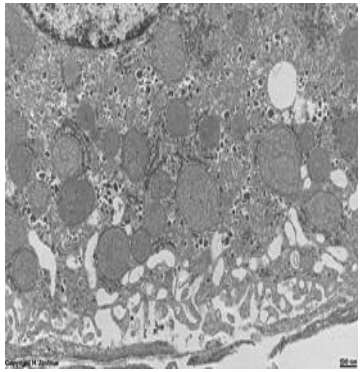


Readers, writers, and erasers are proteins (enzymes)

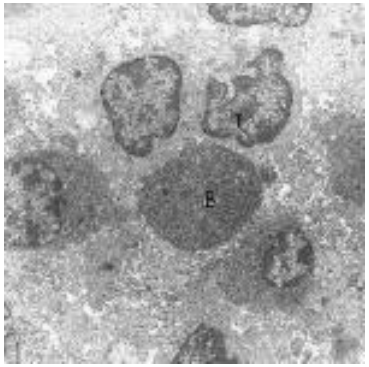
Methylation is preserved during DNA duplication



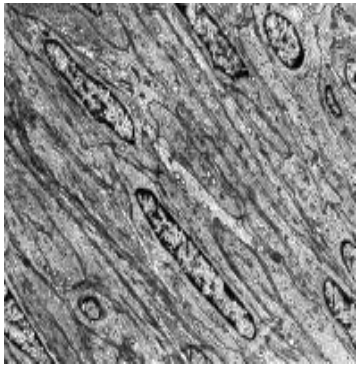
In different cells of the body different parts of the DNA are accessible



Liver



B-Cell



Muscle

Methylation controls file system permissions

Gene: BCL6			
B-cell			r x w
Liver	r	x	-
Brain	-	-	-
Muscle			- - -
Kidney			r x -
...			

How does the cell receive inputs?

Receptors in the cell membrane communicate outer signals into the cell

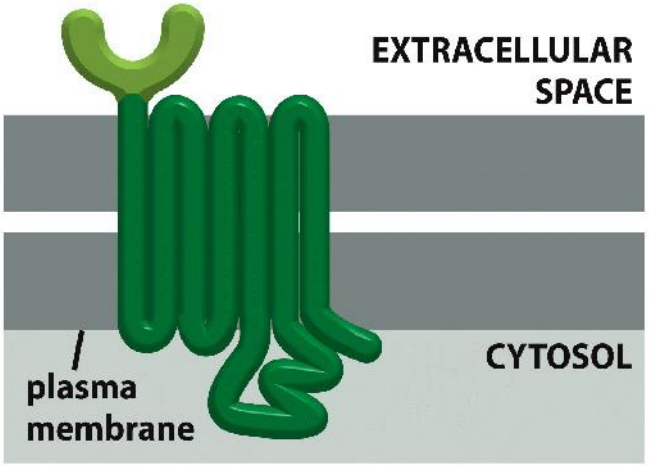
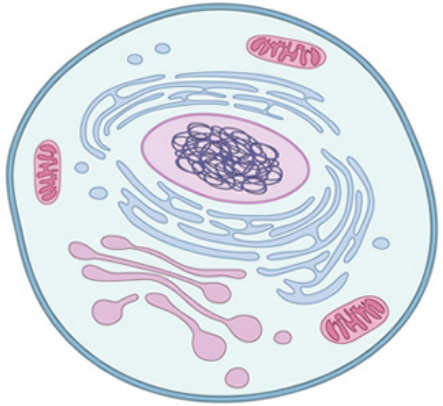


Figure 15-30 *Molecular Biology of the Cell* (© Garland Science 2008)

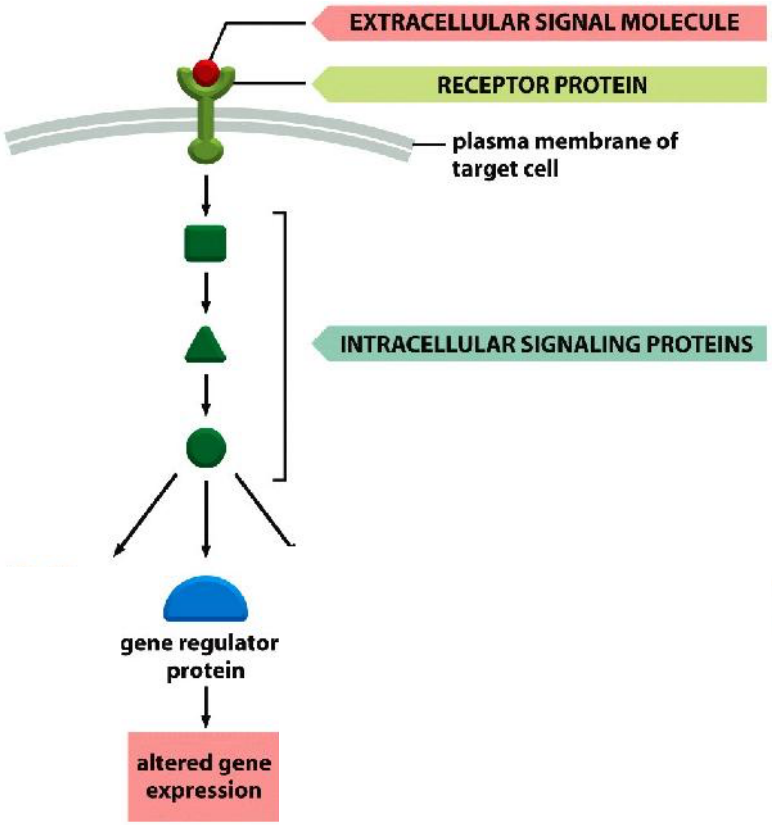
Receptors are the USB ports of a cell

Signals need to be transmitted into the nucleus (molecular signaling)

DNA is in the nucleus



Transcription factors can be activated by signaling pathways in response to incoming signals

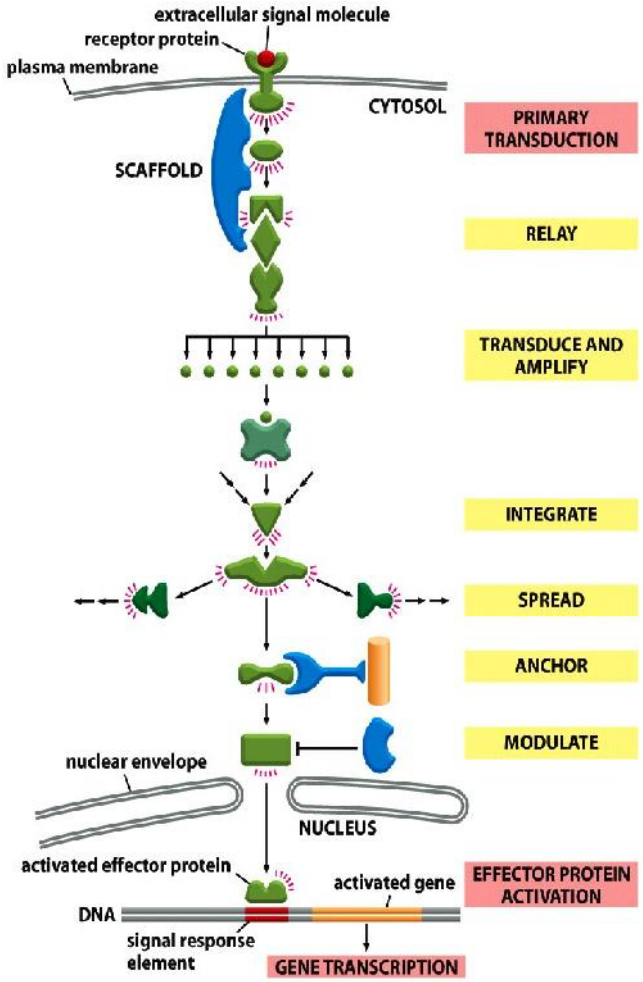


An incoming signal from a receptor activates protein A, which activates protein B, ..., which activates transcription factor X

Different cell types signal differently

Figure 15-1 *Molecular Biology of the Cell* (© Garland Science 2008)

