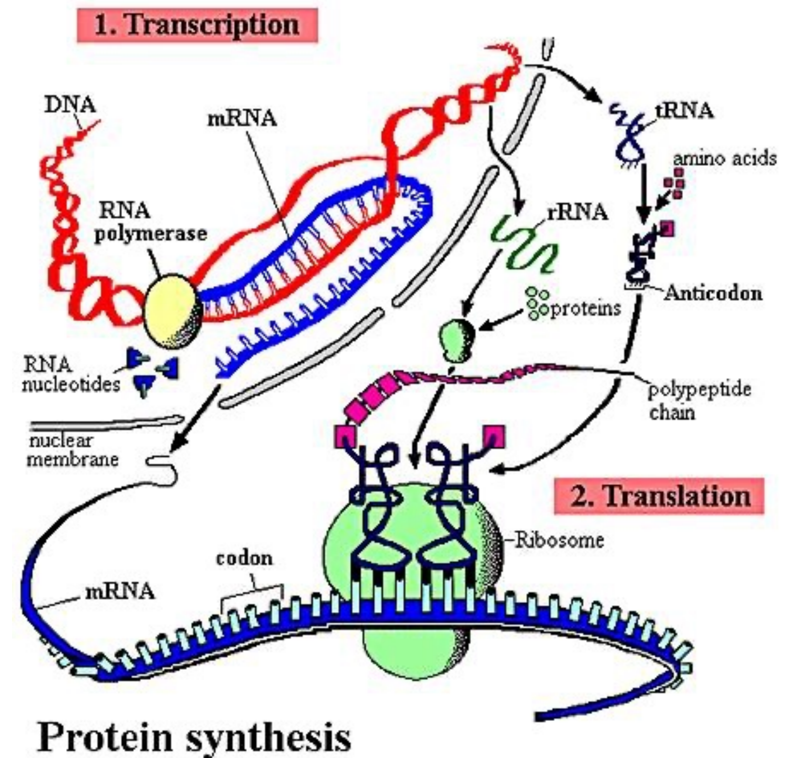
Production of recombinant proteins

Methods and strategies

To begin…
Key steps in protein biosynthesis

- Information on the coding sequence of the protein is stored in DNA
- This information is transferred to RNA ➡ TRANSCRIPTION
- RNA is decoded by the ribosome to produce the protein ➡ TRANSLATION

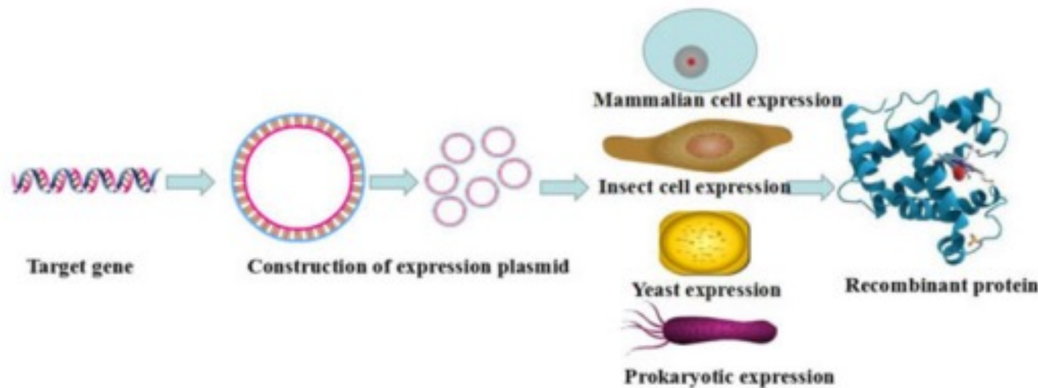
These mechanisms and codes are CONSERVED throughout all kingdoms of life



Protein synthesis

Recombinant proteins: what are they?

