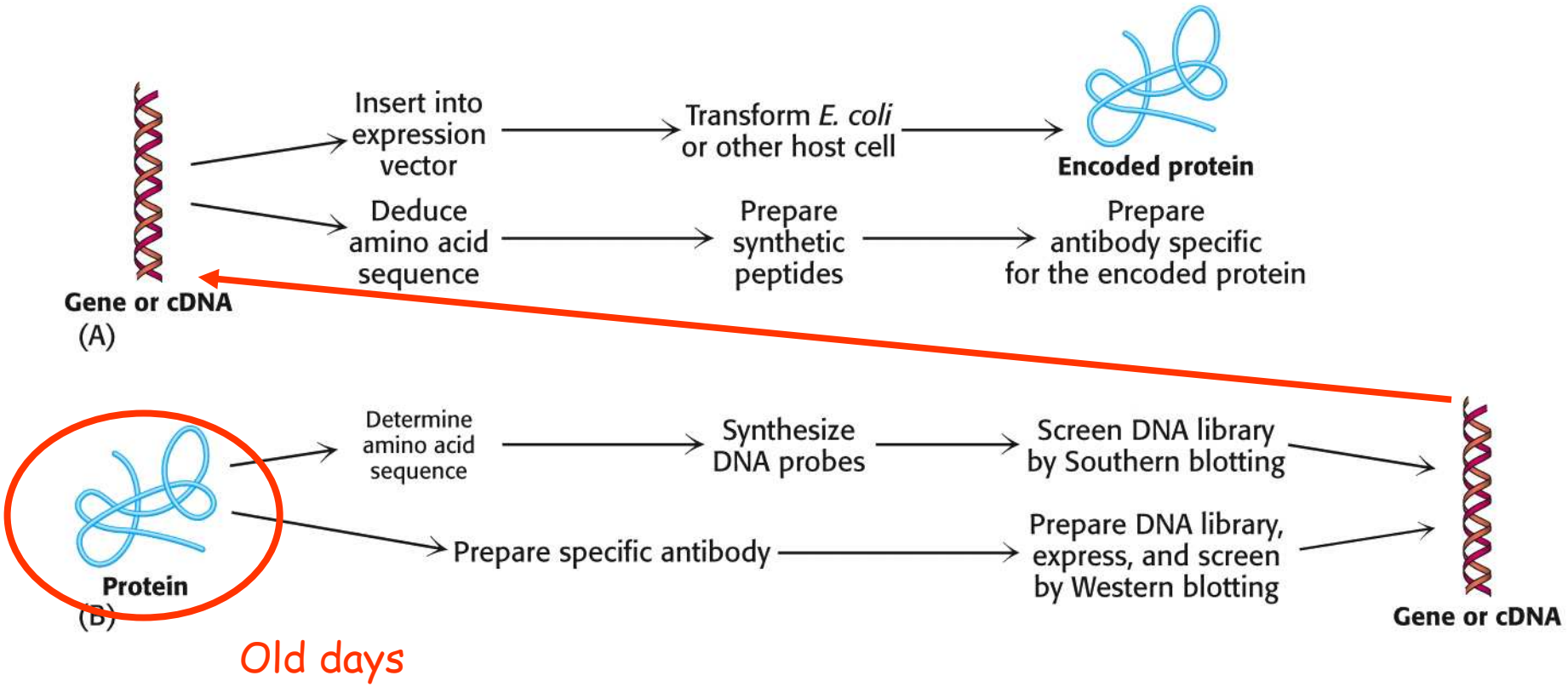
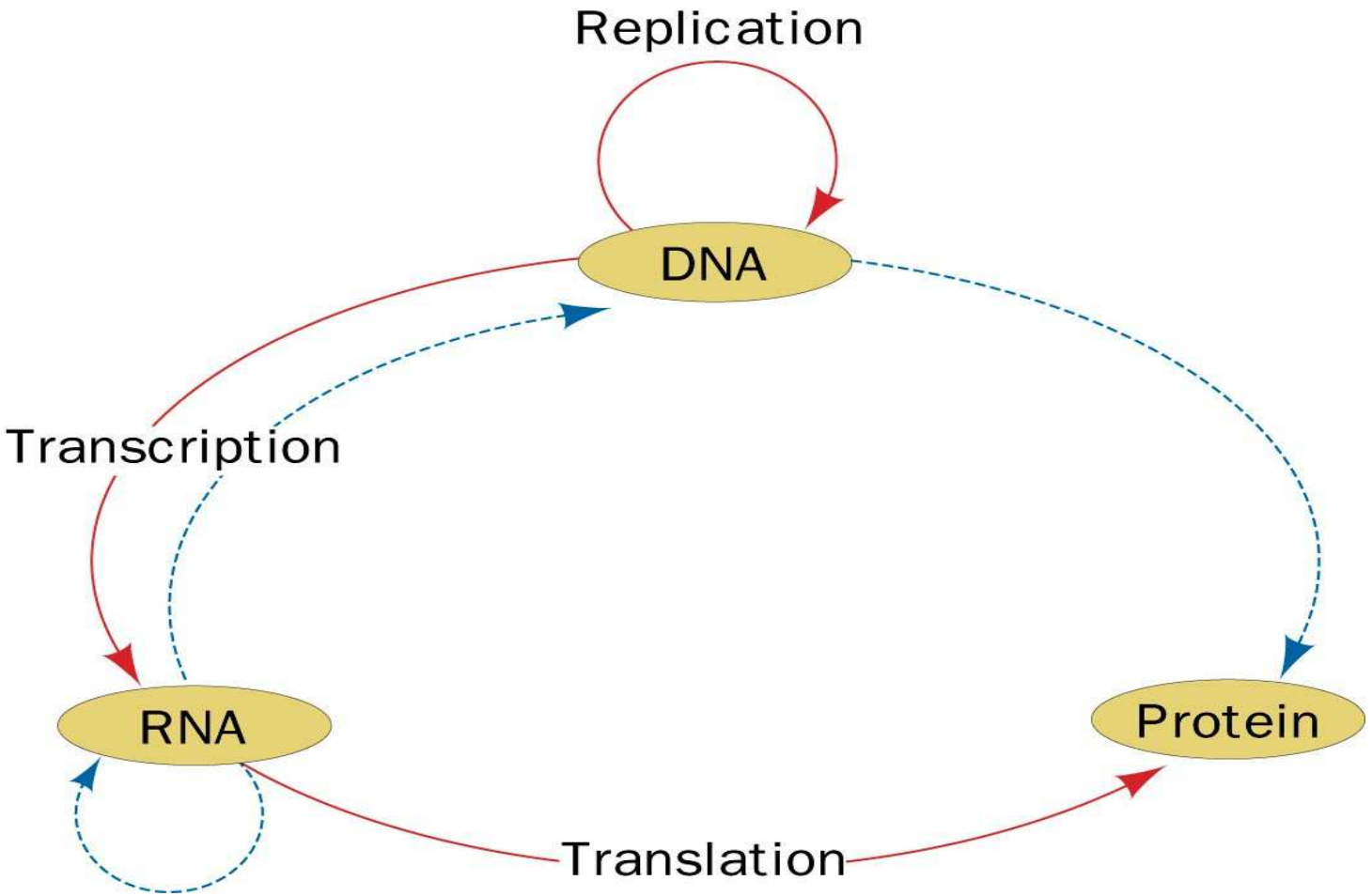
Heterologous gene expression (in prokaryotes)

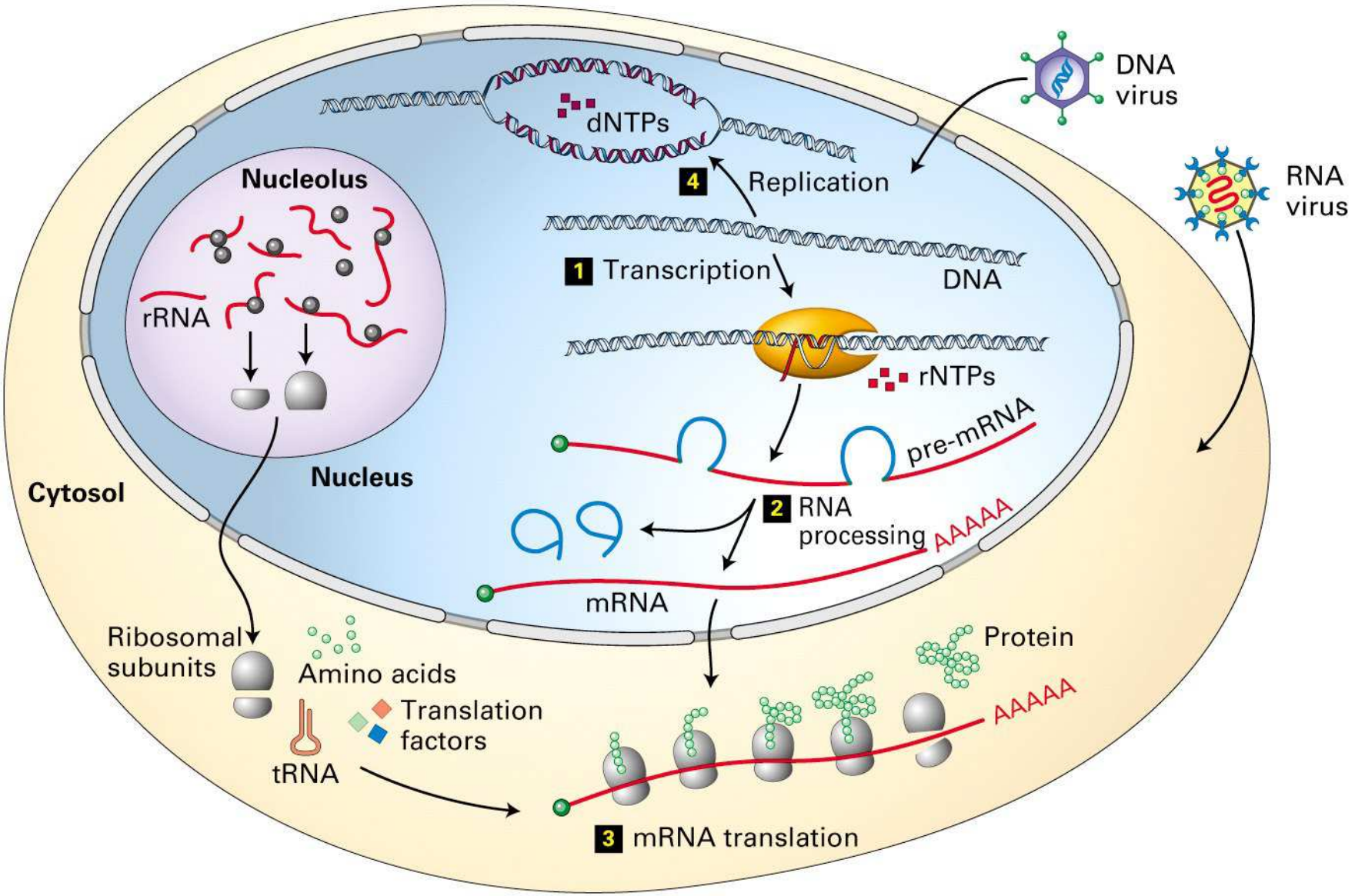
Different ways to clone a gene



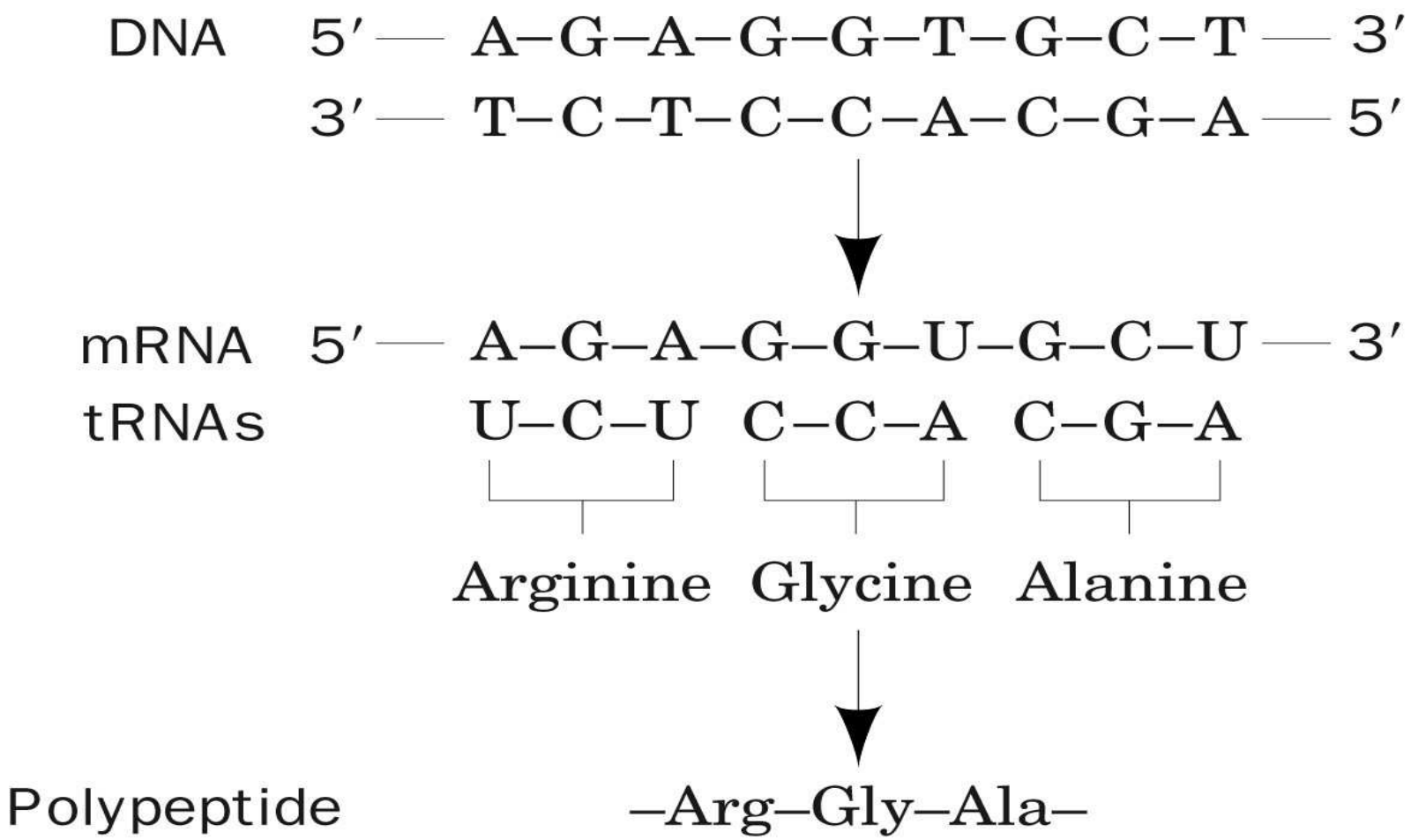
Old days

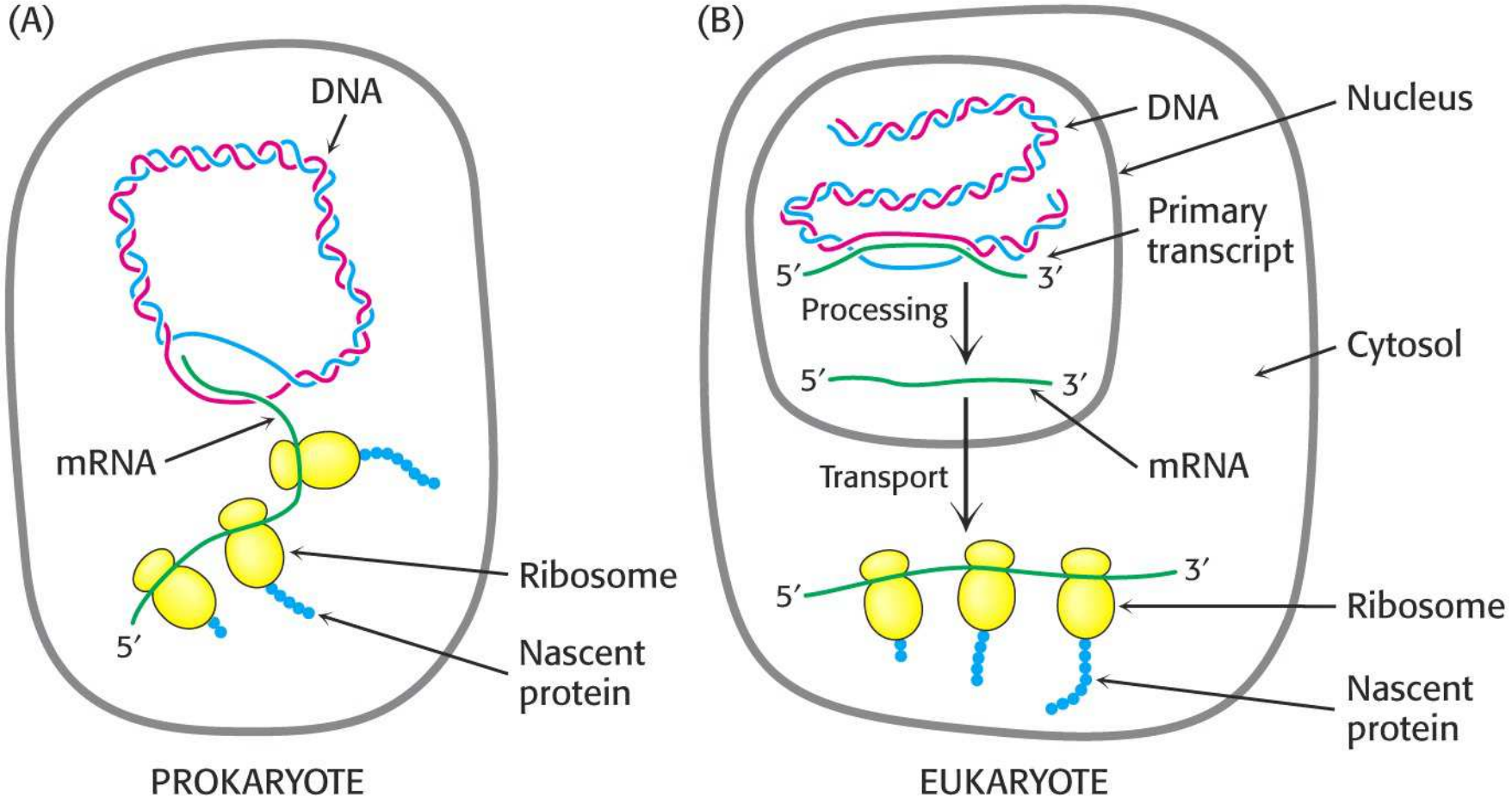
The central dogma



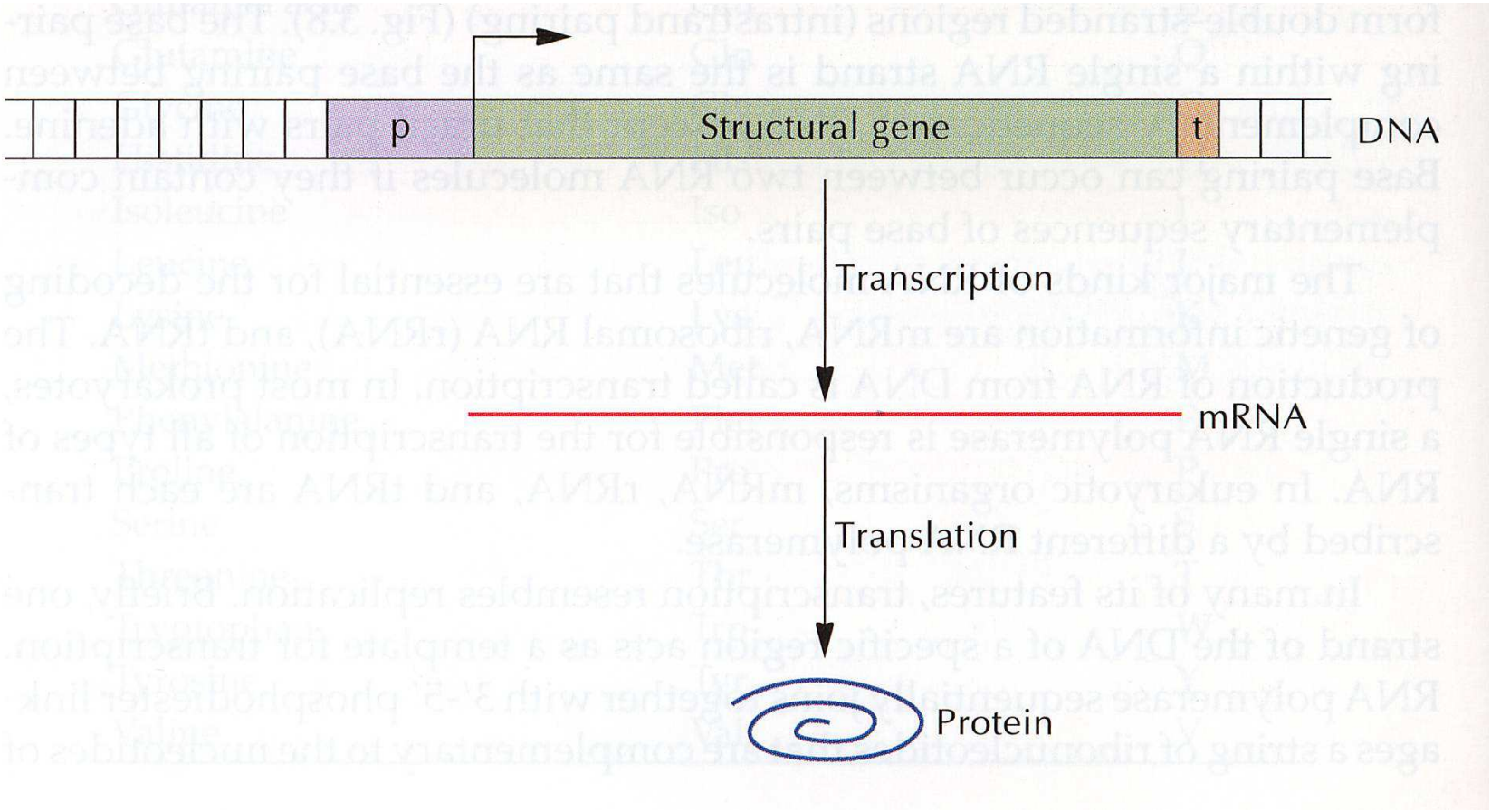

DNA virus
RNA virus
dNTPs
4 Replication
Nucleolus
rRNA
1 Transcription
DNA
rNTPs
pre-mRNA
Nucleus
Cytosol
2 RNA processing
AAAAA
mRNA
Ribosomal subunits
Amino acids
Translation factors
tRNA
Protein
AAAAA
3 mRNA translation