Signaling pathways are sequences of computations

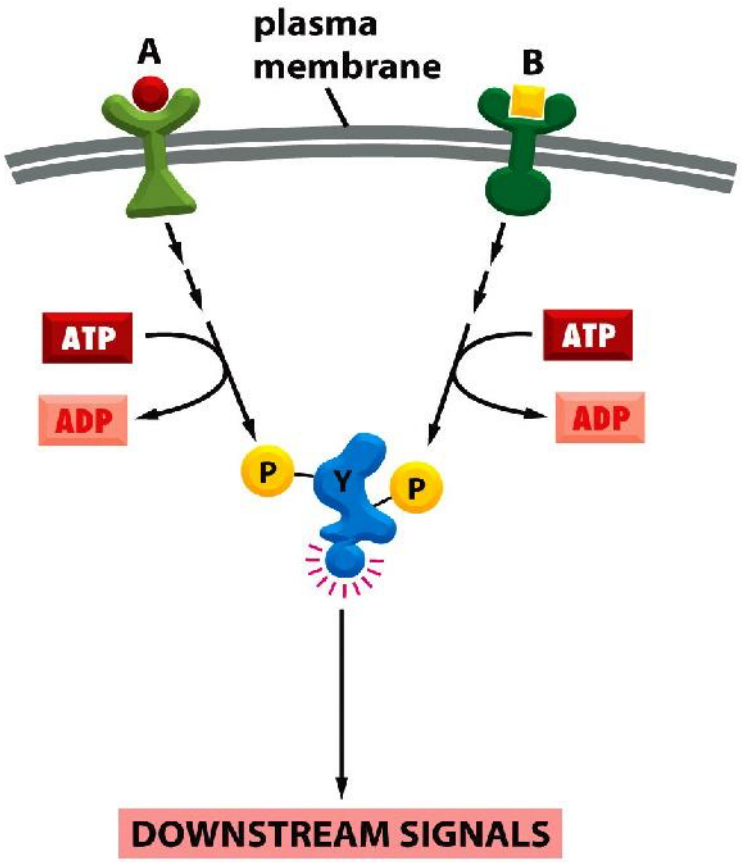


Every step can be controlled

```
If (logical expression)
{
    next signaling step
}
endif
```


Figure 15-17 *Molecular Biology of the Cell* (© Garland Science 2008)

Different signaling pathways can exchange information during runtime (cross talk)

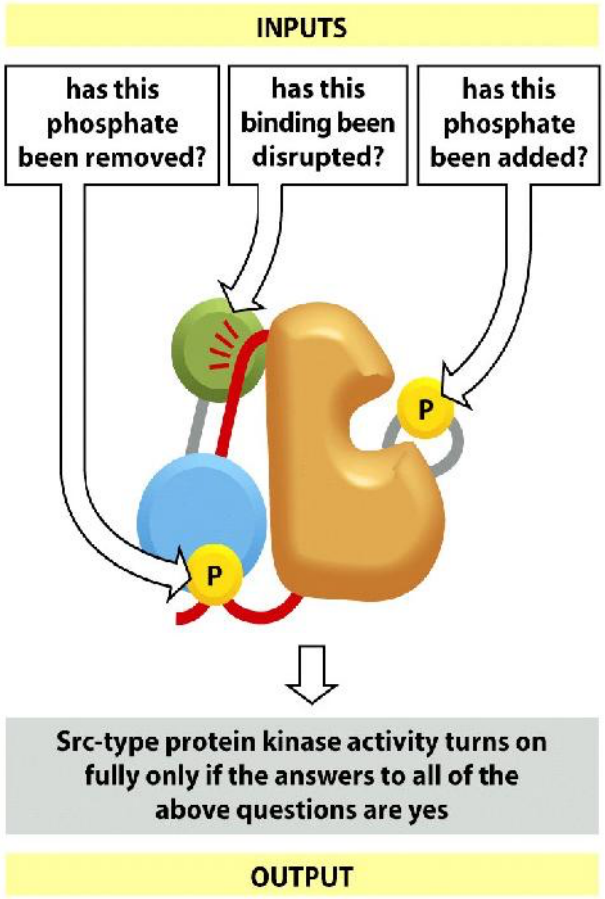


Y is only activated if two pathways send signals

Inter process communication

Figure 15-20 *Molecular Biology of the Cell* (© Garland Science 2008)

Complexes of signaling molecules can integrate the values of several variables




```
Function complex1(a,b,c)
{
  If (a,b,c = true)
  {
    complex1 = true
  }
  Else
  {
    complex1 = false
  }
Endif

Activate d if complex1 = true
```

Figure 3-70 *Molecular Biology of the Cell* (© Garland Science 2008)

Signaling feedback can create oscillators (clocks)

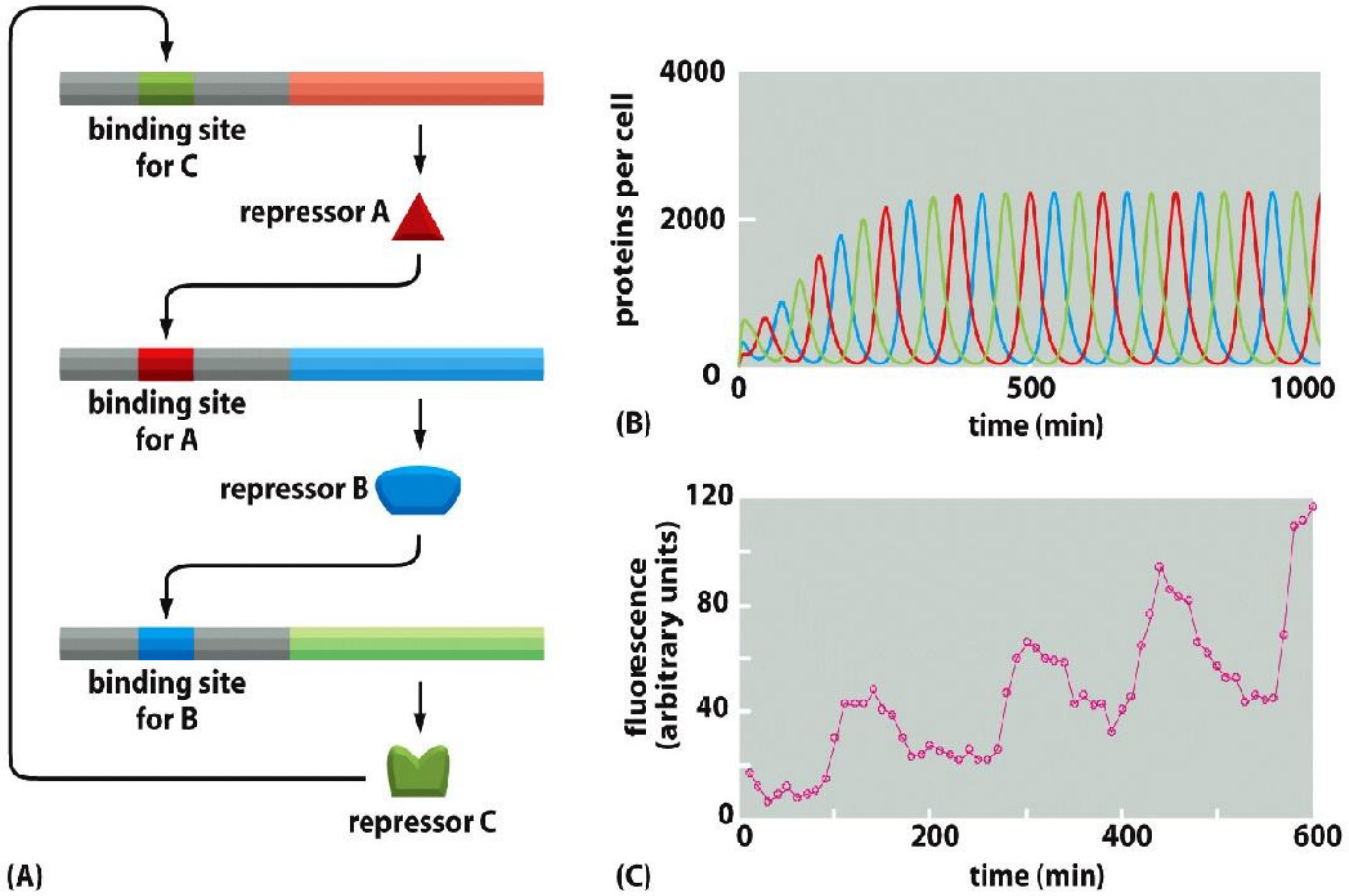


Figure 7-72 *Molecular Biology of the Cell* (© Garland Science 2008)