- A recombinant protein is a protein obtained through the use of recombinant DNA technology



- Recombinant DNA technology: DNA sequences from different sources are joined to produce new DNA molecules

Recombinant proteins: why?

It becomes possible to produce virtually **unlimited** amounts of **any** protein

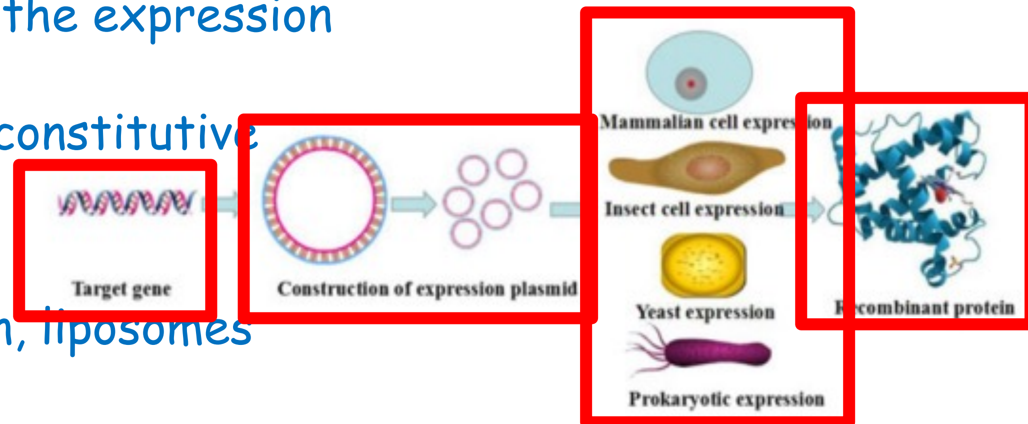
Recombinant proteins: why?

1. Academia: studies of structure-function relationships
   - what does the protein look like? structure
   - how does the protein work? function and mechanism
2. Biopharma: production of drugs (e.g. insulin and hormones, antibodies, clotting factors) or antigens for vaccines (e.g. the hepatitis B surface antigen HBsAg, spike)
3. Industry: production of proteins/enzymes for industrial applications (e.g. detergents, biodiesel)