Gene expression.

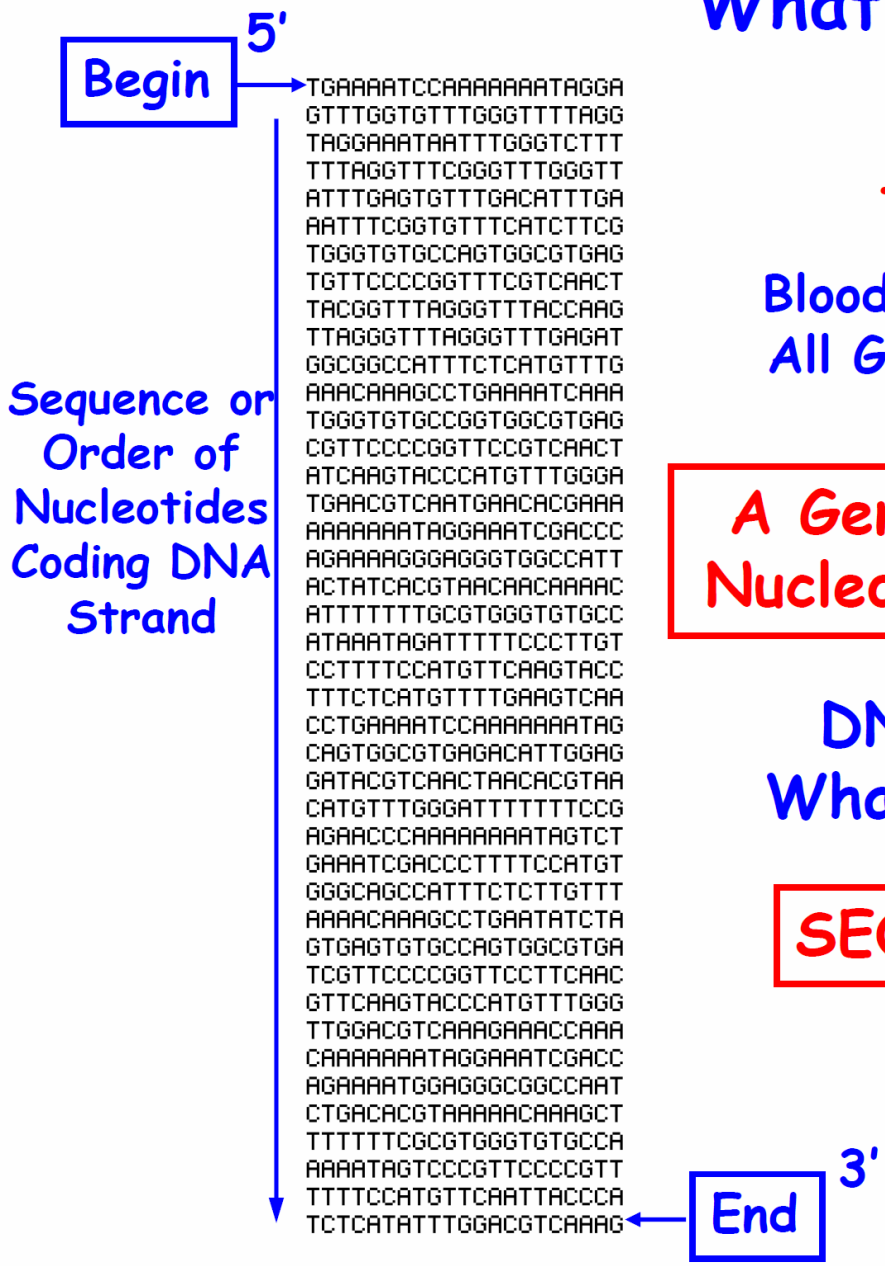



(A)
DNA
mRNA
5′
Ribosome
Nascent protein
PROKARYOTE
(B)
DNA
Nucleus
Primary transcript
5′
3′
Processing
5′
3′
Cytosol
mRNA
Transport
5′
3′
Ribosome
Nascent protein
EUKARYOTE

What is a gene?



What is A Gene?



Sequence or Order of Nucleotides Coding DNA Strand

The β-globin Gene

Blood Protein Carries Oxygen to All Genes From Lungs ⇨ Energy

A Gene is a <u>Unique Sequence</u> of Nucleotides Specifying a Function

DNA Sequence = Biology!
What If Sequence Changed?

SEQUENCE → FUNCTION

Relative to Coding or Sense Strand of Gene

The Circular *E. Coli* Chromosome
One DNA Circle

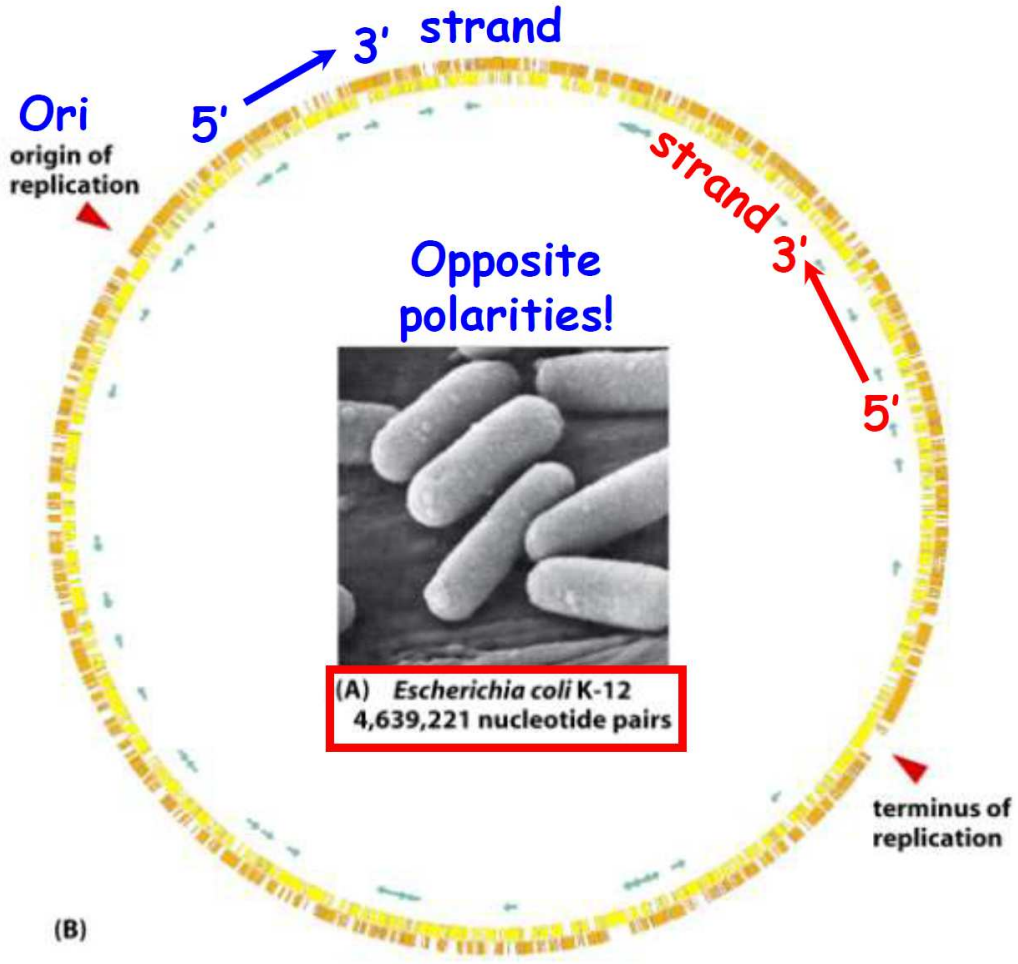
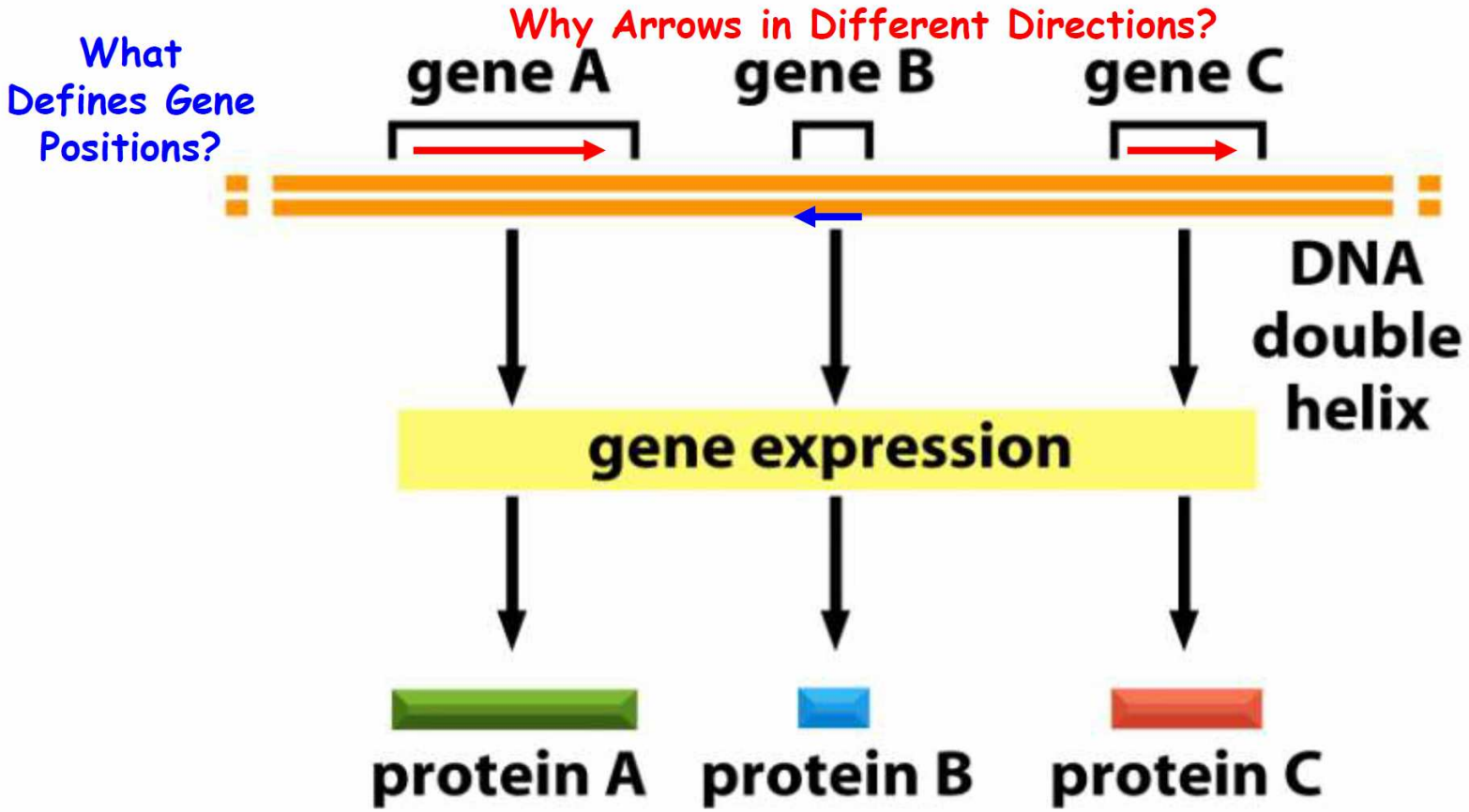


Figure 1-29 *Molecular Biology of the Cell*, Fifth Edition (© Garland Science 2008)

A Chromosome Contains Many Genes That Reside at Specific Positions and Have Unique Functions



Because DNA Contains Two Strands--Genes Can Be Transcribed From Either Strand--But Only One Per Gene

Bacterial Genome Projects Have Provided Remarkable Insight Into Bacterial Genomes and Cell Functions

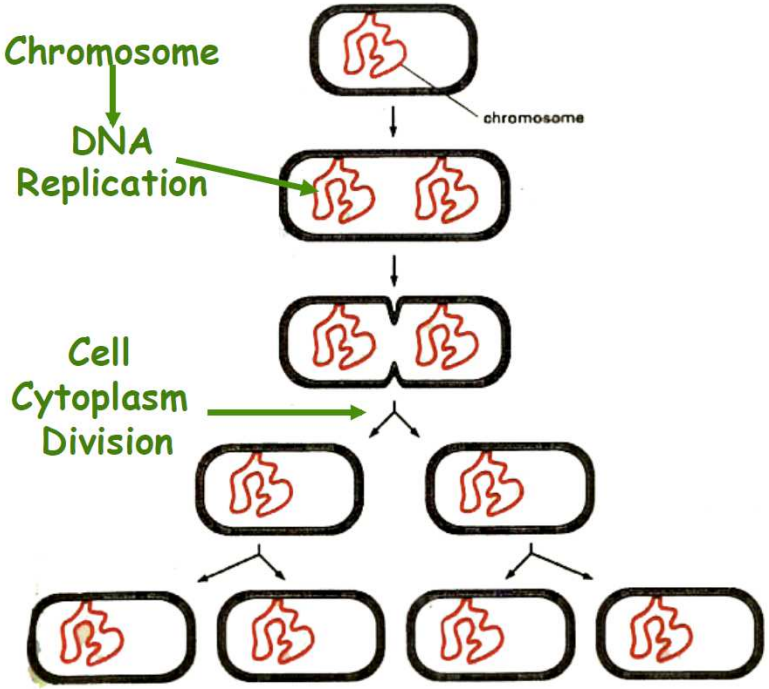
Table 1–1 Some Genomes That Have Been Completely Sequenced

SPECIES	SPECIAL FEATURES	HABITAT	GENOME SIZE (1000s OF NUCLEOTIDE PAIRS PER HAPLOID GENOME)	ESTIMATED NUMBER OF GENES CODING FOR PROTEINS
BACTERIA				
Mycoplasma genitalium	has one of the smallest of all known cell genomes	human genital tract	580	468
Synechocystis sp.	photosynthetic, oxygen-generating (cyanobacterium)	lakes and streams	3573	3168
Escherichia coli	laboratory favorite	human gut	4639	4289
Helicobacter pylori	causes stomach ulcers and predisposes to stomach cancer	human stomach	1667	1590
Bacillus anthracis	causes anthrax	soil	5227	5634
Aquifex aeolicus	lithotrophic; lives at high temperatures	hydrothermal vents	1551	1544
Streptomyces coelicolor	source of antibiotics; giant genome	soil	8667	7825
Treponema pallidum	spirochete; causes syphilis	human tissues	1138	1041
Rickettsia prowazekii	bacterium most closely related to mitochondria; causes typhus	lice and humans (intracellular parasite)	1111	834
Thermotoga maritima	organotrophic; lives at very high temperatures	hydrothermal vents	1860	1877

1400 Bacterial Genomes Have Been Sequenced to Date (January, 2011)

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

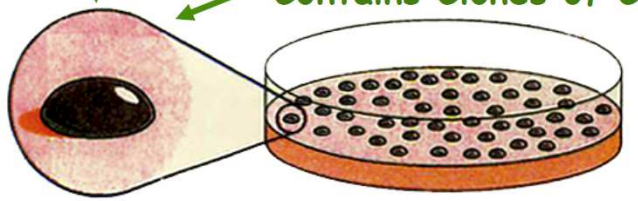
Genes Are Replicated During Each Cell Division



Clones

A "Clone"

Note - Each Clone of Bacteria Contains Clones of Cells



-10^6 cells per colony

~50 colonies per dish

A Bacterial Colony Contains Many Copies of Same Cell, or Clones, Which are Genetically Identical!

Each Daughter Cell Contains The Same Collection of Genes

Clones!