Feed forward loops can smooth signals

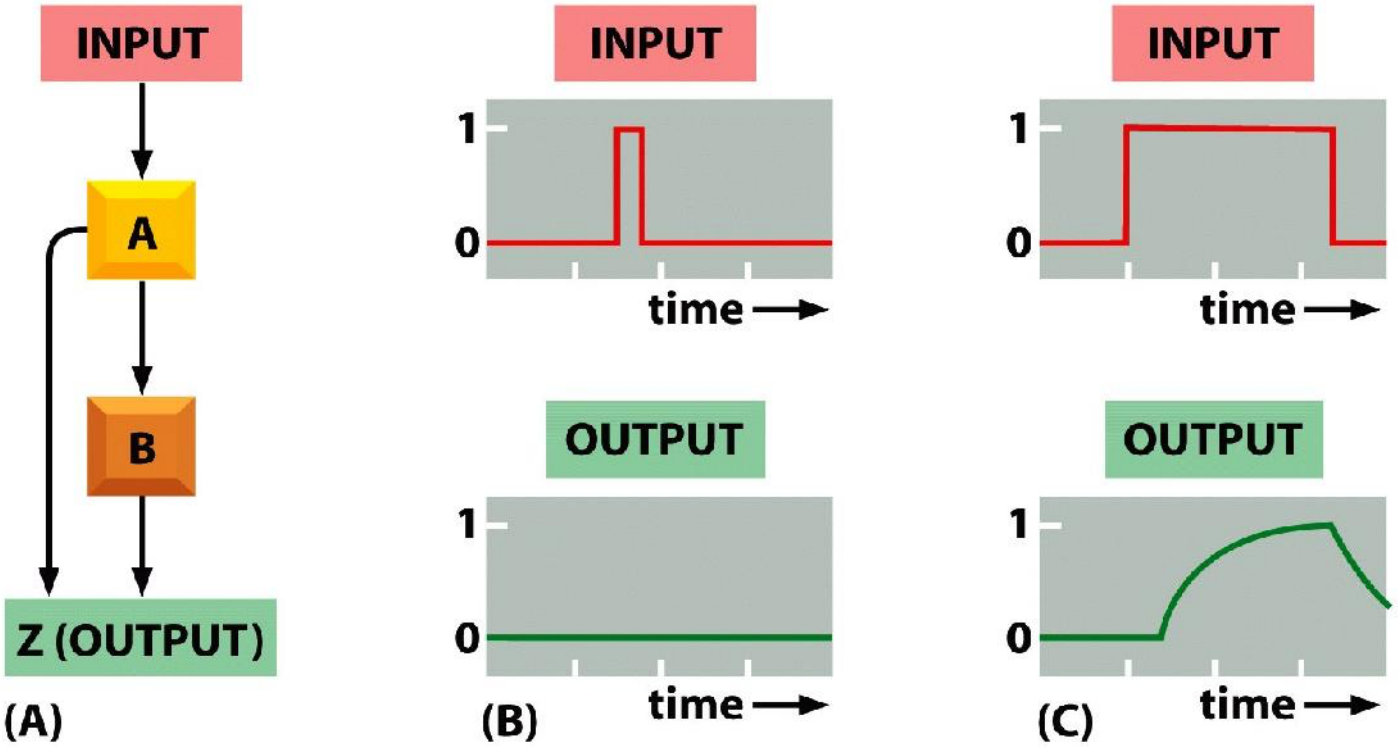


Figure 7-70 *Molecular Biology of the Cell* (© Garland Science 2008)

Cells compute by a complex circuitry of signaling pathways

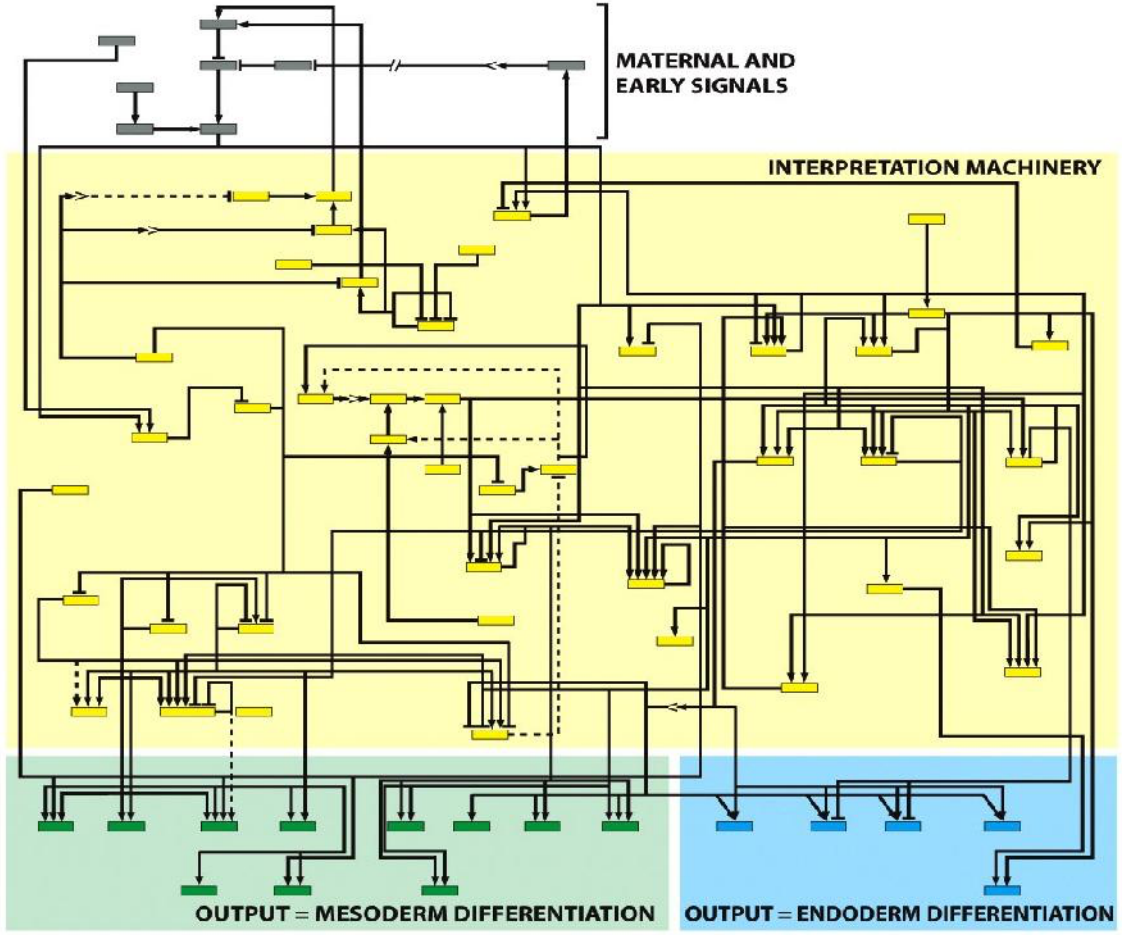
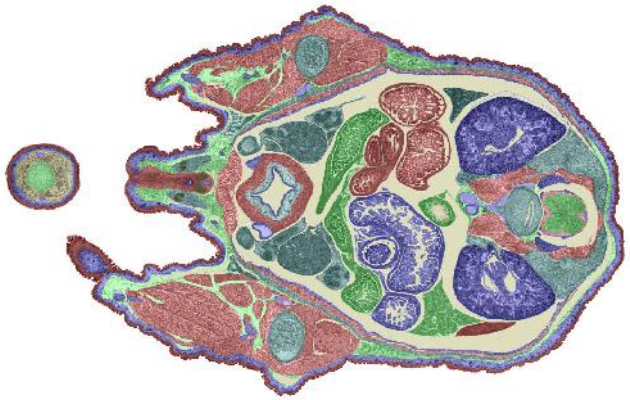


Figure 7-71 *Molecular Biology of the Cell* (© Garland Science 2008)