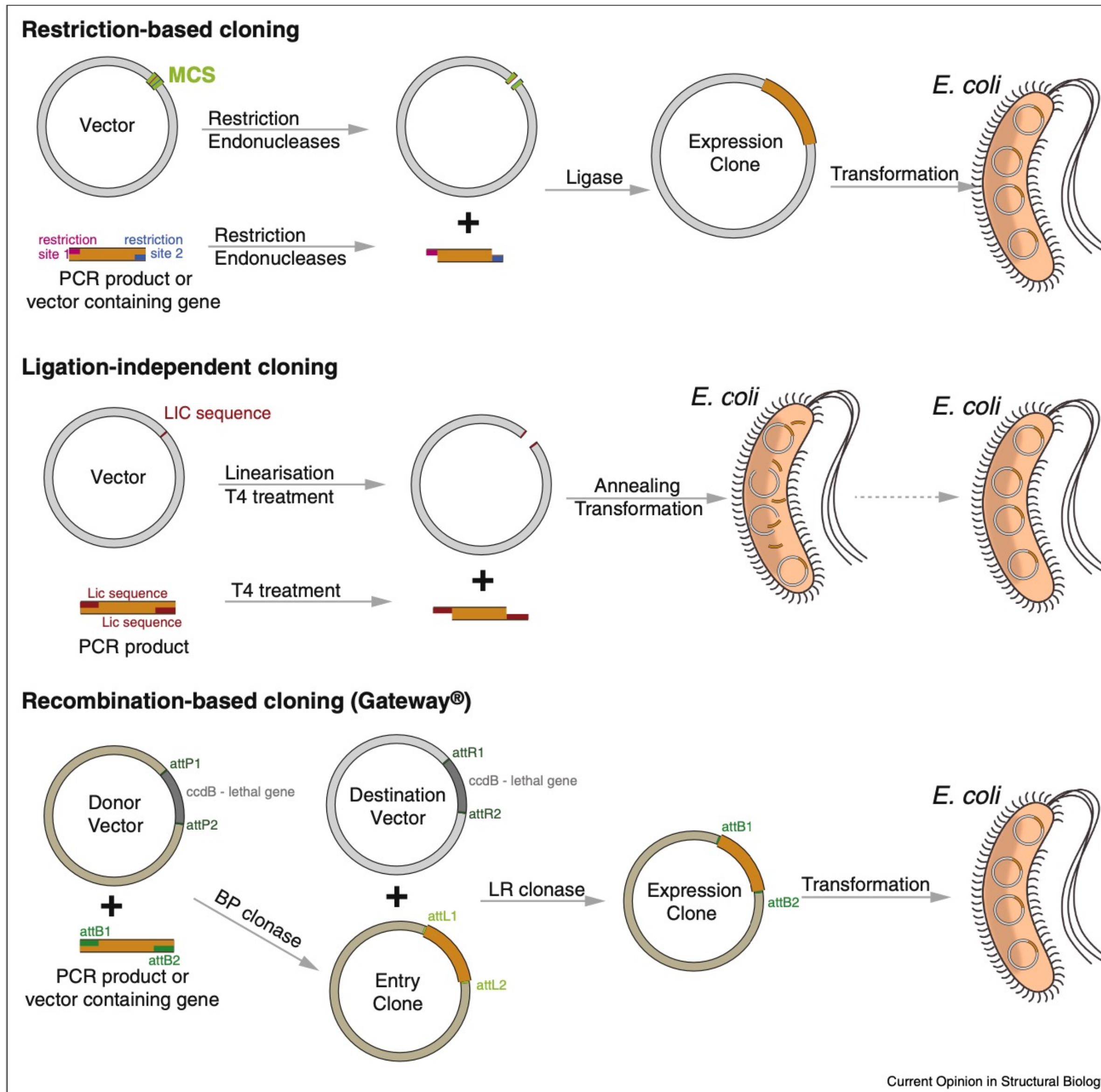
Each of these applications has different requirements as far as quality and quantity of the protein are concerned.

Recombinant proteins: how?

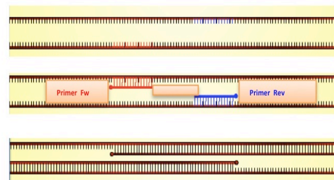
1. Synthesis/cloning of the target gene coding sequence
   - PCR-mediated strategies: availability of DNA/RNA template is essential!
2. Cloning of the coding sequence in the expression vector
   - a plasmid with an inducible or constitutive promoter
3. Trasformation of the host cell
   - thermal shock, electroporation, liposomes
4. Selection of recombinant cells
   - antibiotic resistance, selective advantage
5. Expression and purification of the protein
   - intracellular or secreted



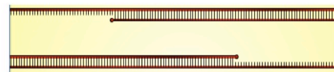
Cloning strategies if the coding sequence is known

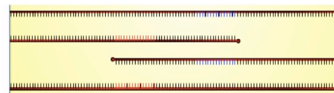


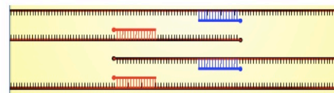
Polymerase chain reaction (PCR)