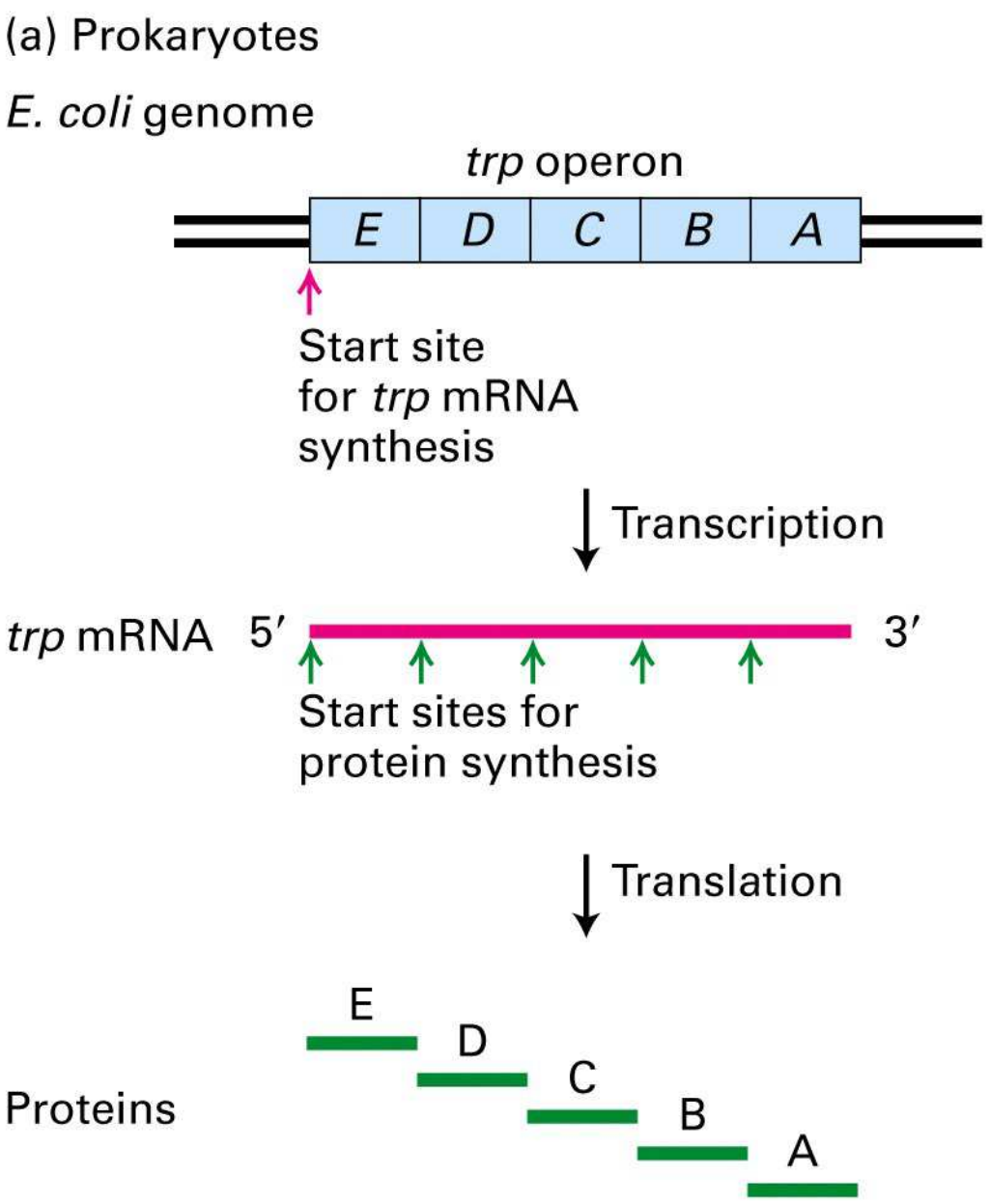
Major Properties of Genetic Material

Replication & Stability

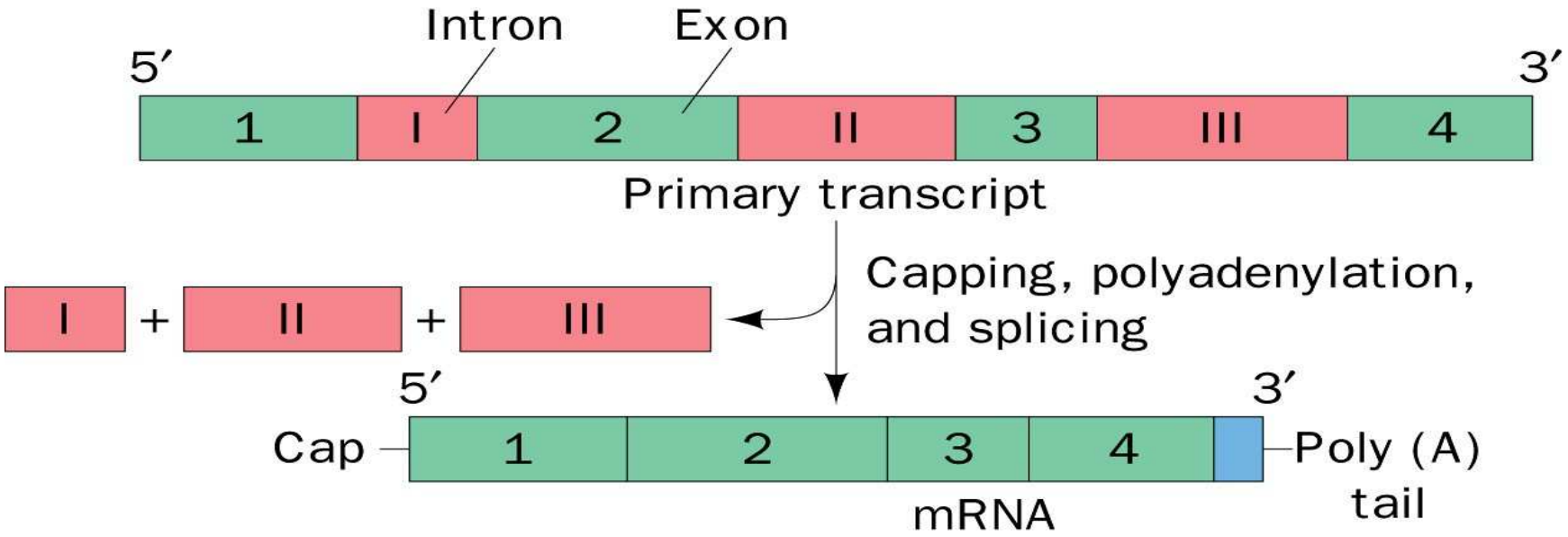
(a) Monocistronic

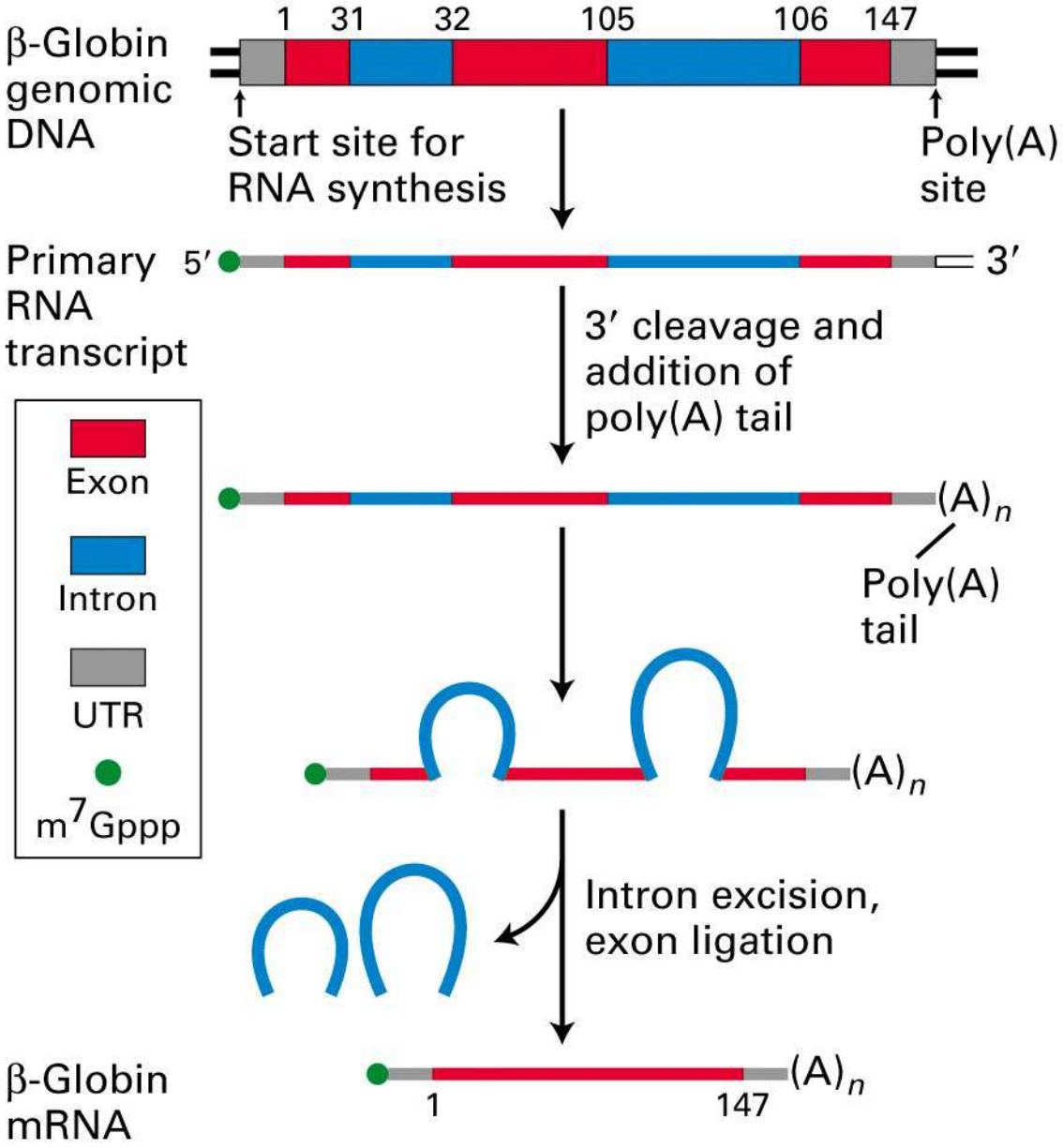
(b) Polycistronic

Prokaryotes operons


(a) Prokaryotes
E. coli genome
trp operon
E
D
C
B
A
Start site
for trp mRNA
synthesis
Transcription
trp mRNA
5′
3′
Start sites for
protein synthesis
Translation
Proteins
E
D
C
B
A

Eukaryotes – Intron-Exon concept




β-Globin genomic DNA
1
31
32
105
106
147
Start site for RNA synthesis
Poly(A) site
Primary RNA transcript
5′
3′
3′ cleavage and addition of poly(A) tail
Exon
Intron
UTR
m7Gppp
(A)n
Poly(A) tail
(A)n
Intron excision, exon ligation
β-Globin mRNA
(A)n
1
147

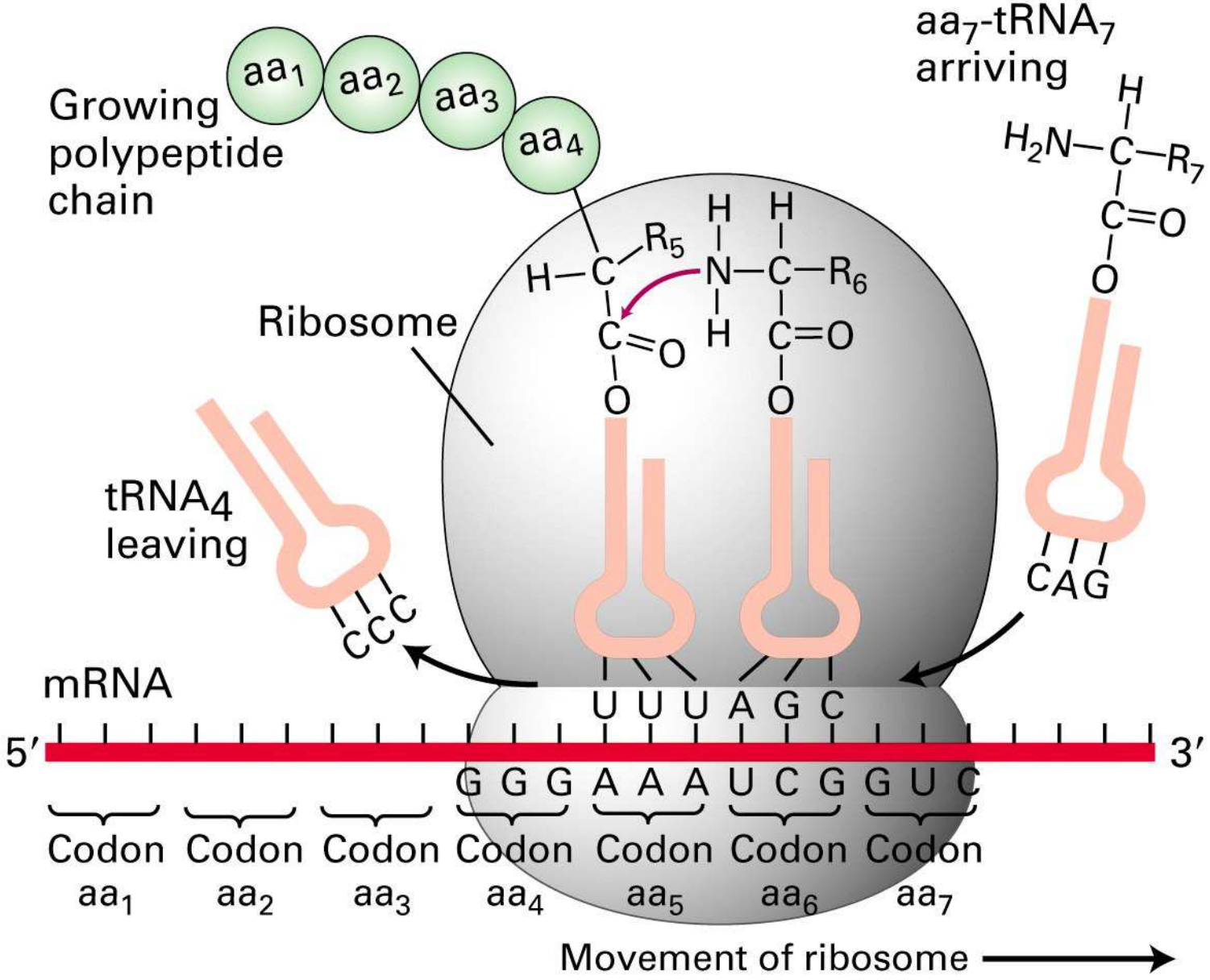

Growing polypeptide chain
aa1
aa2
aa3
aa4
aa7-tRNA7 arriving
Ribosome
tRNA4 leaving
CCC
CAG
mRNA
5′
3′
U U U A G C
G G G A A A U C G G U C
Codon aa1
Codon aa2
Codon aa3
Codon aa4
Codon aa5
Codon aa6
Codon aa7
Movement of ribosome

TABLE 5.4 The genetic code

First position (5′ end)	Second position				Third position (3′ end)
	U	C	A	G	
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	Stop	Stop	A
	Leu	Ser	Stop	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Note: This table identifies the amino acid encoded by each triplet. For example, the codon 5′ AUG 3′ on mRNA specifies methionine, whereas CAU specifies histidine. UAA, UAG, and UGA are termination signals. AUG is part of the initiation signal, in addition to coding for internal methionine residues.

Genetic code degenerate

General consideration about Gene Expression

- Expression Host -> expression System
- Promoter system -> expression vector
- Properties of product -> stability
- Production level

Measuring Gene Expression

Expression studies:

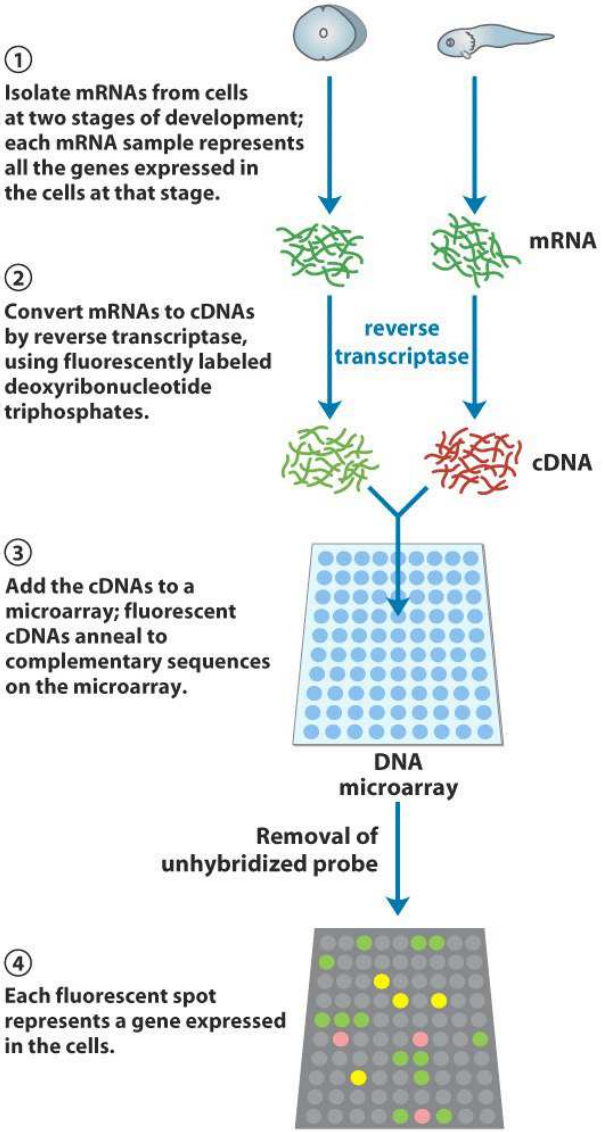
1. Analyzing Transcription
   - Northern blot
   - Micro array
   - real-time PCR
   - Primer extension

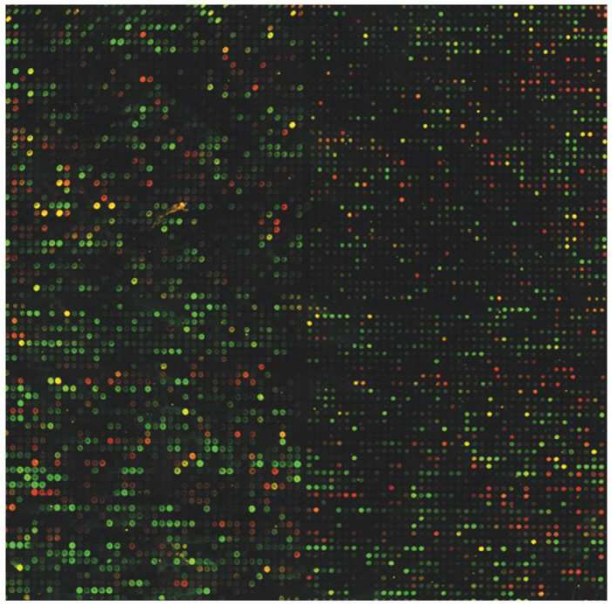
2. In vivo Expression studies
   Use of reporter genes to study regulatory elements

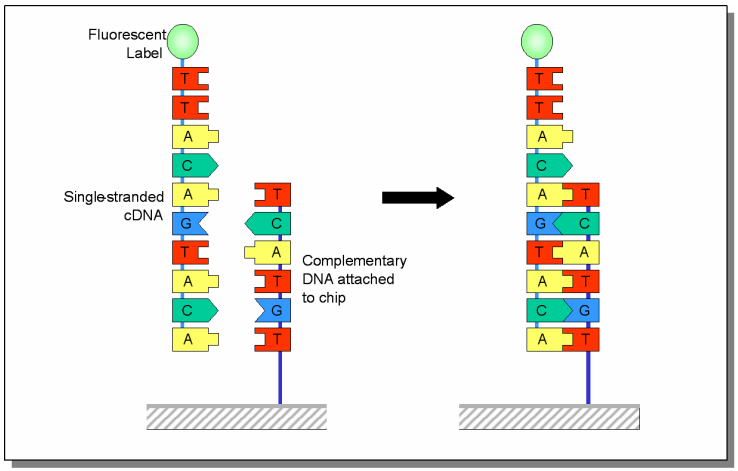
3. Analyzing Translation
   - Western blot - immuno assays
   - 2D electrophoresis
   - proteomics

Studying Transcription

Microarray technique – DNA chips





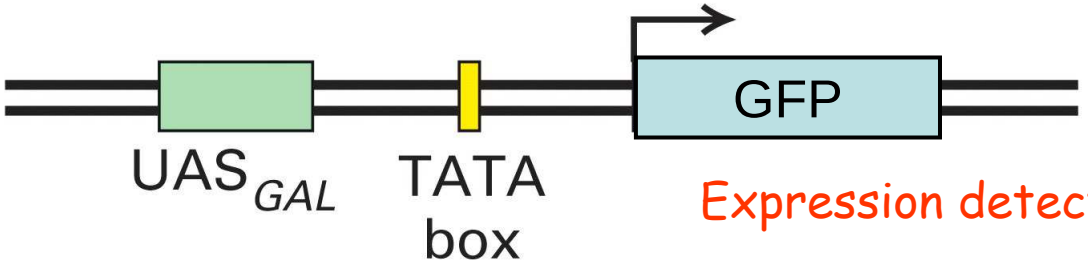


Promoter Studies

Used reporter genes:

- *Lac Z*
- *GFP*
- *Luciferase*

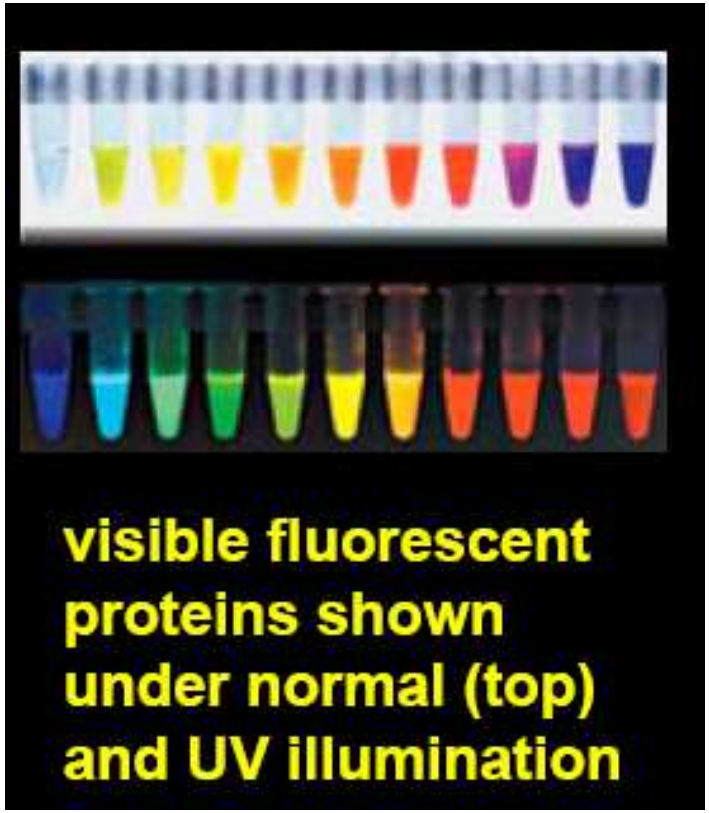
(a) Reporter-gene construct



Promoter

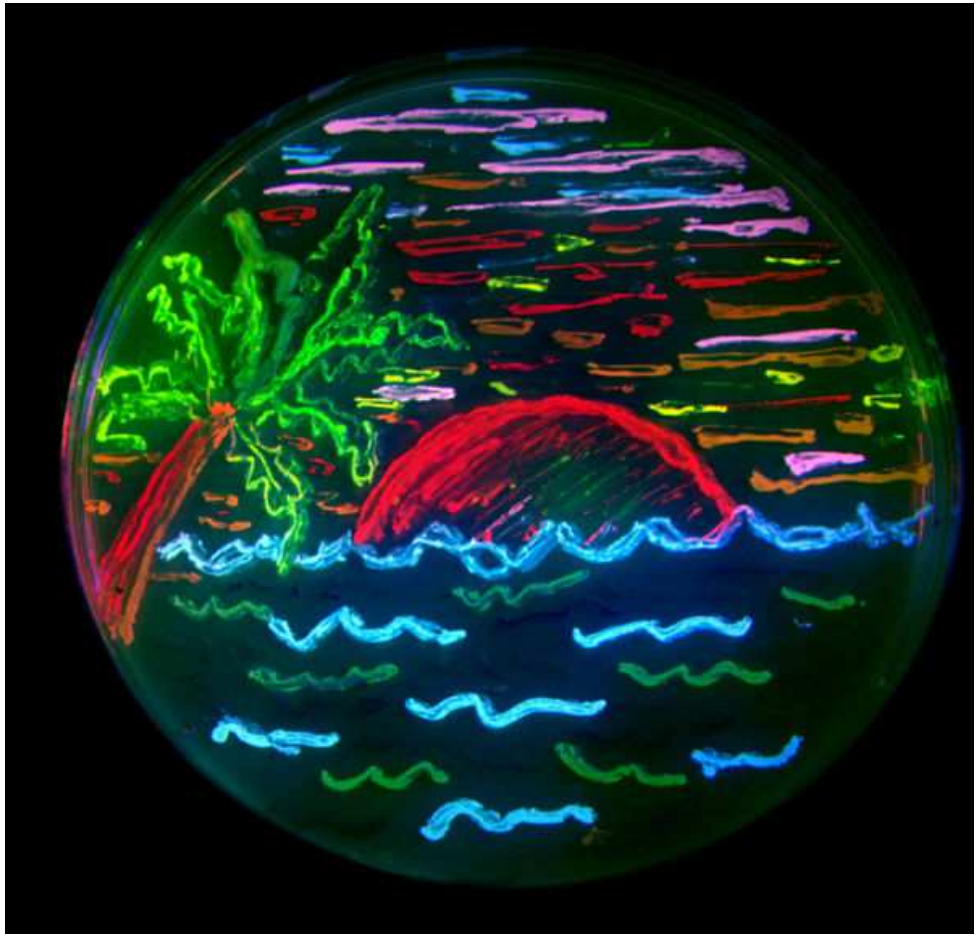
Expression detectable in vivo

GFP and its derivatives



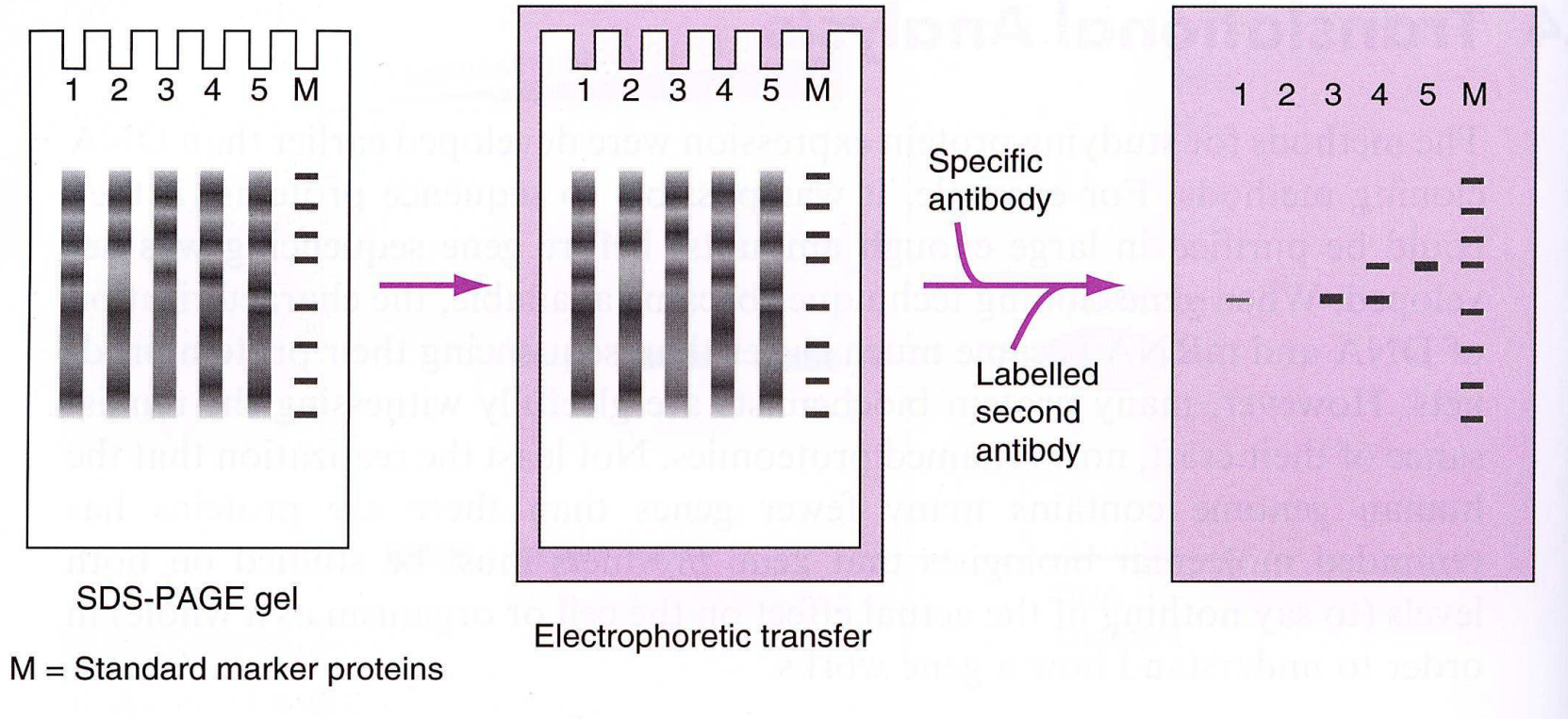
Chem. & Eng. News, Oct. 25, 2004

Engineered proteins are covering all the spectrum



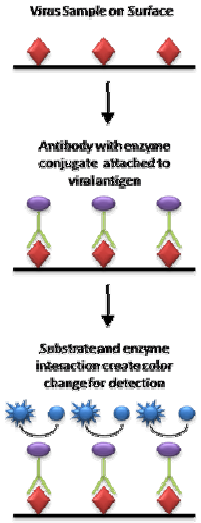
San Diego beach scene drawn with living bacteria expressing 8 different colors of fluorescent proteins.

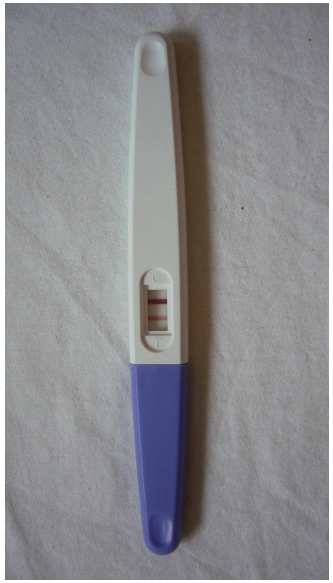
Analyzing Translation – Western Blot



ELISA

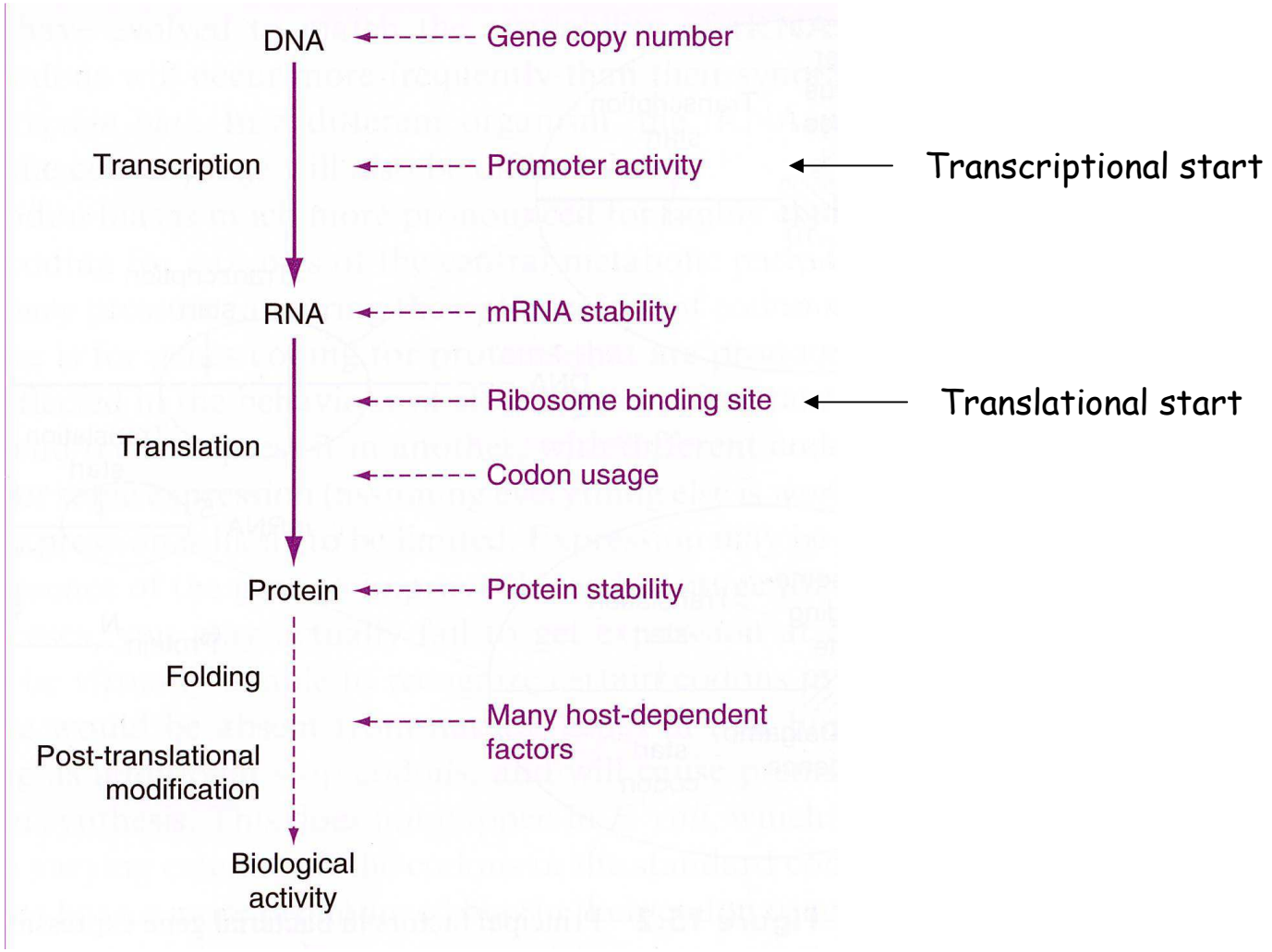
test that uses antibodies and color change to identify an antigen (protein)





Pregnancy testing

Gene Expression in prokaryotes



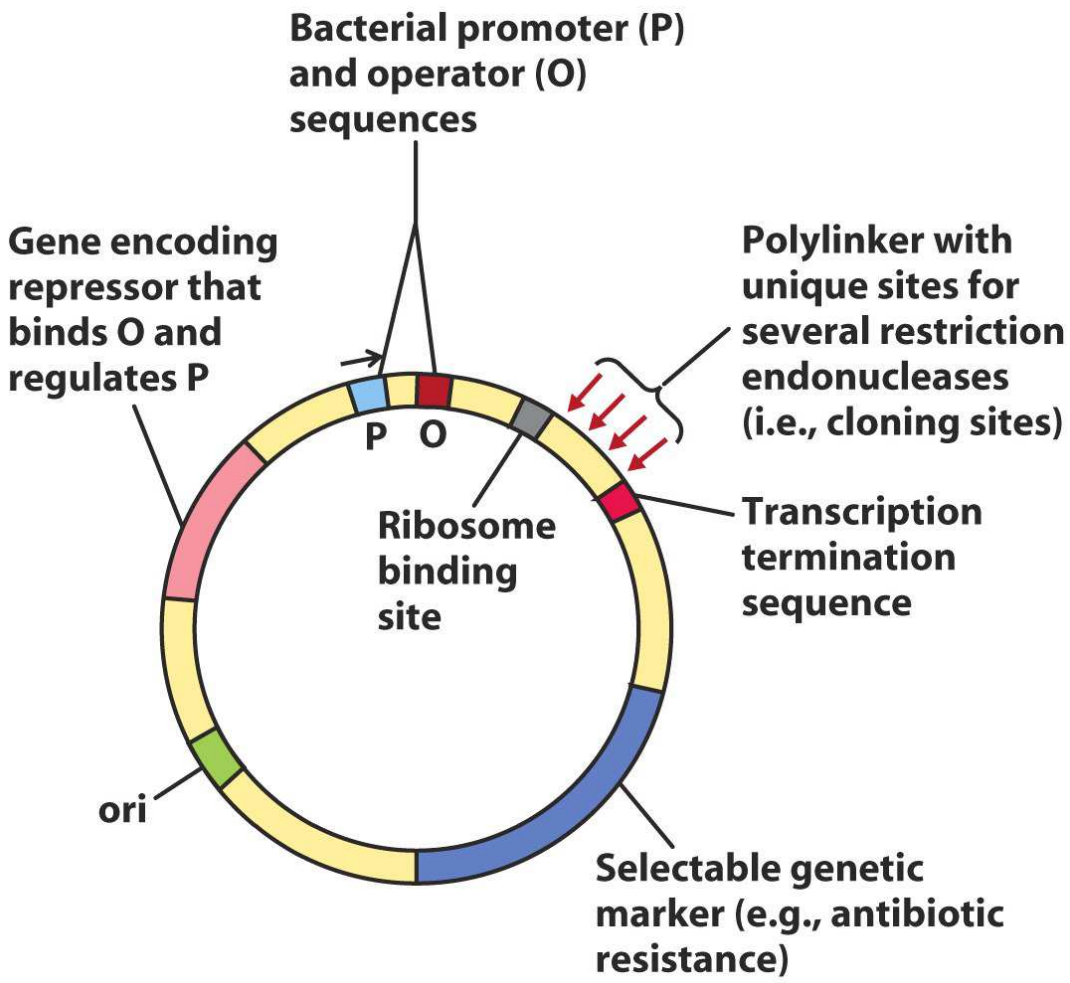
Gene copy number =
Plasmid copy number:

Table 1. Origins of Replication and Copy Numbers of Various Plasmids and Cosmids

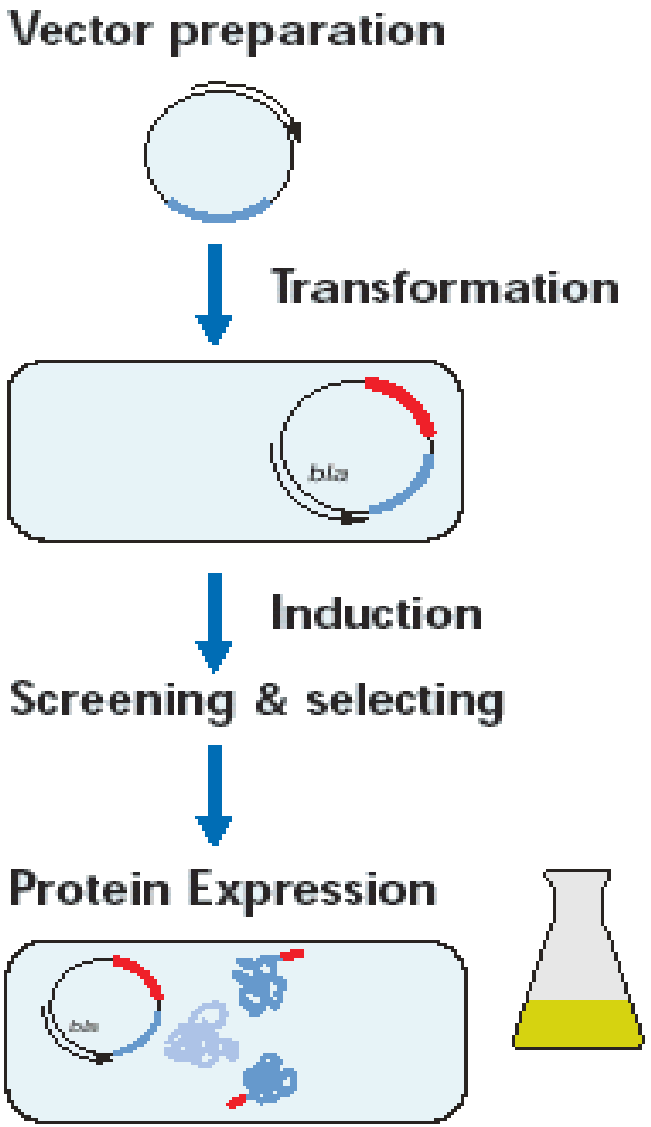
DNA construct	Origin of replication	Copy number	Classification
Plasmids			
pUC vectors	pMB1*	500–700	high copy
pBluescript® vectors	ColE1	300–500	high copy
pGEM® vectors	pMB1*	300–400	high copy
pTZ vectors	pMB1*	>1000	high copy
pBR322 and derivatives	pMB1*	15–20	low copy
pACYC and derivatives	p15A	10–12	low copy
pSC101 and derivatives	pSC101	~5	very low copy
Cosmids			
SuperCos	pMB1	10–20	low copy
pWE15	ColE1	10–20	low copy

* The pMB1 origin of replication is closely related to that of ColE1 and falls in the same incompatibility group. The high-copy plasmids listed here contain mutated versions of this origin.