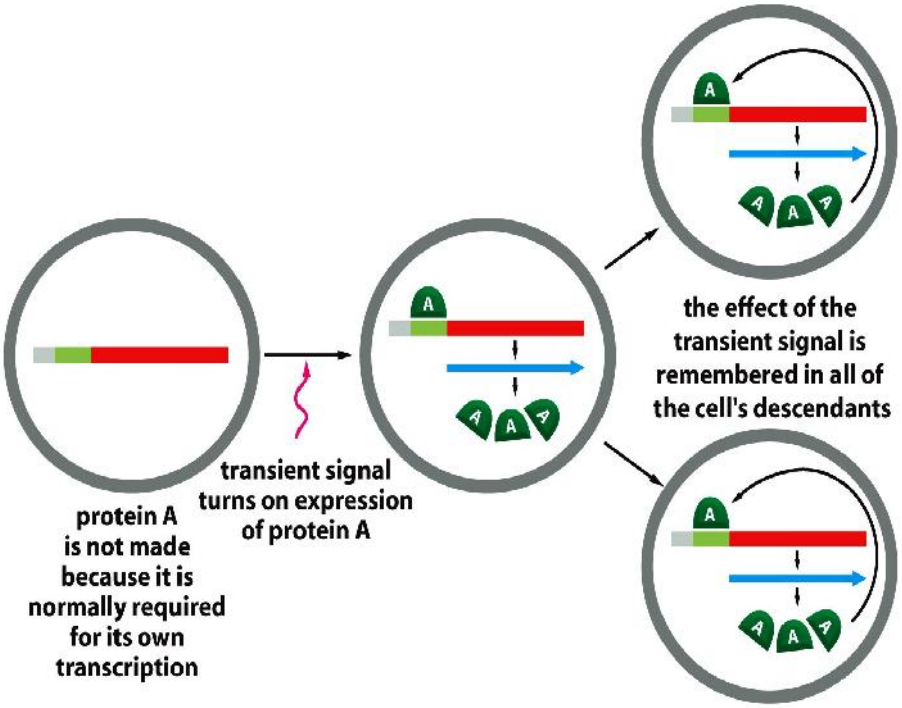
Different cells have identical ROM but different RAM



Stained proteins in different tissues

How does a cell remember whether it is liver or brain?

Cells can create transient memory by positive feedback loops



This is how a cell remembers of which cell type it is

Figure 7-68 *Molecular Biology of the Cell* (© Garland Science 2008)

Cells develop from stem cells in several steps of differentiation

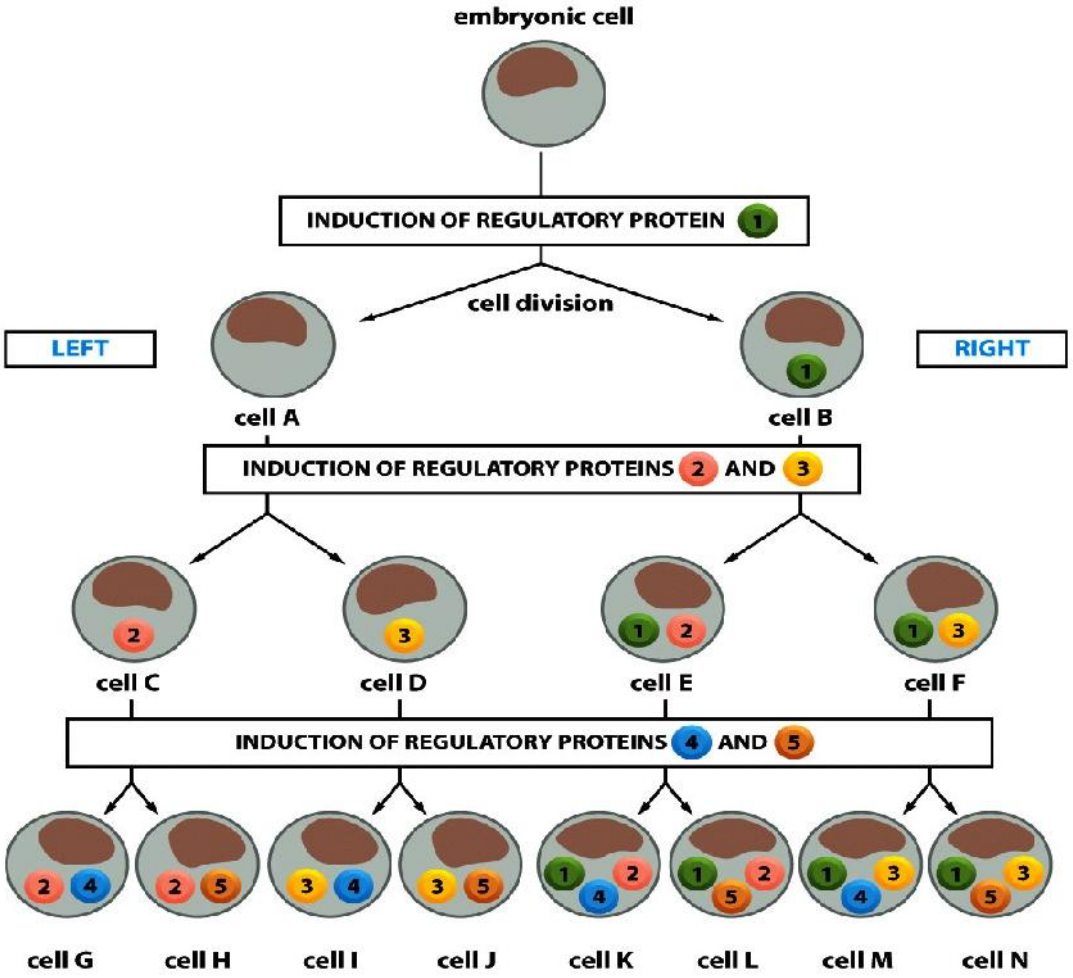
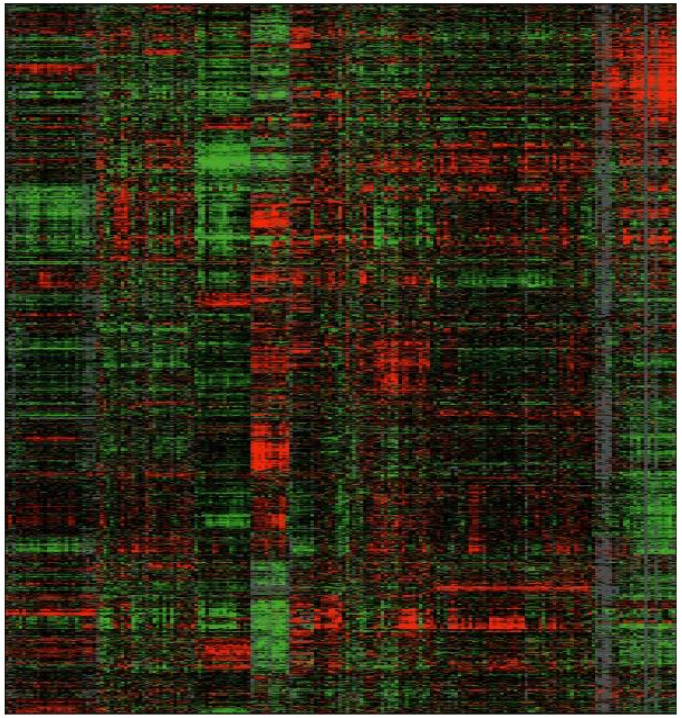


Figure 7-76 *Molecular Biology of the Cell* (© Garland Science 2008)

Methylation is additional memory (RAM) for cells to remember who they are

```
B-cell          r x w
Liver      r x -
Brain      - - -
Muscle          - - -
Kidney          r x -
...
```