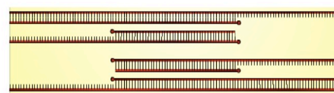


1° ciclo

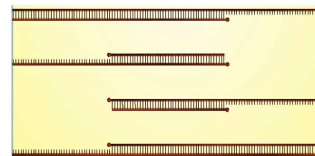


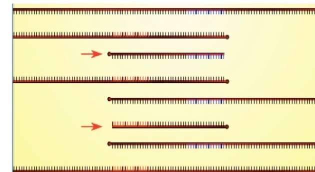


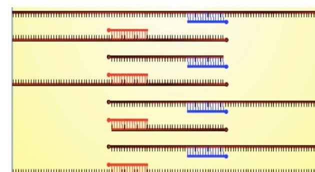


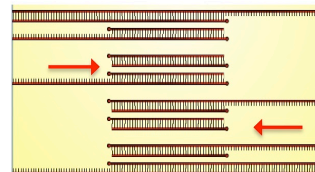


2° ciclo









3° ciclo

Le sequenze target sono indicate dalle frecce rosse

Phases of a PCR:

- Denaturation
- Annealing
- Extension

In the test tube:

- Template
- two primers
- dNTP
- Buffer + $MgCl_2$
- Taq DNA polymerase

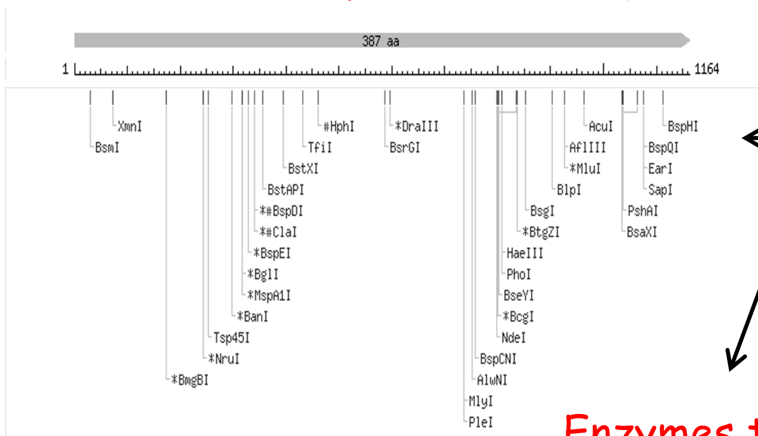
Restriction-based cloning strategy by PCR if the coding sequence is known
1. restriction analysis of the gene

Sequence of the gene

>gi|49175990:3153377-3154540 *Escherichia coli* str. K-12 substr. MG1655 chromosome, complete genome

ATGAACAACTTTAATCTGCACACCCCAACCCGCATTCTGTTTGGTAAAGGCGCAATCGCTGGTTTACGCG
AACAAATTCCTCACGATGCTCGCGTATTGATTACCTACGGCGGCGGCAGCGTGAAAAAAACCGGCGTTCT
CGATCAAGTTCTGGATGCCCTGAAAGGCATGGACGTGCTGGAATTTGGCGGTATTGAGCCAAACCCGGCT
TATGAAACGCTGATGAACGCCGTGAAACTGGTTCGCGAACAGAAAGTGACTTTCCTGCTGGCGGTTGGCG
GCGGTTCTGTACTGGACGGCACCAAATTTATCGCCGCAGCGGCTAACTATCCGGAAAATATCGATCCGTG
GCACATTCTGCAAACGGGCGGTAAAGAGATTAAAAGCGCCATCCCGATGGGCTGTGTGCTGACGCTGCCA
GCAACCGGTTCAGAATCCAACGCAGGCGCGGTGATCTCCCGTAAAACCACAGGCGACAAGCAGGCGTTCC
ATTCTGCCCATGTTCAGCCGGTATTTGCCGTGCTCGATCCGGTTTATACCTACACCCTGCCGCCGCGTCA
GGTGGCTAACGGCGTAGTGGACGCCTTTGTACACACCGTGGAACAGTATGTTACCAAACCGGTTGATGCC
AAAATTCAGGACCGTTTCGCAGAAGGCATTTTGCTGACGCTAATCGAAGATGGTCCGAAAGCCCTGAAAG
AGCCAGAAAACTACGATGTGCGCGCCAACGTCATGTGGGCGGCGACTCAGGCGCTGAACGGTTTGATTGG
CGCTGGCGTACCGCAGGACTGGGCAACGCATATGCTGGGCCACGAACTGACTGCGATGCACGGTCTGGAT
CACGCGCAAACACTGGCTATCGTCCTGCCTGCACTGTGGAATGAAAAACGCGATACCAAGCGCGCTAAGC
TGCTGCAATATGCTGAACGCGTCTGGAACATCACTGAAGGTTCCGATGATGAGCGTATTGACGCCGCGAT
TGCCGCAACCCGCAATTTCTTTGAGCAATTAGGCGTGCCGACCCACCTCTCCGACTACGGTCTGGACGGC
AGCTCCATCCCGGCTTTGCTGAAAAAACTGGAAGAGCACGGCATGACCCAACTGGGCGAAAATCATGACA
TTACGTTGGATGTCAGCCGCCGTATATACGAAGCCGCCCGC**TAA**