Prokaryotic Expression vector



Protein expression in prokaryotic systems



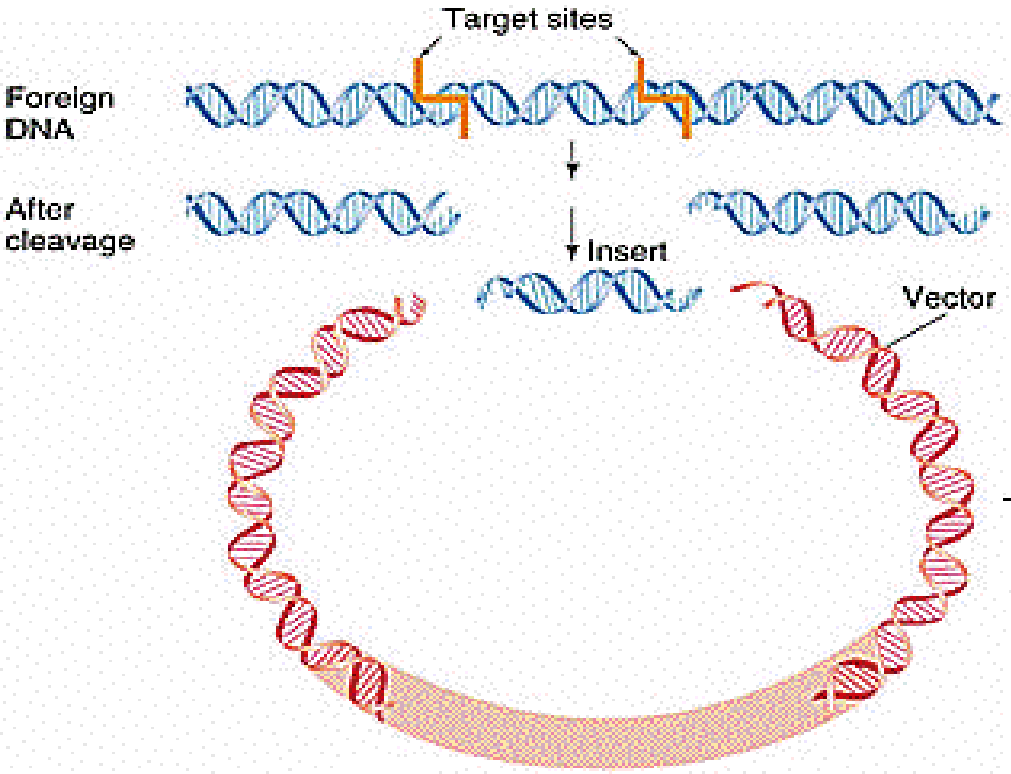
So, this new story would be about vectors again.

Bacterial expression vectors have some distinct features:

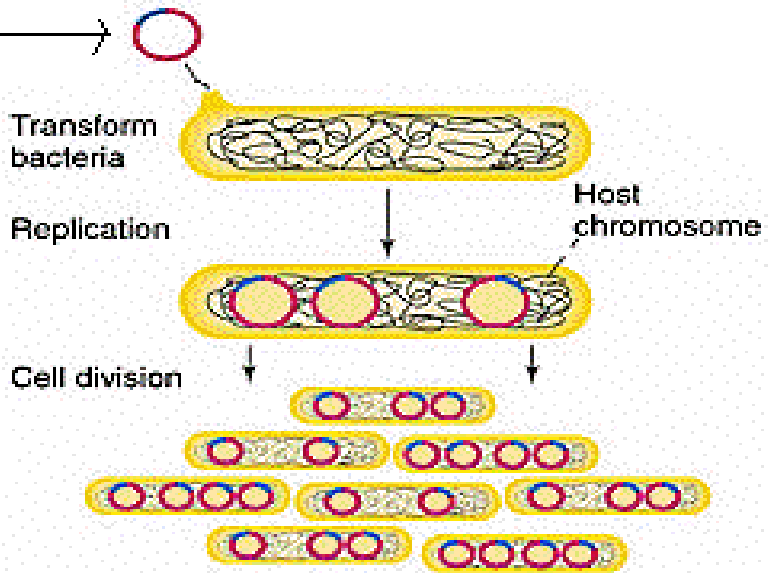
Inducible promoter systems;

Protein fusions including fused tags;

Transformation of plasmid DNA in competent *E. coli* cells



Competent (here) = able to uptake DNA

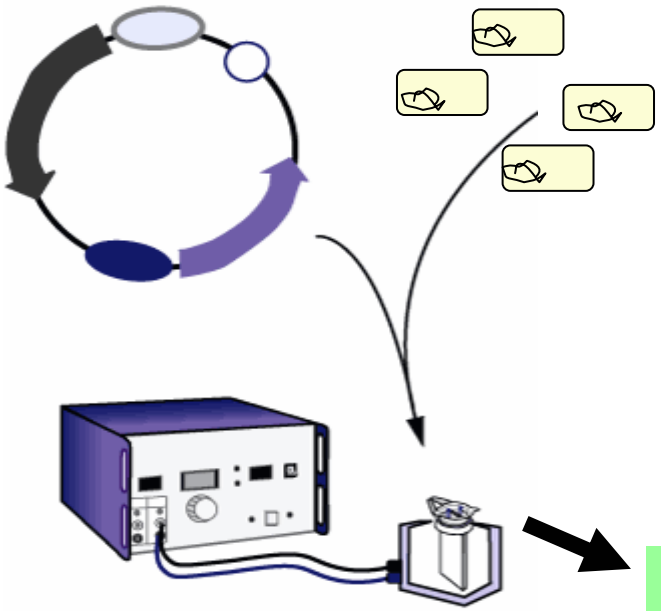


Transformation of plasmid DNA to competent *E. coli* cells

-- Electroporation and **electroporation-competent cells**

-- Heat shock transformation and **chemically competent cells**

Electroporation



Chemical transformation

treating E. coli CaCl2 will batter the membranes and essentially make the bacteria very unhappy.

CaCl2 is gaping holes in the membrane

BRIEF HEAT SHOCK

Recovery in rich growing media

General advices for one who wants to produce (eukaryotic) protein in prokaryotes

Most obvious and common mistakes:

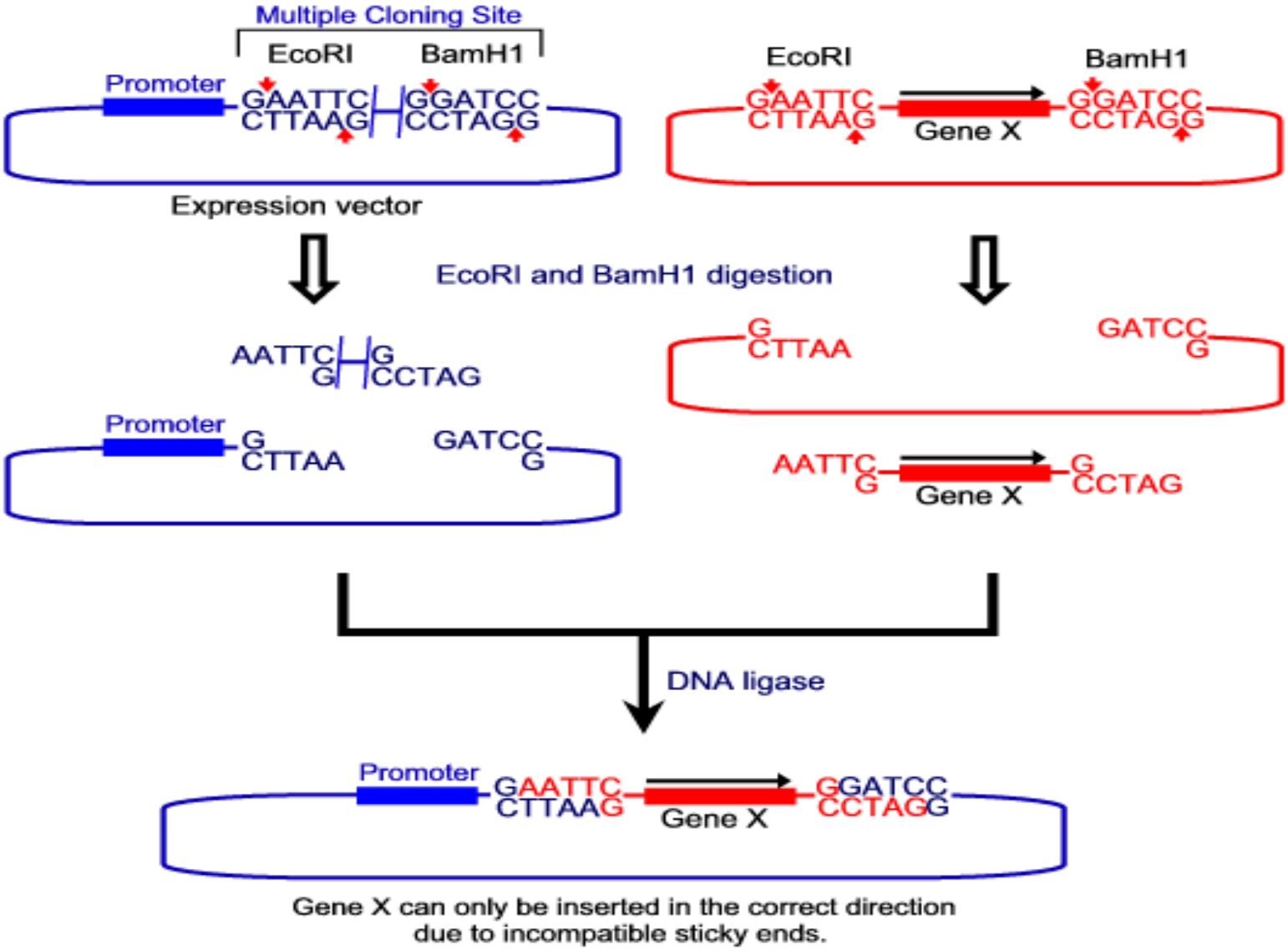
1. Do not forget to cut out the intron
Not an issue when you clone a cDNA

2. Check orientation of insert

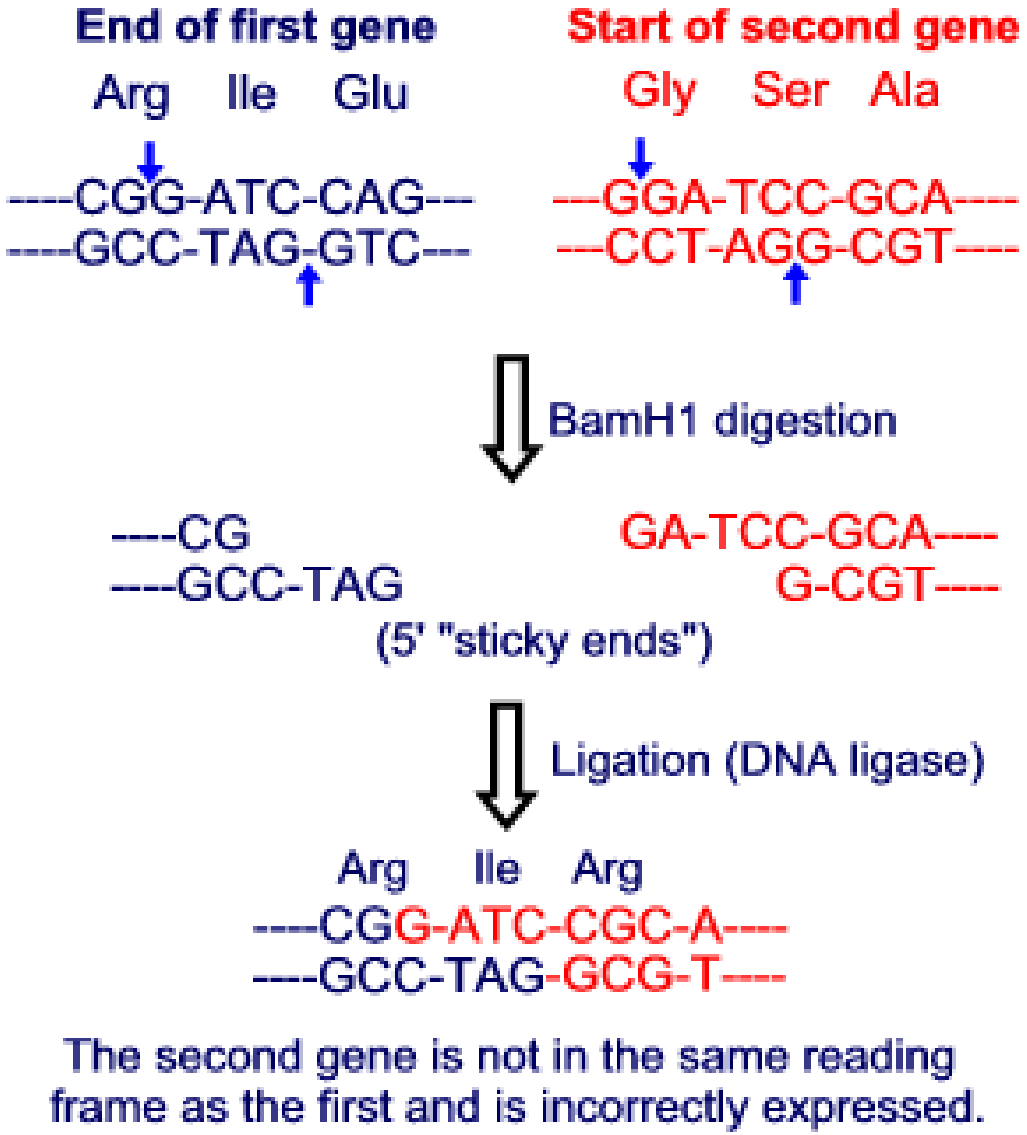
3. Do fusions with something In-frame

4. No Post-translation modification
= no product activity

Orientation of insert (could go backward, if cloned with same-type sticky ends) – use incompatible sticky ends



Fusion proteins

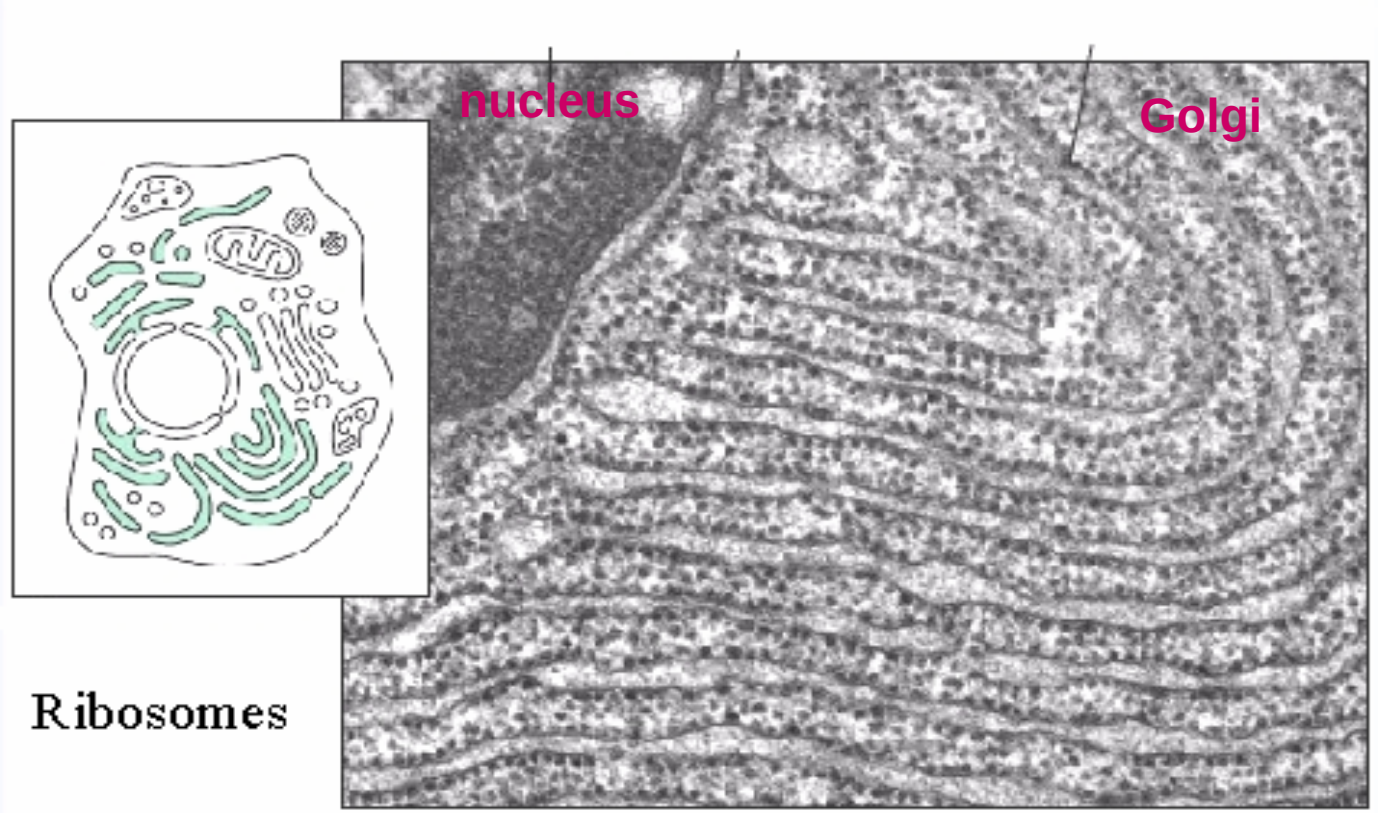


When expressing a fusion proteins, ensure that both of them are in the same reading frame

PostTranslational modification

Eukaryotic cells have Golgi system

Prokaryotic cells do not have it



Synthesis of secreted and membrane proteins

Efficiency of expression in E.coli

Dependent of:

1. **Type of transcription promoter and terminator**
2. **Affinity of mRNA and prokaryotic ribosome**
3. **Amount of copies of transgene and its localization (chromosome or plasmid)**
4. **Cellular localisation of the protein end-product**
5. **Efficiency of translation in the host organism**
6. **Stability of protein product in the host organism**

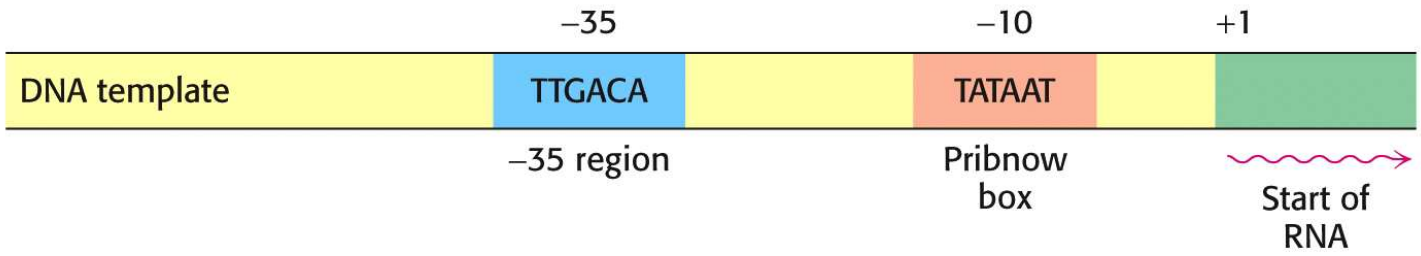
Systems could be optimized on gene to gene basis.
No universal strategy possible