A gene expression profile is a mirror of the RAM of a cell



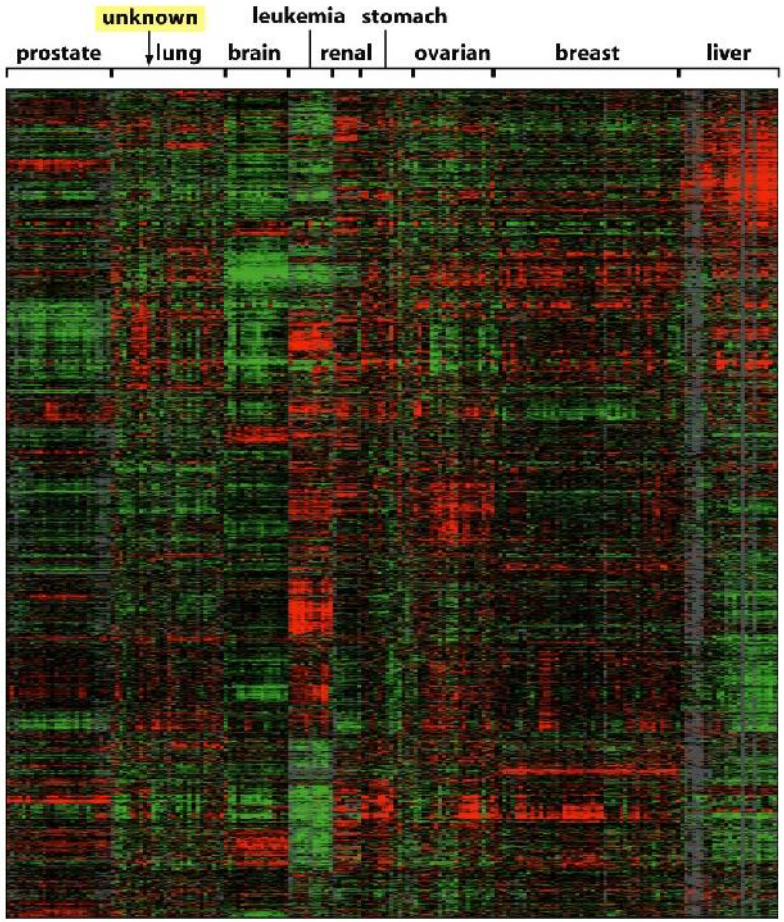
Column: cell type (core dump)
Row: gene

Color: mRNA abundance (expression)
red = high
green = low

```
admin:utils core list

      Size          Date              Core File Name
==============================================================
 232032 KB    2009-05-09 18:20:00    core.6047.6.ccm.1241907597
 528508 KB    2009-10-29 10:17:18    core.14040.11.ccm.1256303364
 238276 KB    2009-11-30 11:11:50    core.21377.11.ccm.1259597500
admin:
Control-C pressed
```


Figure 7-3 *Molecular Biology of the Cell* (© Garland Science 2008)

The information whether a cell is a liver cell or a neuron is reflected in RNA and protein abundances

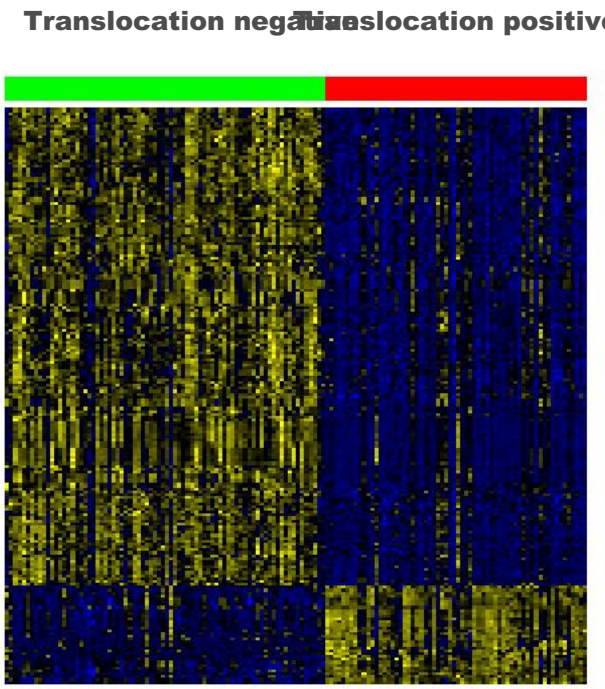


Every cell type has a transcriptional identity

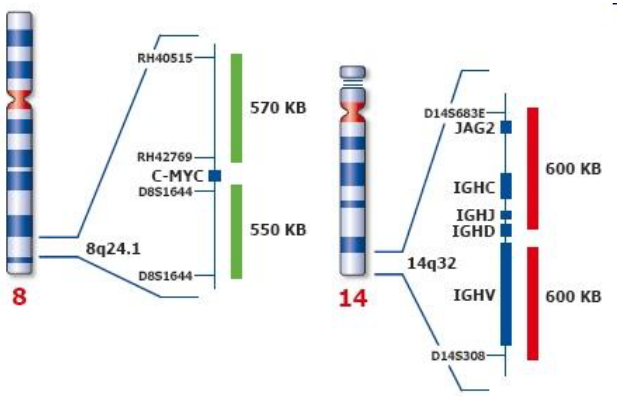
Tell me which programs you use and I tell you who you are

Figure 7-3 *Molecular Biology of the Cell* (© Garland Science 2008)

Genetic aberrations can reprogram the transcriptional identity of a cell



The IG-Myc translocation in lymphomas



Changes on the genome effect only two genes. However changes in the code effect the expression of more then a thousand genes.

Consequences of one wrong line of code

Cells process input signals to output signals

Input: A growth factor binds to a membrane receptor

Output: The cell grows

Signaling networks driven by protein interactions do the computation

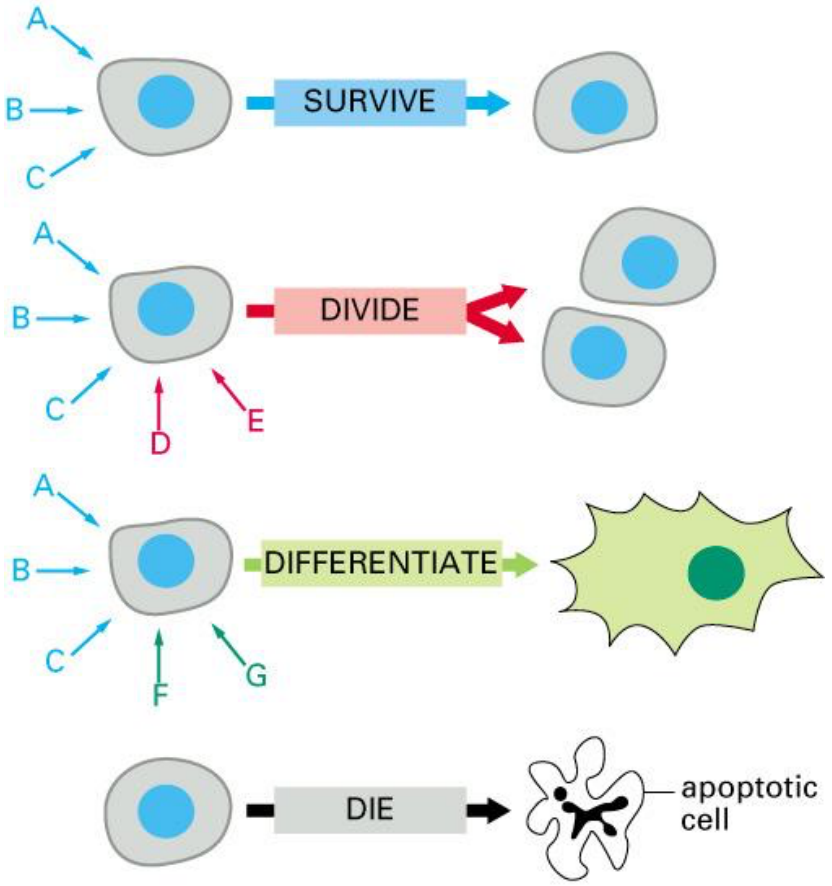


Figure 15–8. Molecular Biology of the Cell, 4th Edition.