Factors affecting transcription

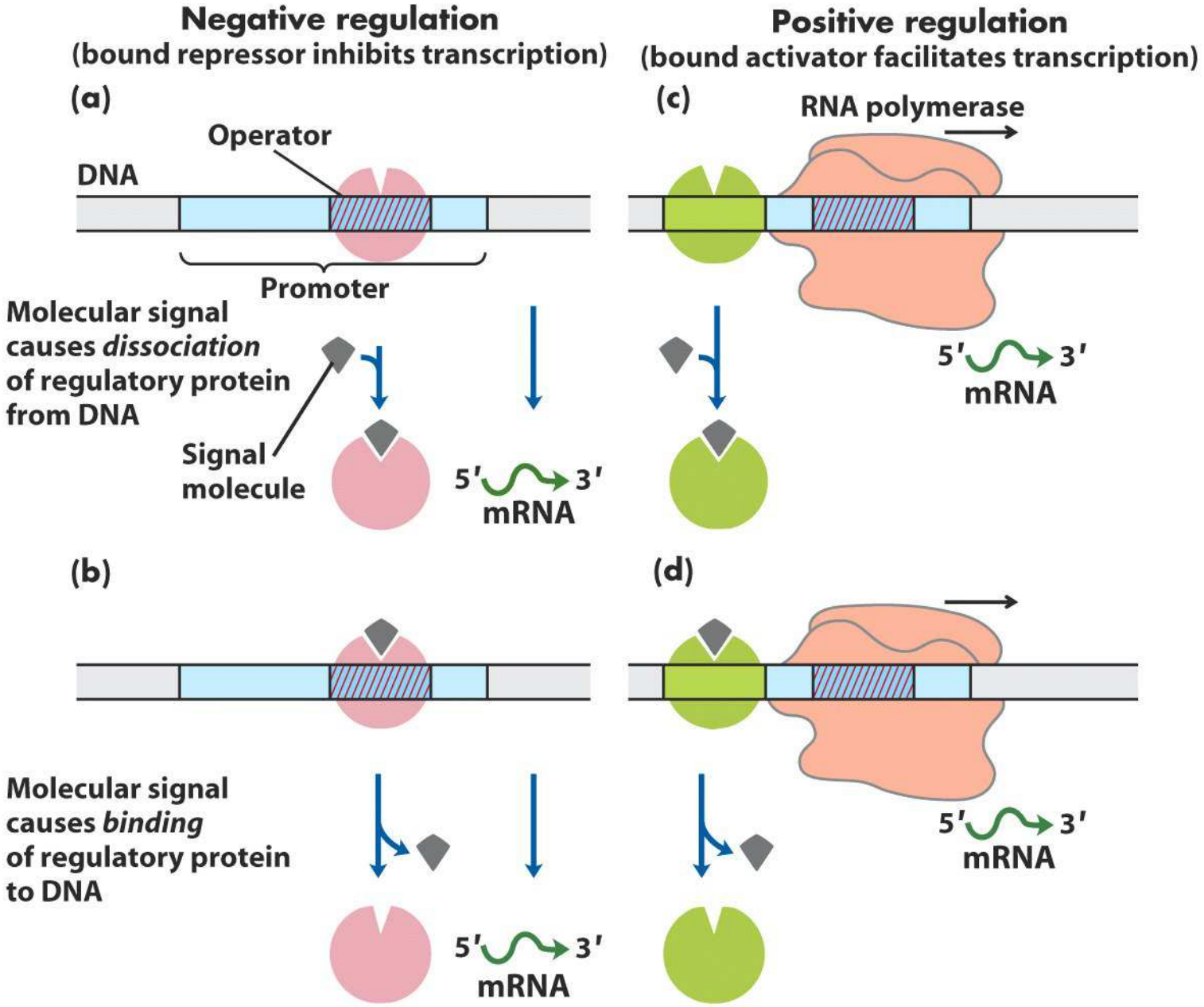
1.Promoters (including regulated ones)
PROKARYOTIC!!!!

2. Terminators
PROKARYOTIC!!!!

Promoters

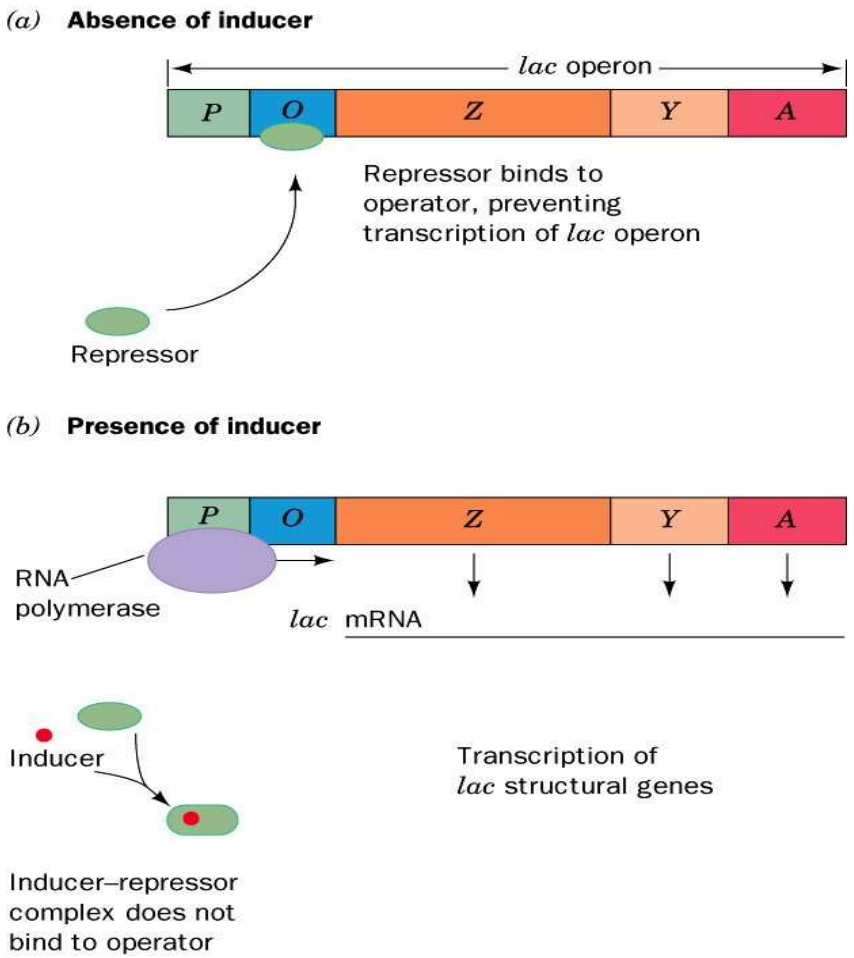


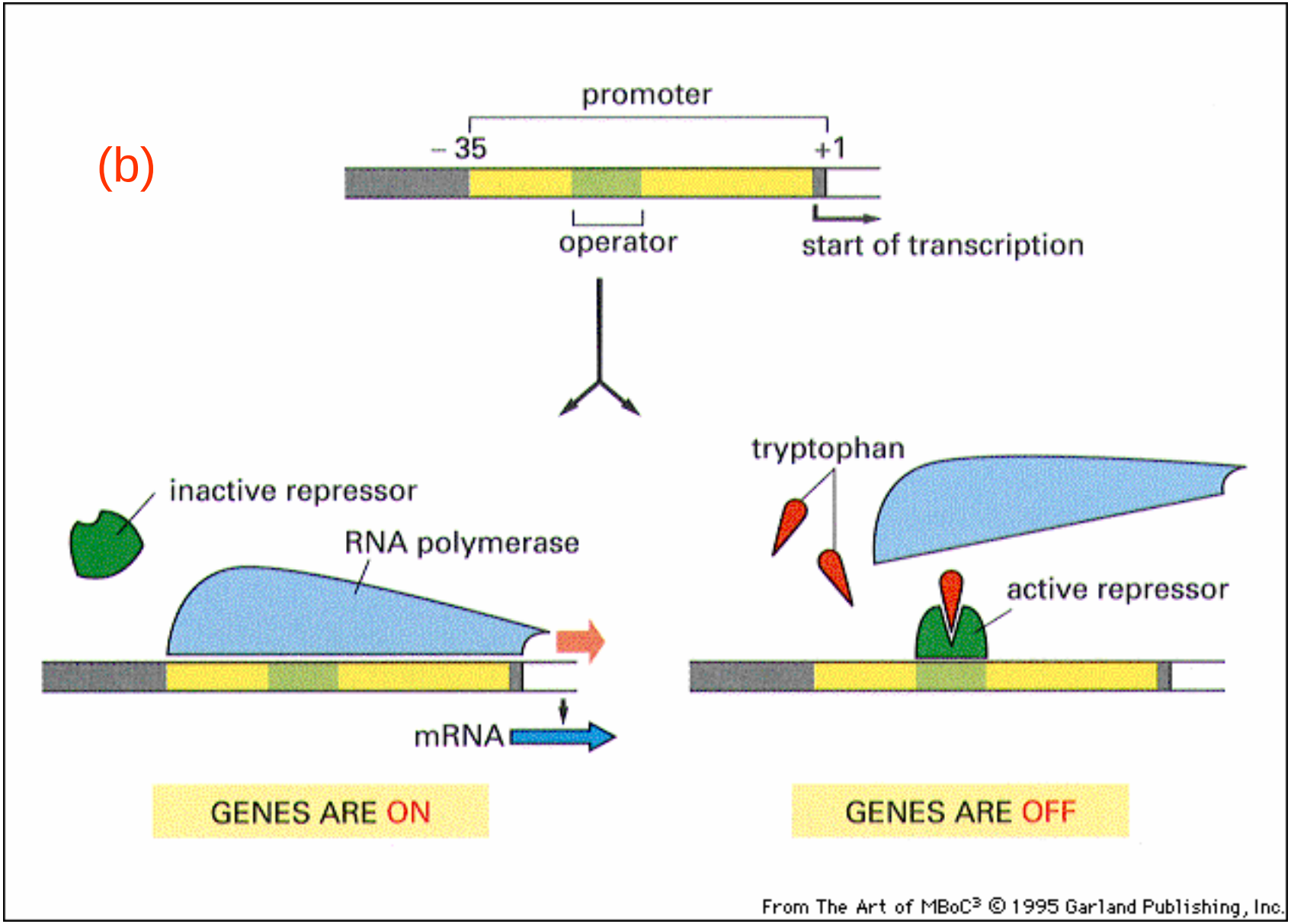
(A) **Prokaryotic promoter site**


Negative regulation
(bound repressor inhibits transcription)
(a)
Operator
DNA
Promoter
Molecular signal
causes *dissociation*
of regulatory protein
from DNA
Signal
molecule
5′
3′
mRNA
(b)
Molecular signal
causes *binding*
of regulatory protein
to DNA
5′
3′
mRNA
Positive regulation
(bound activator facilitates transcription)
(c)
RNA polymerase
5′
3′
mRNA
(d)
5′
3′
mRNA

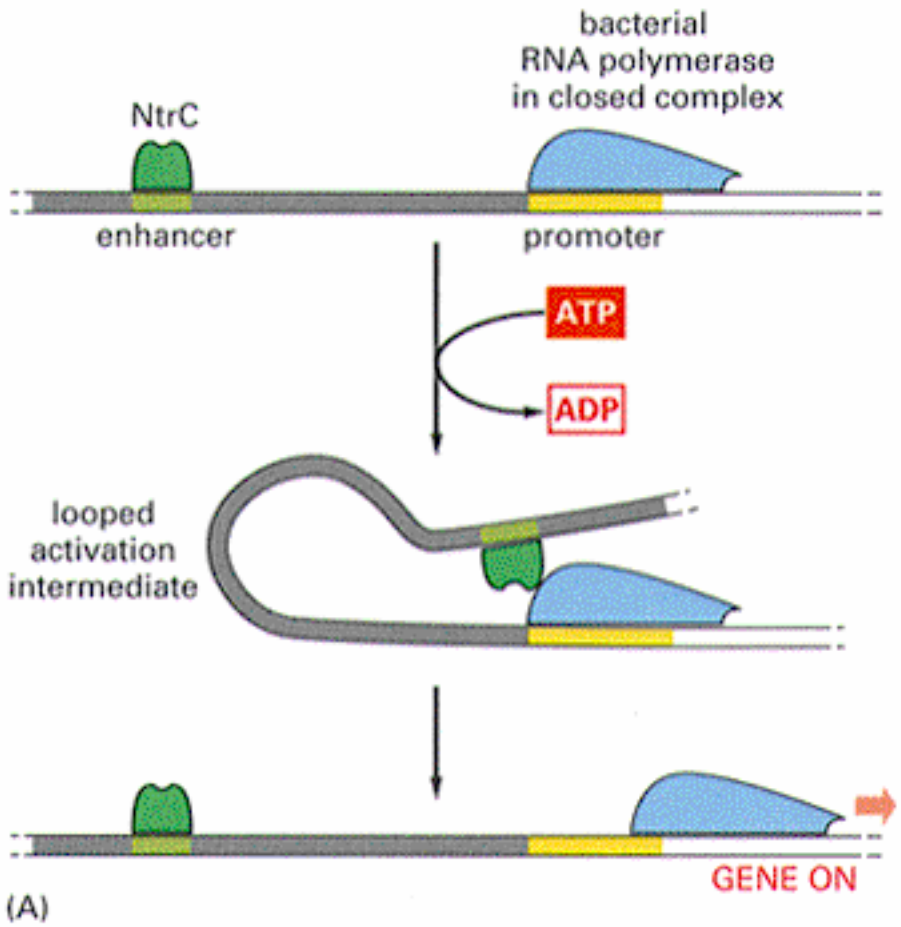
Control of transcription of the *lac* operon.

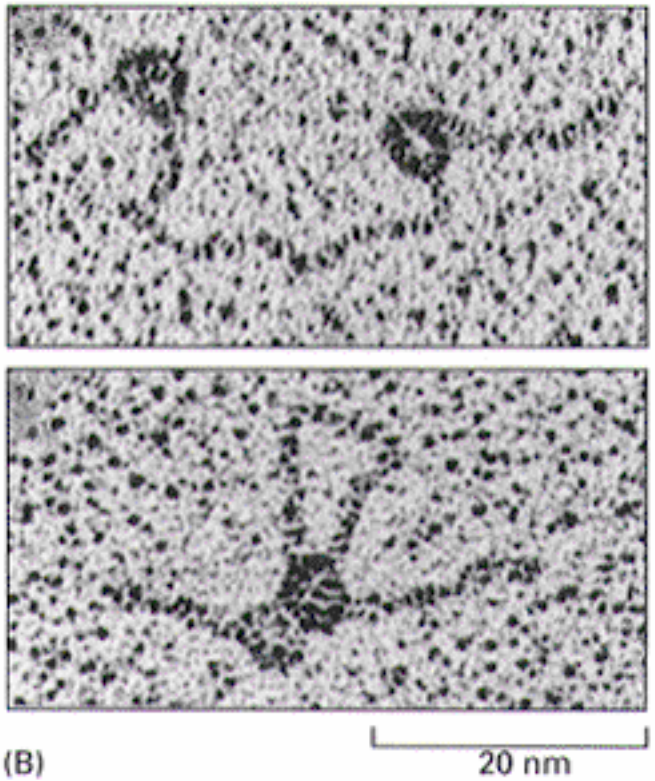
(a)




(b)
promoter
– 35
+1
operator
start of transcription
inactive repressor
RNA polymerase
mRNA
GENES ARE ON
tryptophan
active repressor
GENES ARE OFF
From The Art of MBoC³ © 1995 Garland Publishing, Inc.

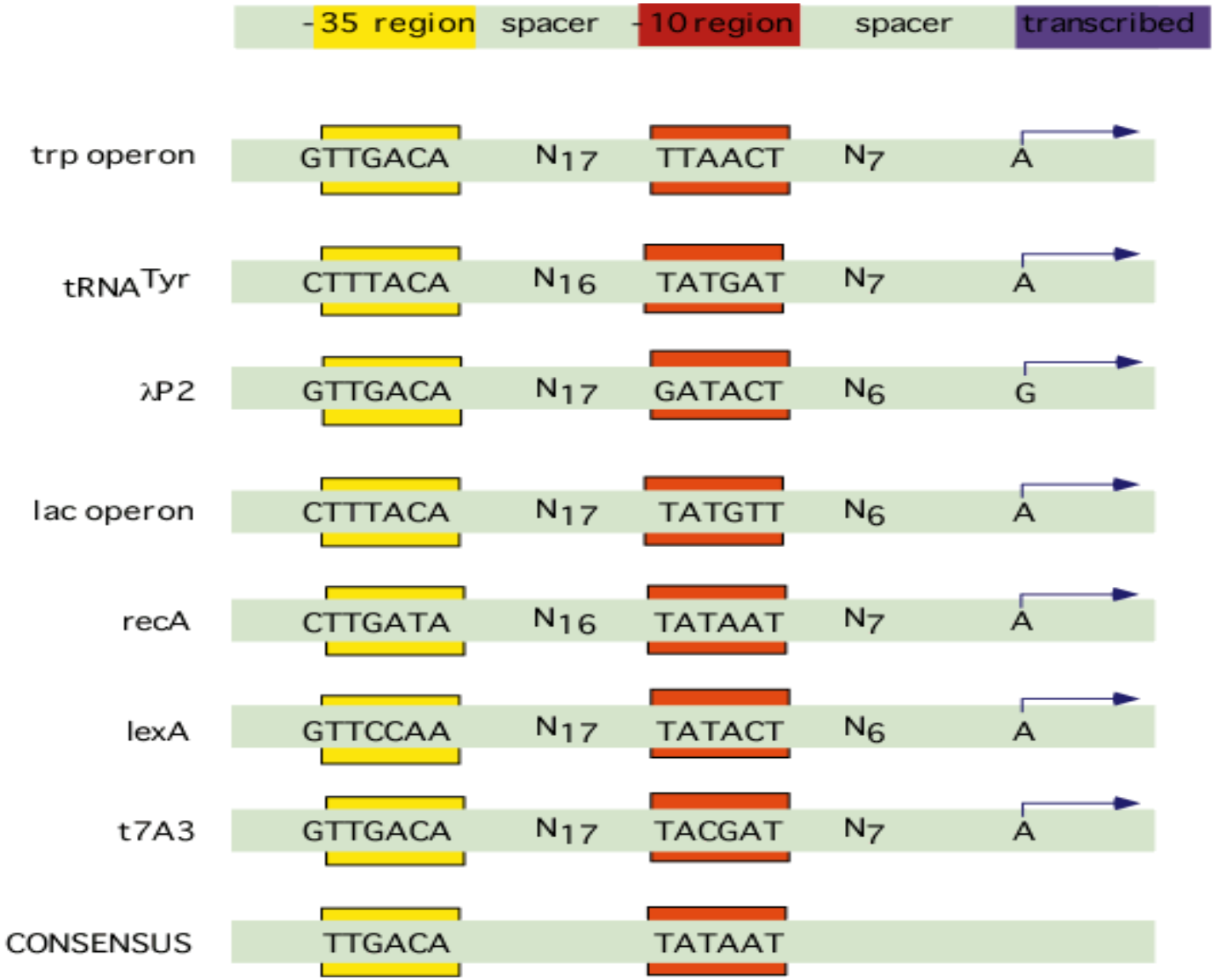
Nitrogen regulatory protein C - glutamine synthetase gene





From The Art of MBoC³ © 1995 Garland Publishing, Inc.

Variations between prokaryotic promoters are minimal



Inducible bacterial promoters

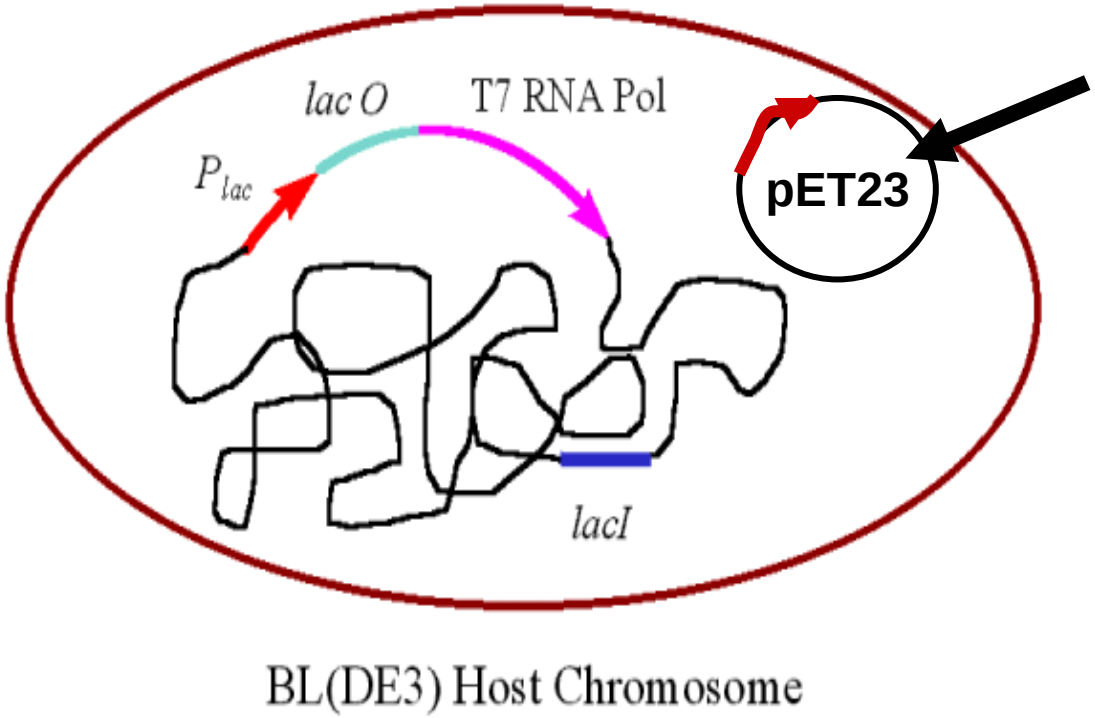
Why not to use constitutive, always strong promoter?

Because recombinant (alien) protein is often toxic for bacterial cell.

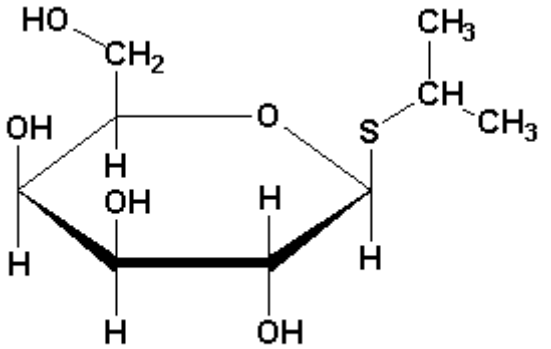
Bacteria tend to expel harmful plasmids

Induction

BL(DE3) inducible system and pET vectors (invented in 1984 by Bill Studier, on sale by Novagen)



Gene of interest is expressed from strong T7 promoter only if T7 RNA polymerase is made



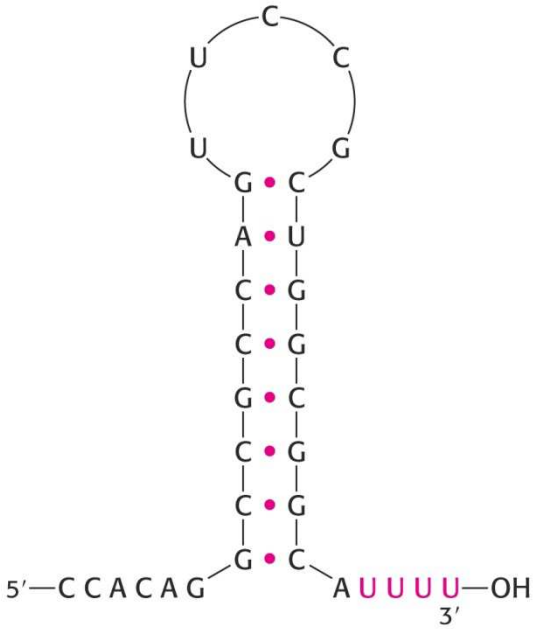
Isopropyl Thiogalactoside (IPTG)

1) T7 RNA polymerase gene is integrated in chromosome under the control of a *lac* promoter and operator

2) lactose analogue, IPTG, causes the host to produce T7 RNA polymerase

- **3) The *E. coli* host genome also carries the *lacI* (repressor) gene**

Terminators

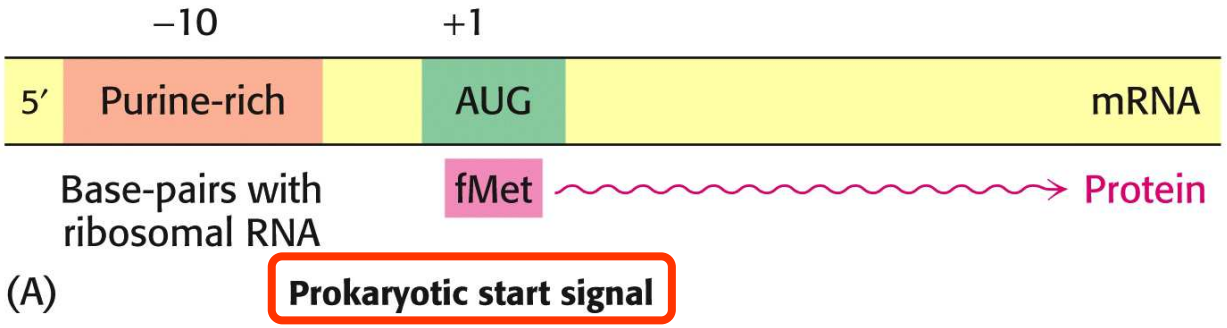


Factors affecting translation

1. Ribosome binding site (RBS)

2. Codon bias

3. Stability of the transcript



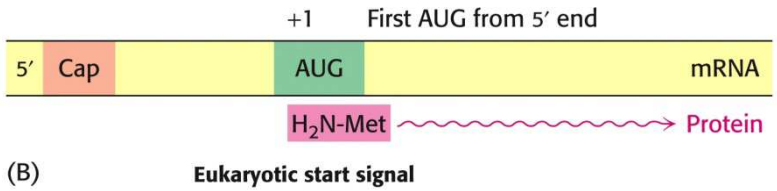
(A) **Prokaryotic start signal**

5′		3′	
AGCAC	GAGG	GAAAUCUGAUG	GAACGCUAC	*E. coli trpA*
UUUGGAU	GGAG	UGAAACGAUG	GCGAUUGCA	*E. coli araB*
GGUAAC	CAGGU	AACAACCAUG	CGAGUGUUG	*E. coli thrA*
CAAUUCAG	GGUGG	UGAAUGUG	AAACCAGUA	*E. coli lacI*
AAUCUU	GGAGG	CUUUUUUAUG	GUUCGUUCU	ϕX174 phage A protein
UAACU	AAGGA	UGAAAUGCAUG	UCUAAGACA	Qβ phage replicase
UCCU	AGGAGG	UUUGACCUAUG	CGAGCUUUU	R17 phage A protein
AUGUAC	UAAGGAGG	UUGUAUG	GAACAACGC	λ phage *cro*

Pairs with 16S rRNA

Pairs with initiator tRNA

Ribosome binding site (RBS)
Shine-Delgarno seq. AG rich



(B) **Eukaryotic start signal**

Codon usage

Optimal codons in fast-growing microorganisms, like *Escherichia coli* or *Saccharomyces cerevisiae* (baker's yeast), reflect the composition of their respective genomic tRNA pool. It is thought that optimal codons help to achieve faster translation rates and high accuracy

CODON USAGE IN *E. COLI* GENES[1]

	Codon	Amino acid[2]	%[3]	Ratio[4]	Codon	Amino acid	%	Ratio	Codon	Amino acid	%	Ratio	Codon	Amino acid	%	Ratio	
U	UUU	Phe (F)	1.9	0.51	UCU	Ser (S)	1.1	0.19	UAU	Tyr (Y)	1.6	0.53	UGU	Cys (C)	0.4	0.43	U
	UUC	Phe (F)	1.8	0.49	UCC	Ser (S)	1.0	0.17	UAC	Tyr (Y)	1.4	0.47	UGC	Cys (C)	0.6	0.57	C
	UUA	Leu (L)	1.0	0.11	UCA	Ser (S)	0.7	0.12	UAA	STOP	0.2	0.62	UGA	STOP	0.1	0.30	A
	UUG	Leu (L)	1.1	0.11	UCG	Ser (S)	0.8	0.13	UAG	STOP	0.03	0.09	UGG	Trp (W)	1.4	1.00	G
C	CUU	Leu (L)	1.0	0.10	CCU	Pro (P)	0.7	0.16	CAU	His (H)	1.2	0.52	CGU	Arg (R)	2.4	0.42	U
	CUC	Leu (L)	0.9	0.10	CCC	Pro (P)	0.4	0.10	CAC	His (H)	1.1	0.48	CGC	Arg (R)	2.2	0.37	C
	CUA	Leu (L)	0.3	0.03	CCA	Pro (P)	0.8	0.20	CAA	Gln (Q)	1.3	0.31	CGA	Arg (R)	0.3	0.05	A
	CUG	Leu (L)	5.2	0.55	CCG	Pro (P)	2.4	0.55	CAG	Gln (Q)	2.9	0.69	CGG	Arg (R)	0.5	0.08	G
A	AUU	Ile (I)	2.7	0.47	ACU	Thr (T)	1.2	0.21	AAU	Asn (N)	1.6	0.39	AGU	Ser (S)	0.7	0.13	U
	AUC	Ile (I)	2.7	0.46	ACC	Thr (T)	2.4	0.43	AAC	Asn (N)	2.6	0.61	AGC	Ser (S)	1.5	0.27	C
	AUA	Ile (I)	0.4	0.07	ACA	Thr (T)	0.1	0.30	AAA	Lys (K)	3.8	0.76	AGA	Arg (R)	0.2	0.04	A
	AUG	Met (M)	2.6	1.00	ACG	Thr (T)	1.3	0.23	AAG	Lys (K)	1.2	0.24	AGG	Arg (R)	0.2	0.03	G
G	GUU	Val (V)	2.0	0.29	GCU	Ala (A)	1.8	0.19	GAU	Asp (D)	3.3	0.59	GGU	Gly (G)	2.8	0.38	U
	GUC	Val (V)	1.4	0.20	GCC	Ala (A)	2.3	0.25	GAC	Asp (D)	2.3	0.41	GGC	Gly (G)	3.0	0.40	C
	GUA	Val (V)	1.2	0.17	GCA	Ala (A)	2.1	0.22	GAA	Glu (E)	4.4	0.70	GGA	Gly (G)	0.7	0.09	A
	GUG	Val (V)	2.4	0.34	GCG	Ala (A)	3.2	0.34	GAG	Glu (E)	1.9	0.30	GGG	Gly (G)	0.9	0.13	G
	U				C				A				G				

[1] The data shown in this table is from the Arabidopsis Research Companion on the World Wide Web (//weeds/mgh.harvard.edu). Codon frequencies for many other bacteria can be found at http://morgan.angis.su.oz.au/Angis/Tables.html.

[2] The letter in parenthesis represents the one-letter code for the amino acid.

[3] % represents the average frequency this codon is used per 100 codons.

[4] Ratio represents the abundance of that codon relative to all of the codons for that particular amino acid.

Codon Usage in E. coli & humans

Codon	Amino acid	Frequency of use in:	
		E. coli	Humans
GAG	Glutamic acid	0.30	0.59
GAA	Glutamic acid	0.70	0.41
CGG	Arginine	0.08	0.19
CGA	Arginine	0.05	0.10
CGU	Arginine	0.42	0.09
CGC	Arginine	0.37	0.19
AGG	Arginine	0.03	0.22
AGA	Arginine	0.04	0.21
CCG	Proline	0.55	0.11
CCA	Proline	0.20	0.27
CCU	Proline	0.16	0.29
CCC	Proline	0.10	0.33
UGA	Stop	0.30	0.61
UAG	Stop	0.09	0.17
UAA	Stop	0.62	0.22

Codon Optimization Strategies

- Chemically synthesize new gene
  – Alter sequence of the gene of interest
  to match donor codons to the codons
  most frequently used in host organism
- Express in different host
  – choose host with better matching codon usage
- Use an engineered host cell
  that overexpresses low abundance tRNAs

Factors affecting protein stability

1.Overall level of protease activity in bacterial cells

2. N-terminal amino acid affects protein half-life

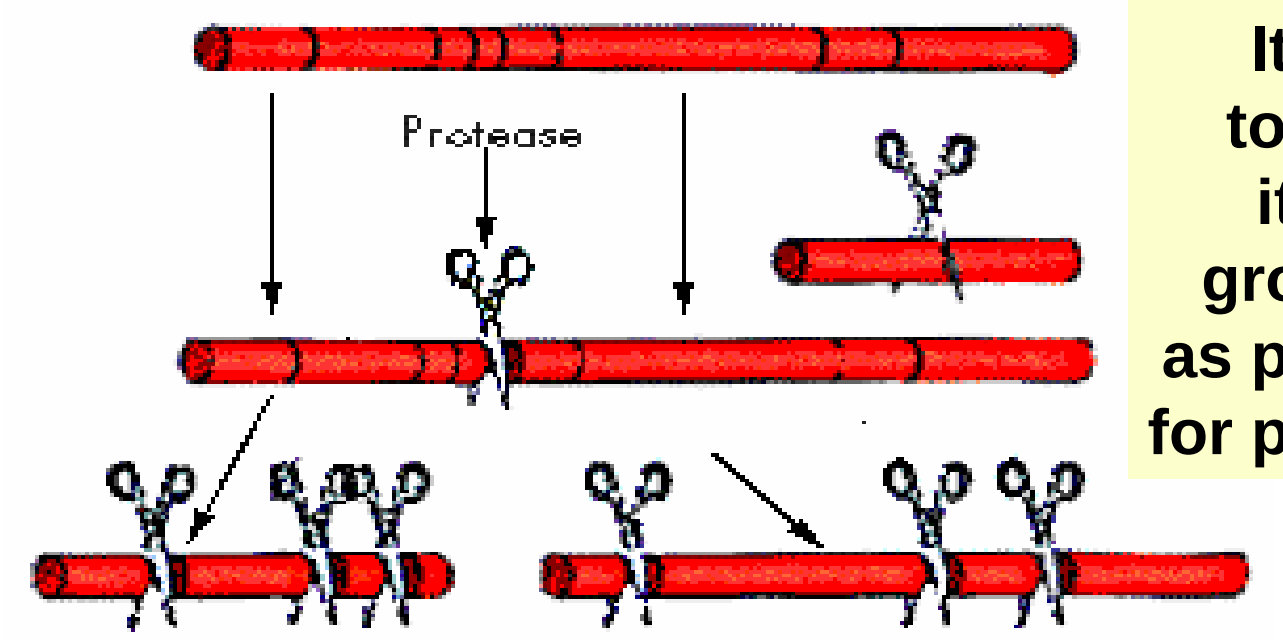
3. Internal regions containing clusters of certain amino acids can increase proteolysis

P proline
E glutamic acid
S serine
T threonine

…. Mutate PEST aminoacids….

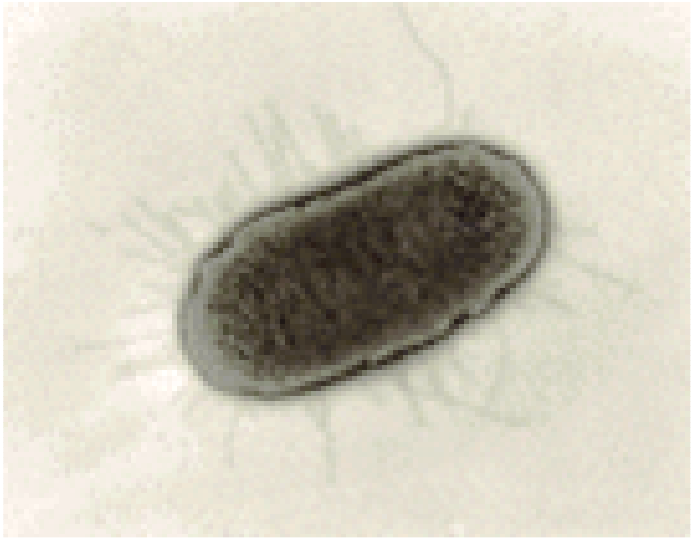
Protease-deficient host strains

**BL21, the work horse of *E. coli* expression,
is deficient in two proteases
encoded by the *lon* (cytoplasmic)
and *omp*T (periplasmic) genes.**



**It is dangerous
to kill proteases,
it makes E.coli
grow much slowly
as proteases needed
for proper metabolism**

Bacterial host engineering



Escherichia coli (E. coli) is a type of bacteria normally found in the intestines of people and animals.

Although most strains of E. coli are harmless, some can cause illness or even death. The most serious form is E. coli **0157:H7**.

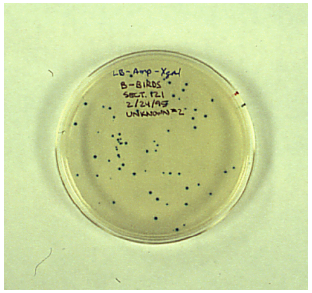
E. coli leads to about 73,000 cases of infection and 61 deaths each year in the United States.

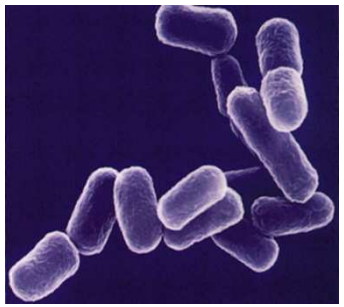
E.Coli K12 strain has been used for further engineering

The K12 strain was first isolated in 1921 from the stool of a malaria patient and it has been maintained in laboratory stocks as a pure strain.