A cell has both fast and slow modes to respond to a stimulus

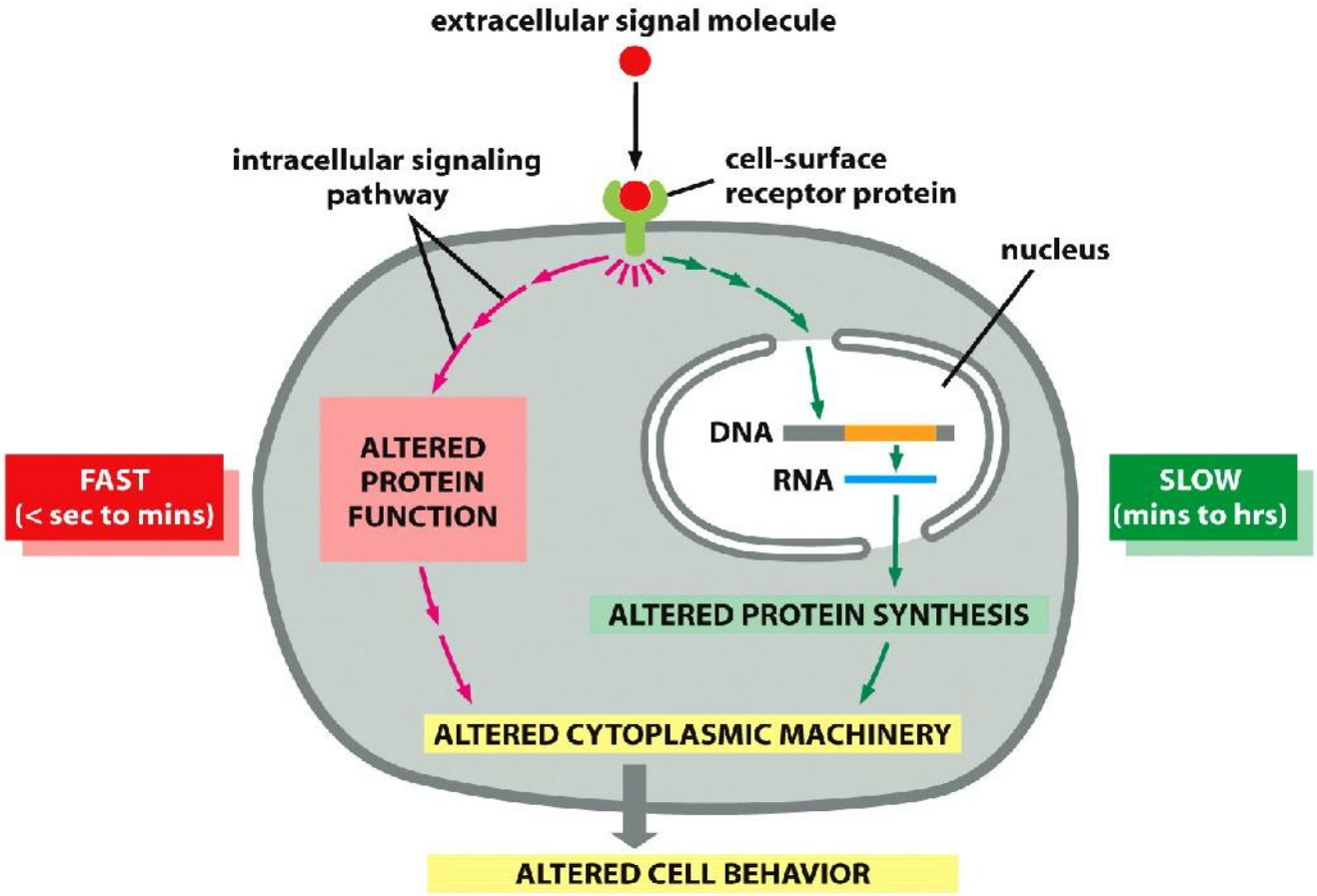
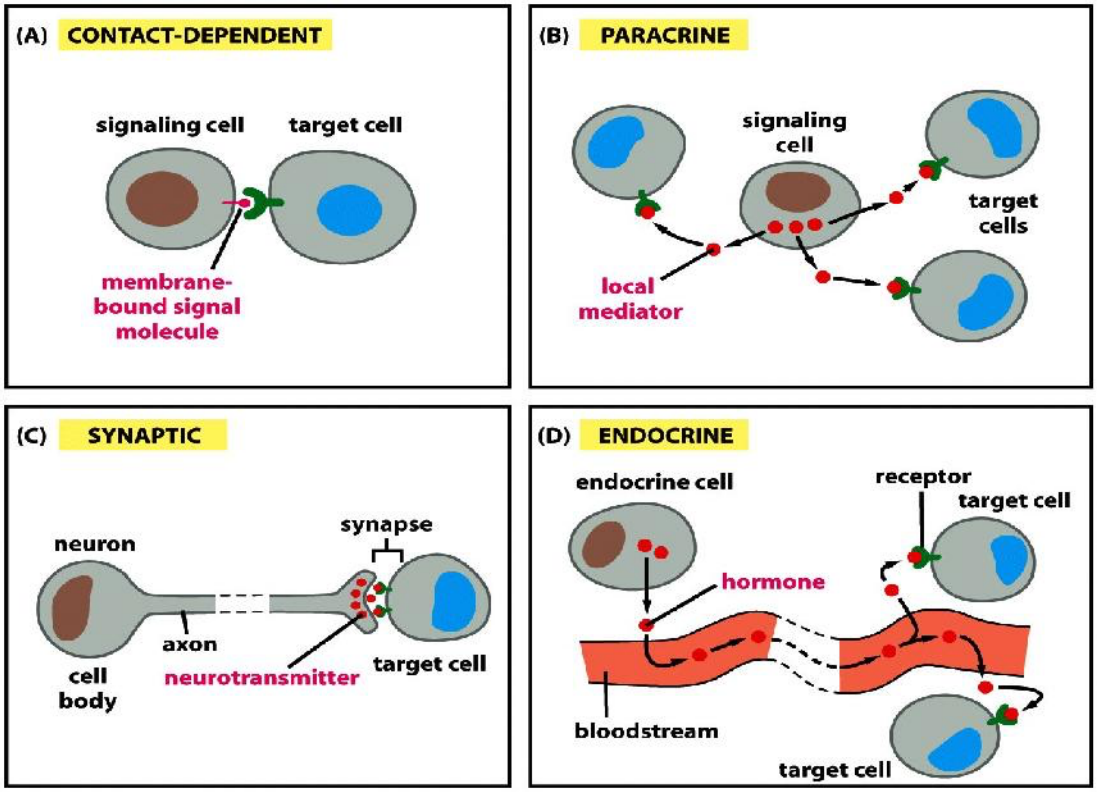


Figure 15-6 *Molecular Biology of the Cell* (© Garland Science 2008)

Cells talk to each other



Multicellular organisms are compute grids

Figure 15-4 *Molecular Biology of the Cell* (© Garland Science 2008)

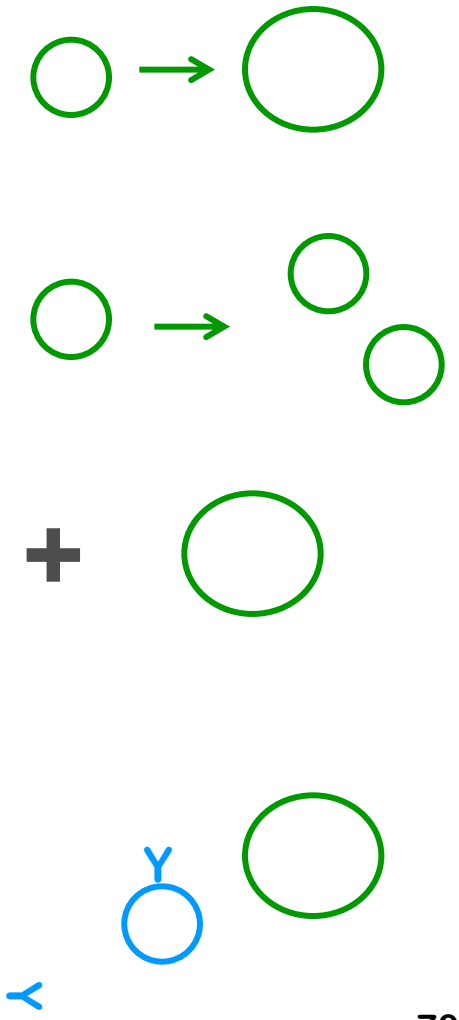
Tumors arise from dysfunctional cellular signal processing

Normal cells grow when receiving growth signals. Tumor cells grow without these signals

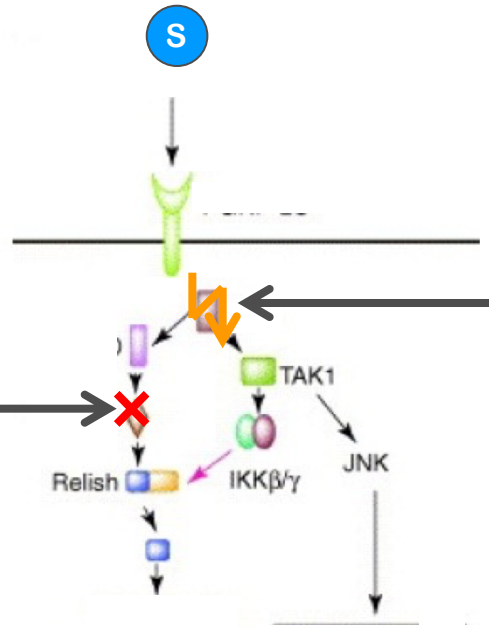
They proliferate although they should not

They receive signals of cell death but do not respond to them properly

They can escape immune responses by sending out signals that modulate the immune system



In cancer mutations perturb signaling



Mutations can block the signal flow

Mutations can introduce constitutive active signals

You can compare cells with computers

hard drive (DNA)

transient memory (RNA, protein, methylation, …)

file system (DNA binding sites)

executable (proteins)

compilers (ribosomes)

networks (tissues, bodies, ecosystems)

internet (life on earth)

debugging (medicine)

How are genomes programed ?

You have code to calculate the GC skew

```
computeGCskew <- function(x) {

  if (is.character(x)) {
    x <- alphabetFrequency(DNAString(x))
  }
  res <- (x["G"]-x["C"])/(x["G"]+x["C"])
  names(res) <- NULL
  return(res)
}
```

How would you generate code to calculate the AT skew?

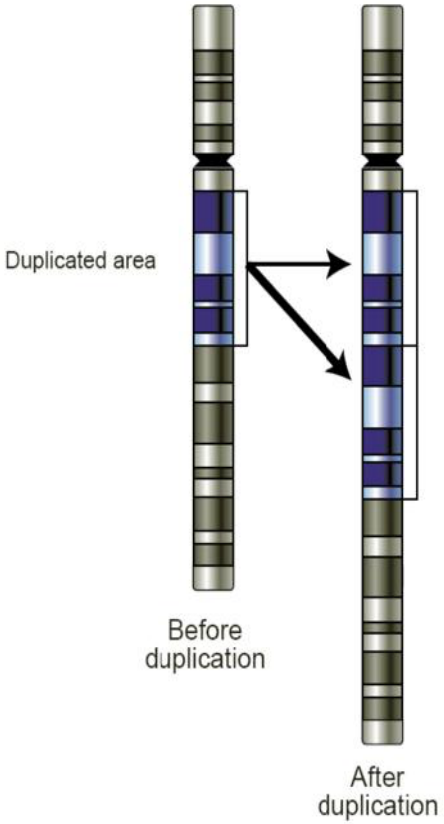
Copy the old code and make some changes

```
computeGCskew <- function(x) {

  if (is.character(x)) {
    x <- alphabetFrequency(DNAString(x))
  }
  res <- (x["G"]-x["C"])/(x["G"]+x["C"])
  names(res) <- NULL
  return(res)
}
```

```
computeATskew <- function(x) {

  if (is.character(x)) {
    x <- alphabetFrequency(DNAString(x))
  }
  res <- (x["A"]-x["T"])/(x["A"]+x["T"])
  names(res) <- NULL
  return(res)
}
```

Thats is how it is done in genomes too

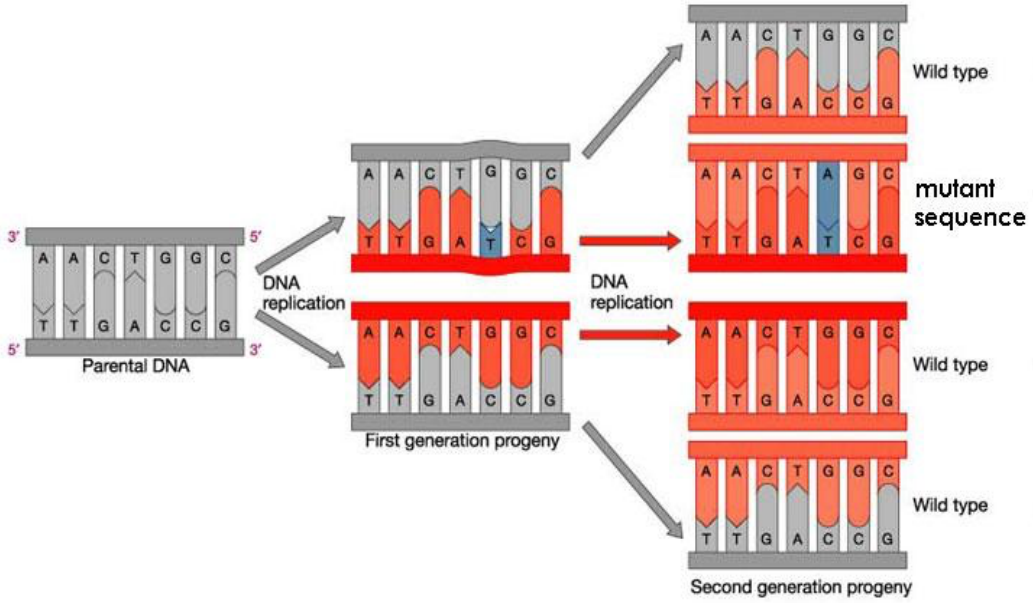
Gene duplications are rare "errors" during DNA replication



Now the cell has the same code twice.

How can it make the necessary alterations?

Mutations randomly alter existing code

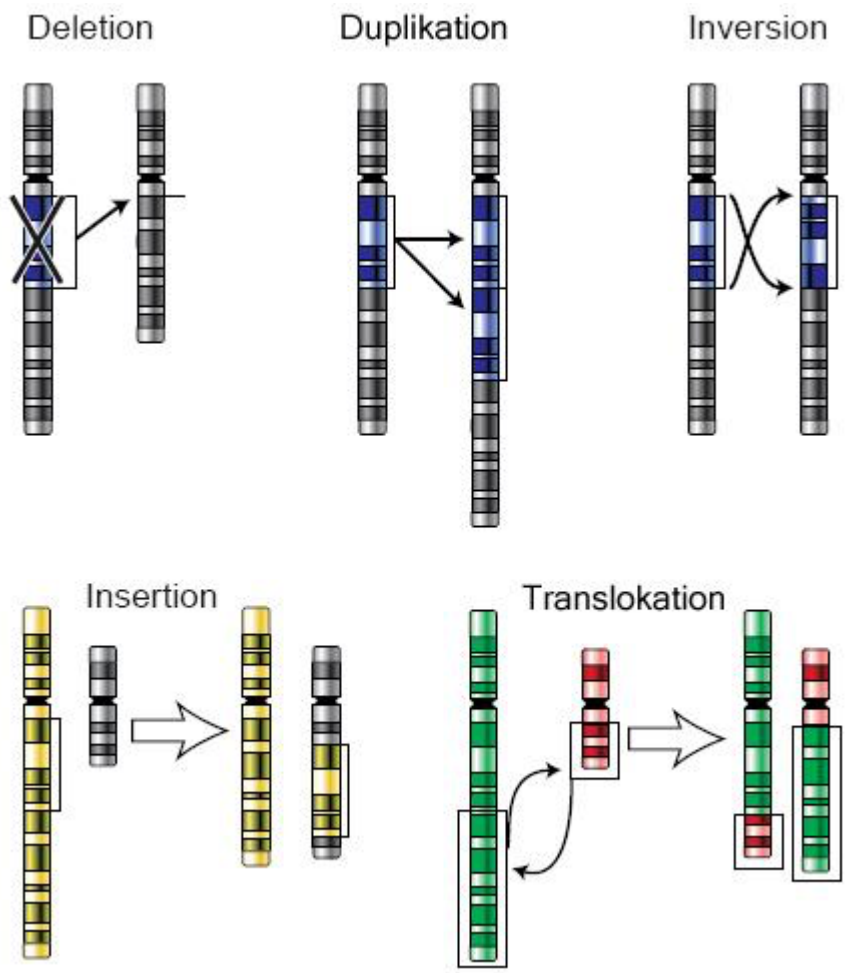


Mutations are "errors" during DNA replication

They can be beneficial

Point mutations affect single bases

Mutations can affect long sequences



If mutations are random modifications of genomes, why don't we have random genomes?