NIH Recombinant Advisory Committee (RAC) (1973)

Asilomar Conference on Recombinant DNA (February 1975)





Every strain comes with description of its genotype:
DH5alpha (recA-; hsdR-; LacIq; uvrA-; mcrA-……)

Most strain in molecular biology are *recA- endA- hsdR- …*

Additional changes in K12 E.coli for ease of the laboratory practice

1. Bacterial restriction modification systems have been removed.

(To prevent its interferention with the replication of foreign DNA in bacteria).

hsdR/hsdM/hsdS (EcoK) restriction system

Degrades DNA not methylated at the sequence 5'-AAC-(N)5-GTGC-3'

hsdM **recognises unmethylated DNA**
hsdM **is also involved in methylation of DNA**
hsdR **encodes an endonulease**
hsdS **encodes DNA sequence specific protein**

hsdR- **or** ***hsdS-*** ***mutants*** **facilitate propagation of any foreign DNA**

mcrA/mcrB/mrr complex

E.coli DNA is methylated by ***dcm, dam*** **and** ***hsdM***

mcrA/mcrB/mrr cleaves DNA methylated by other systems

-mcrA-/mcrB-
strains are good for cloning eukaryotic DNA

2. DNA recombination systems are modified to prevent rearrangements (RecA-)

(to prevent deletions and rearrangements)

***recA* is a core recombination protein**

***recA*- strains allow cloning of repetitive sequences**

***recA*-/*recB*-/*recC*- are enhanced strains with very low recombination efficiency**

***uvrC/umuC* are involved in DNA repair**

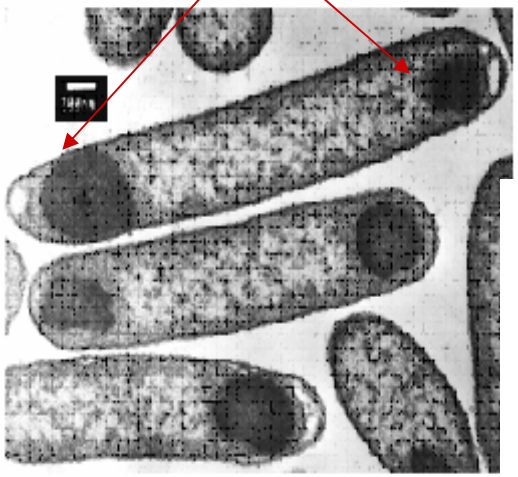
***uvrC*-/*umuC*- are good for cloning of inverted repeats**

3. Endonuclease activity has been mutated (EndA-)

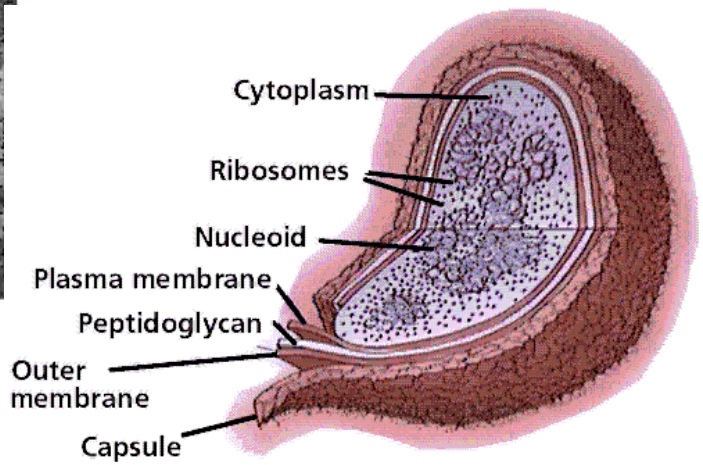
(to increase plasmid yields and improve the quality of DNA – no nicks)

Where your expressed protein will be located?

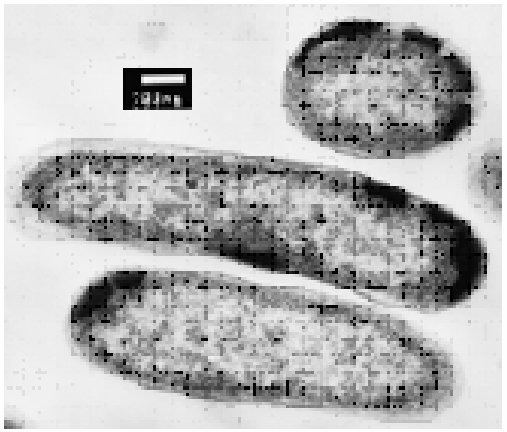
Inclusion bodies (insoluble)



Cytoplasm (soluble)



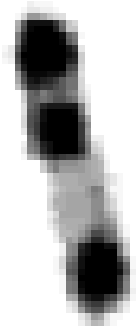
Periplasmatic space (soluble or insoluble)



1. Inclusion bodies (most common case)

-- Inclusion bodies are formed through the accumulation of folding intermediates rather than from the native or unfolded proteins.

-- It is not possible to predict which proteins will be produced as inclusion bodies.



-- Production of inclusion bodies not dependent on the origin of protein, the used promoters, the hydrophobicity of target proteins...

Good side of inclusion bodies

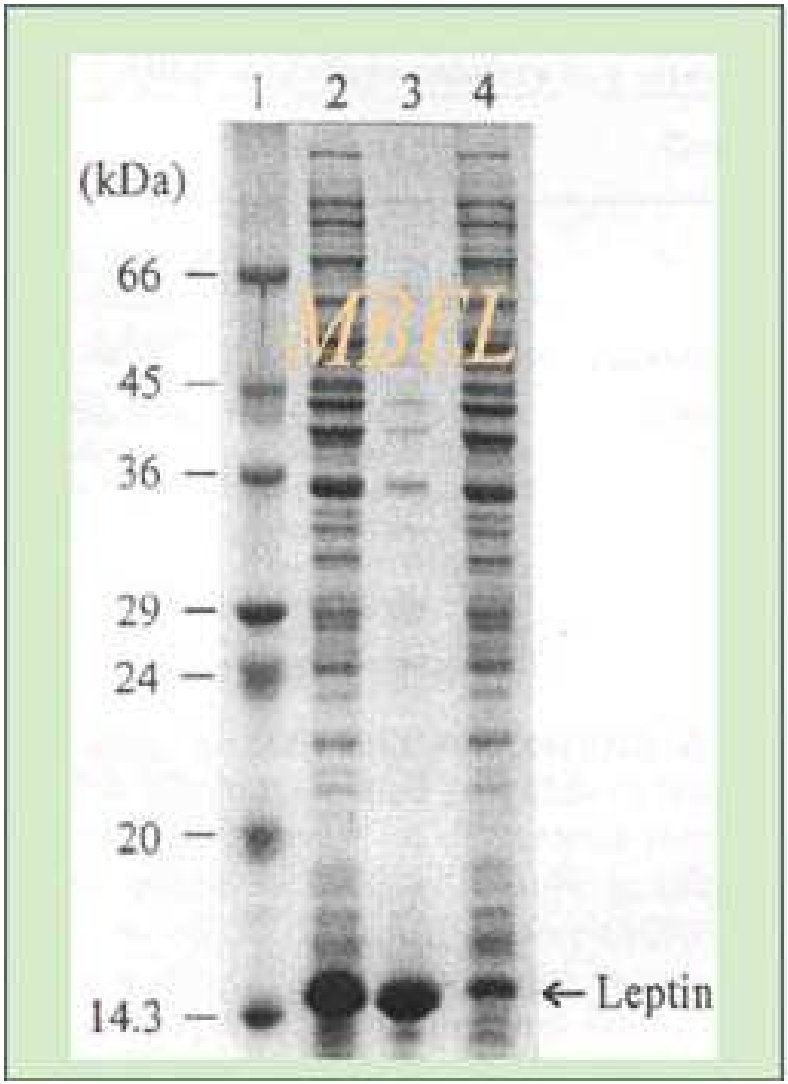
1) inclusion bodies can be accumulated in the cytoplasm to much higher level (greater than 25%) than production as soluble form;

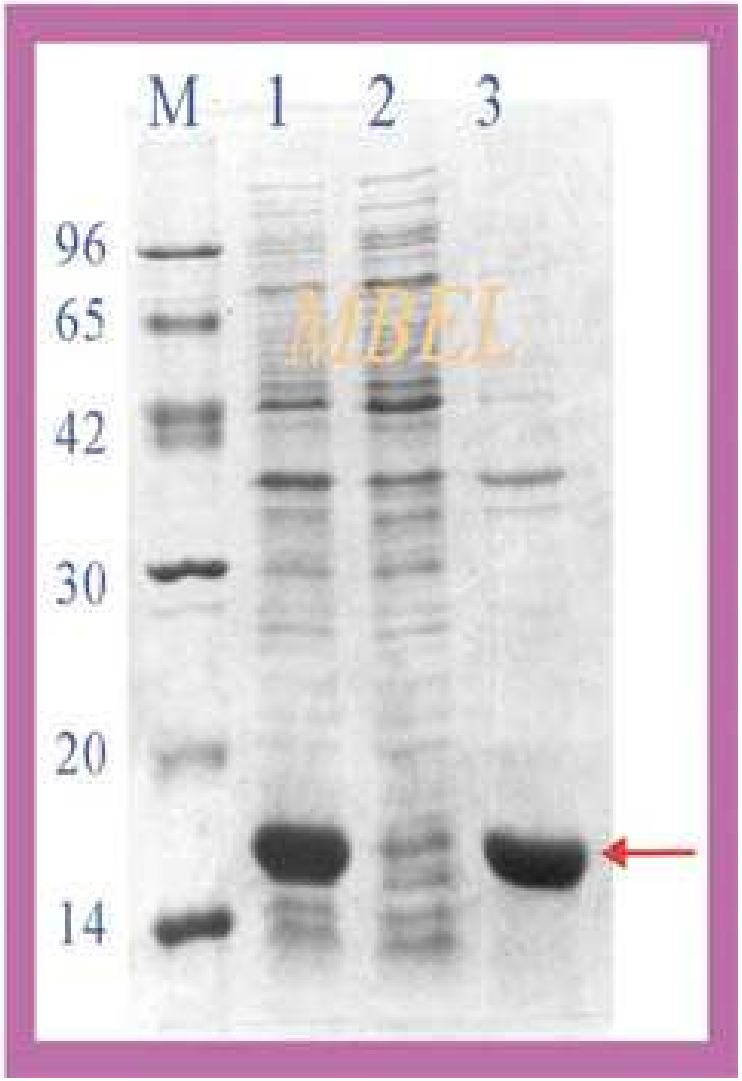
2) inclusion bodies is initially isolated in a highly purified, solid, and concentrated state by simple physical operation (centrifugation).

3) inclusion bodies have no biological activity. For toxic proteins it may be the only one available;

4) inclusion bodies are resistant to proteolysis That results in the high yield of protein production.

SDS-PAGE analysis of recombinant protein produced as inclusion body





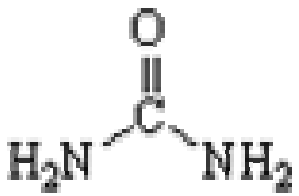
hG-CSF

Recovery of proteins from inclusion bodies

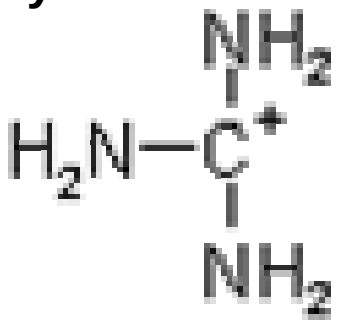
Is not a straightforward process, but road of trials and errors

Solubilization

Choice of solubilizing agents, e.g., urea, guanidine HCl, or detergents, plays a key role in solubilization efficiency



Urea



Guandinium

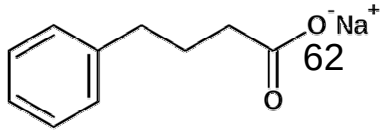
Refolding

-- Refolding is initiated by reducing concentration of denaturant used to solubilize IBs.

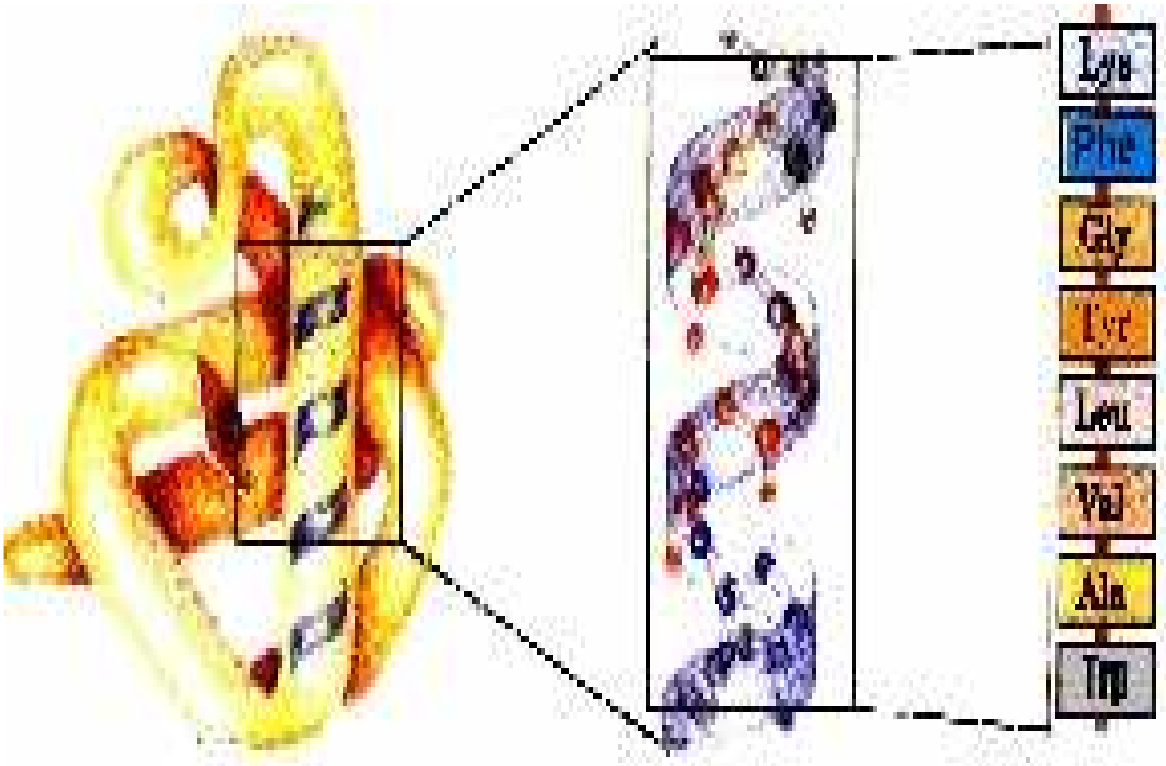
-- Refolding competes with other reactions, such as misfolding and aggregation (both are leading to bad results)

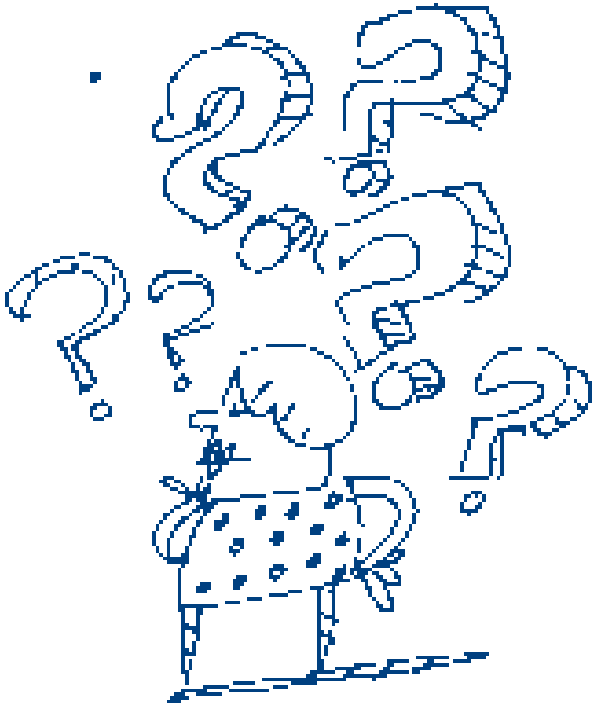
-- Chaperones are helpful in refolding (including chemical chaperones)

Phenylbutyrate



Question of questions – how to purify your protein?





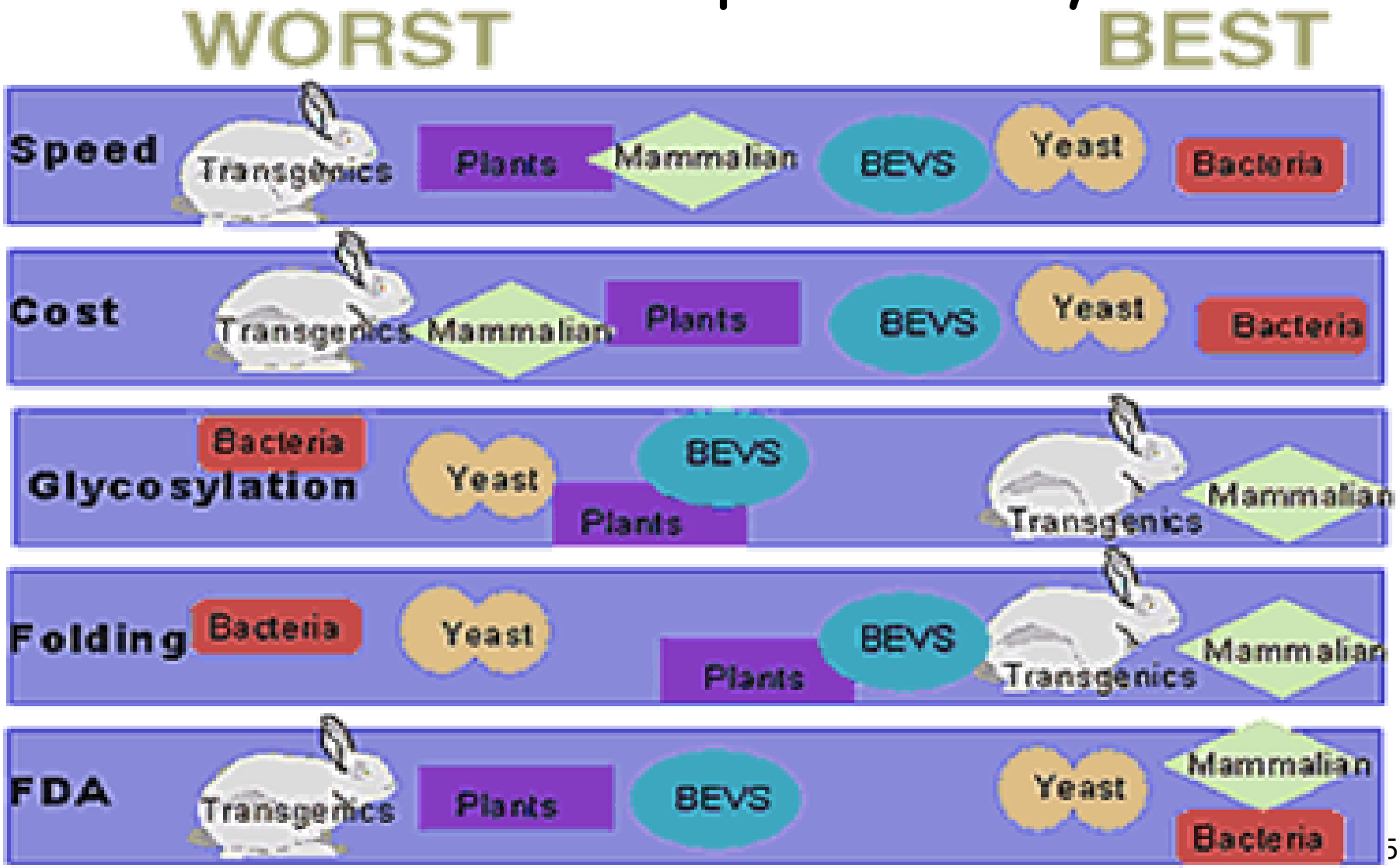
Roman SOBOTKA will provide answers

Comparison of expression systems

Table 1
Comparison of Expression Systems

Desired characteristics	Expression system			
	Bacteria	Yeast	Insect	Mammalian cell culture
Cell growth	Rapid	Rapid	Slow	Slow
Complexity of growth medium	Minimum	Minimum	Complex	Complex
Cost of growth medium	Low	Low	High	High
Expression level	High	Low to high	Low to high	Low to moderate
Extracellular expression	Secretion to periplasm	Secretion to medium	Secretion to medium	Secretion to medium
Posttranslational modifications				
Protein folding	Refolding usually required	Refolding may be required	Proper folding	Proper folding
N-linked glycosylation	None	High mannose	Simple, no sialic acid	Complex
O-linked glycosylation	No	Yes	Yes	Yes
Phosphorylation	No	Yes	Yes	Yes
Acetylation	No	Yes	Yes	Yes
Acylation	No	Yes	Yes	Yes
γ-Carboxylation	No	No	No	Yes

Competitiveness of different expression systems



Baculovirus Expression Vector System (BEVS) – insects; Mammalian = cell lines