Only reproducing sequences survive

```
computeGCskew <- function(x) {

  ix (is.character(x)) {
    x <- alphabetFrequency(DNAString(x=)
  }
  res <- (x["G"]-x["C"])/(x["G"]+x["C"])
  names(res) <- NULL
  return(res)
}
```

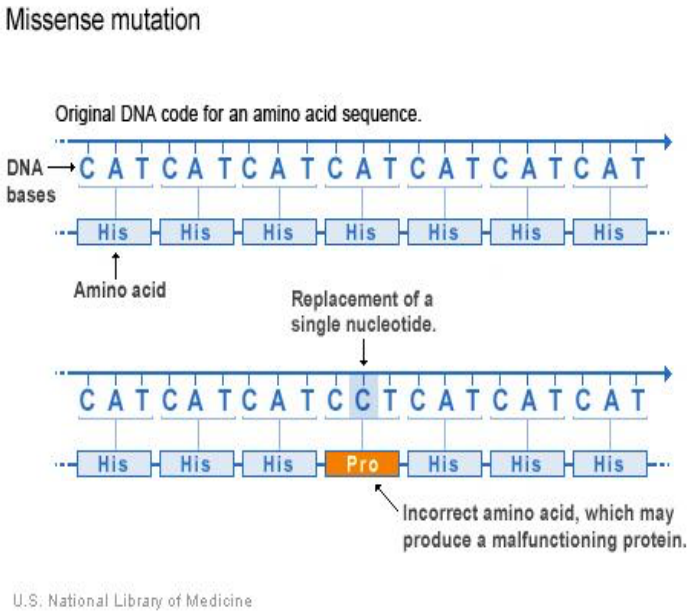
If the program does not run any more, the organism might have problems

The individual might die young without children

→ The genome disappears from the gene pool

Natural selection is the parser of genomes

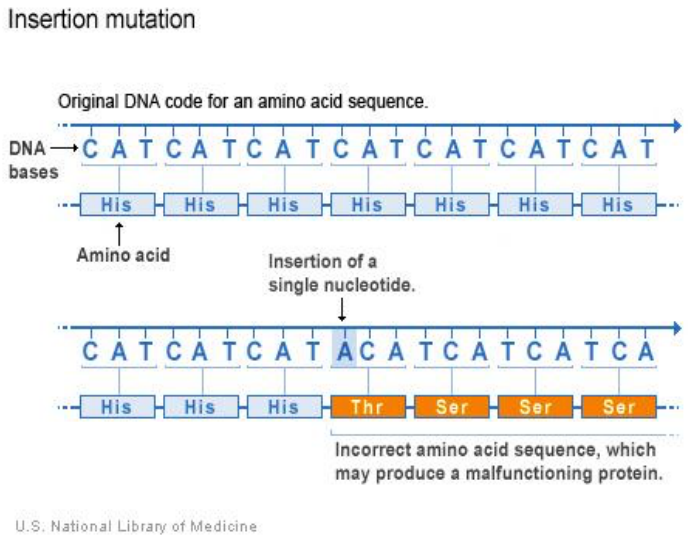
Mutations in DNA can cause the inclusion of a different amino acid into a protein altering its function




```
computeGCskew <- function(x) {

  if (is.character(x)) {
    x <- alphabetFrequency(DNAString(x))
  }
  res <- (y["G"]-x["C"])/(x["G"]+x["C"])
  names(res) <- NULL
  return(res)
}
```

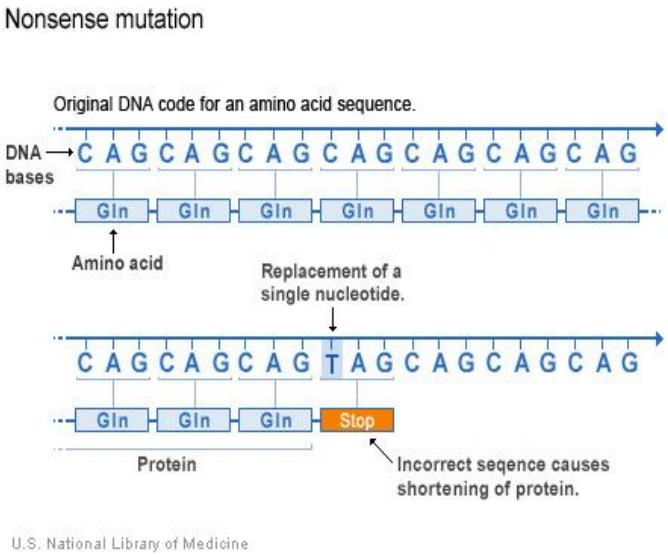
This might be a problem or the invention of a new cool program
Natural selection will decide what it is

Single DNA Mutations can completely reprogram a protein




```
computeGCskew <- function(x) {

  if (is.character(x)) {
    y <- length(x)-7
  return(y)}
}
```

Mutations in DNA can introduce early stop codons




```
computeGCskew <- function(x) {

  if (is.character(x)) }
```

This might be a problem or the invention of a new cool program
Natural selection will decide what it is

Mutations can insert code

ATTCTGGCT
ATTCTTTGGCT

```
computeGCskew <- function(x) {

  if (is.character(x)) {
    x <- alphabetFrequency(DNAString(x))
  }
  res <- (x["G"]-x["C"])/(x["G"]+x["C"])
  res <- res/2
  names(res) <- NULL
  return(res)
}
```

This might be a problem or the invention of a new cool program
Natural selection will decide what it is

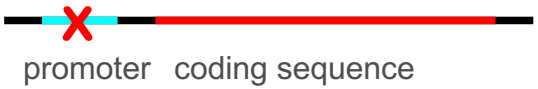
Mutations can delete code

ATTCTTTGGCT
ATTCTGGCT

```
computeGCskew <- function(x) {

  if (is.character(x)) {
    x <- alphabetFrequency(DNAString(x))
  }
  res <- (x["G"]+x["C"])
  names(res) <- NULL
  return(res)
}
```

This might be a problem or the invention of a new cool program
Natural selection will decide what it is

Mutations in transcription factor binding sites might prevent the cell from finding the gene




```
computeGCspew <- function(x) {

  if (is.character(x)) {
    x <- alphabetFrequency(DNAString(x))
  }
  res <- (x["G"]-x["C"])/(x["G"]+x["C"])
  names(res) <- NULL
  return(res)
}
```

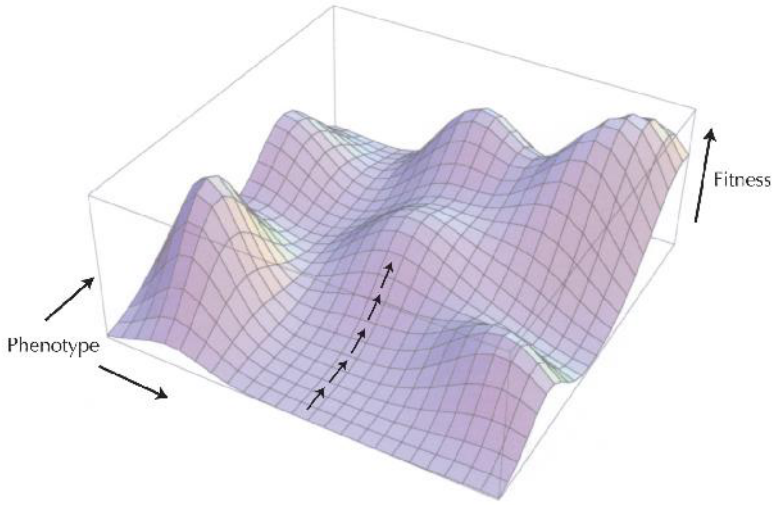
This might be a problem or the invention of a new cool program
Natural selection will decide what it is

If a gene is not used anymore selective pressure is missing and mutations accumulate

```
xompteGjCspew <- puncton:x) {

  if (isi.cactor(=)=== {
    x <- ahabetelprequincy(tring())
  ?
  res <- (xp"G"]kx[""])/x[(G"]lx[])
  prames(pis) <- BULL
  reburn(krass!
}
```

pseudo genes = dead code

Evolution is a highly parallelized optimization algorithm



Every single cell evolves on its own

Mutations are small local changes in the genome

Evolution can get stuck in local optima.

We are not optimal

The programming of genomes also known as evolution is a process of trial and error

Variation: **random, uniform, unstructured**

Selection: **introduces structure**
but does not get rid of all randomness

Genomic code is a mess

“Cells evolved to survive and not for scientists to understand”

End of Chapter 2