Restriction analysis of the sequence



Enzymes that do NOT cut the sequence

AatII Acc65I AccI AclI AfeI AflII AhdI AleI ApaI ApaLI AscI AseI AsiSI AvaI AvrII
BaeGI BaeI **BamHI** BanII BbsI BbvCI BciVI BclI BfaI BfuAI BglII BmtI BpmI Bpu10I BpuEI
BsaAI BsaBI BsaI BseRI BsiEI BsiWI BsmAI BsmBI BsmFI BsoBI BspMI BsrBI BsrDI BssSI
BstBI BstEII BstNI BstYI BstZ17I Bsu36I BtsI CspCI DraI DrdI EaeI EagI EciI Eco53kI
EcoNI EcoO109I **EcoRI** EcoRV FseI FspI HincII **HindIII** HpaI Hpy99I KasI KpnI MfeI MscI
NaeI NarI NcoI NgoMIV NheI NmeAIII **NotI** NsiI NspI PacI PaeR7I PciI PflFI PflMI PmeI
PmlI PpuMI PsiI PspGI PspOMI PspXI PstI PvuI PvuII RsrII **SacI** SacII **SalI** SbfI ScaI
SexAI SfcI SfiI SfoI SgrAI SmaI SmlI SnaBI SpeI SphI SspI StuI StyI SwaI TliI TspMI
Tth111I XbaI XcmI **XhoI** XmaI ZraI

Restriction-based cloning strategy by PCR if the coding sequence is known
2. design of *primers*

Multiple cloning site of the vector

BamHI *EcoRI* *SacI* *SalI* *HindIII* *NotI* *XhoI*

ATCGGAATTAATTC**GGATCCGAATTCGAGCTC**C**GTCGACAAGCTTGCGGCCGC**A**CTCGAG**CACCACCAC

Sequence of *primers* for cloning of the gene by PCR

BamHI Forward 5'-CAGGGATCC**ATG**AACAACTTTAATCTGCAC-3'

HindIII Reverse 5'-CAGAAGCTT**TTA**GCGGGCGGCTTCGTATATAC-3'

Primers contain 5' extensions that introduce restriction sites useful for cloning in the vector

Restriction-based cloning strategy by PCR if the coding sequence is known

Sequence of primers for cloning of the gene by PCR

BamHI Forward 5'-CAGGGATCCATGAACAACTTTAATCTGCAC-3'
HindIII Reverse 5'-CAGAAGCTTTTAGCGGGCGGCTTCGTATATAC-3'

PCR on genomic DNA or Reverse Transcriptase-PCR (RT-PCR) on mRNA

Digestion of the PCR product with restriction enzymes BamHI and HindIII

Ligation in the vector digested with BamHI and HindIII

Transformation of *E. coli*, selection of colonies, extraction of plasmid DNA and confirmation of the presence of the sequence of our gene

Cloning strategies if the coding sequence is NOT known

Screening of cDNA or genomic DNA libraries

hybridization of the library with a gene-specific DNA probe, identification and recovery of the plasmid/phage that contains the target sequence

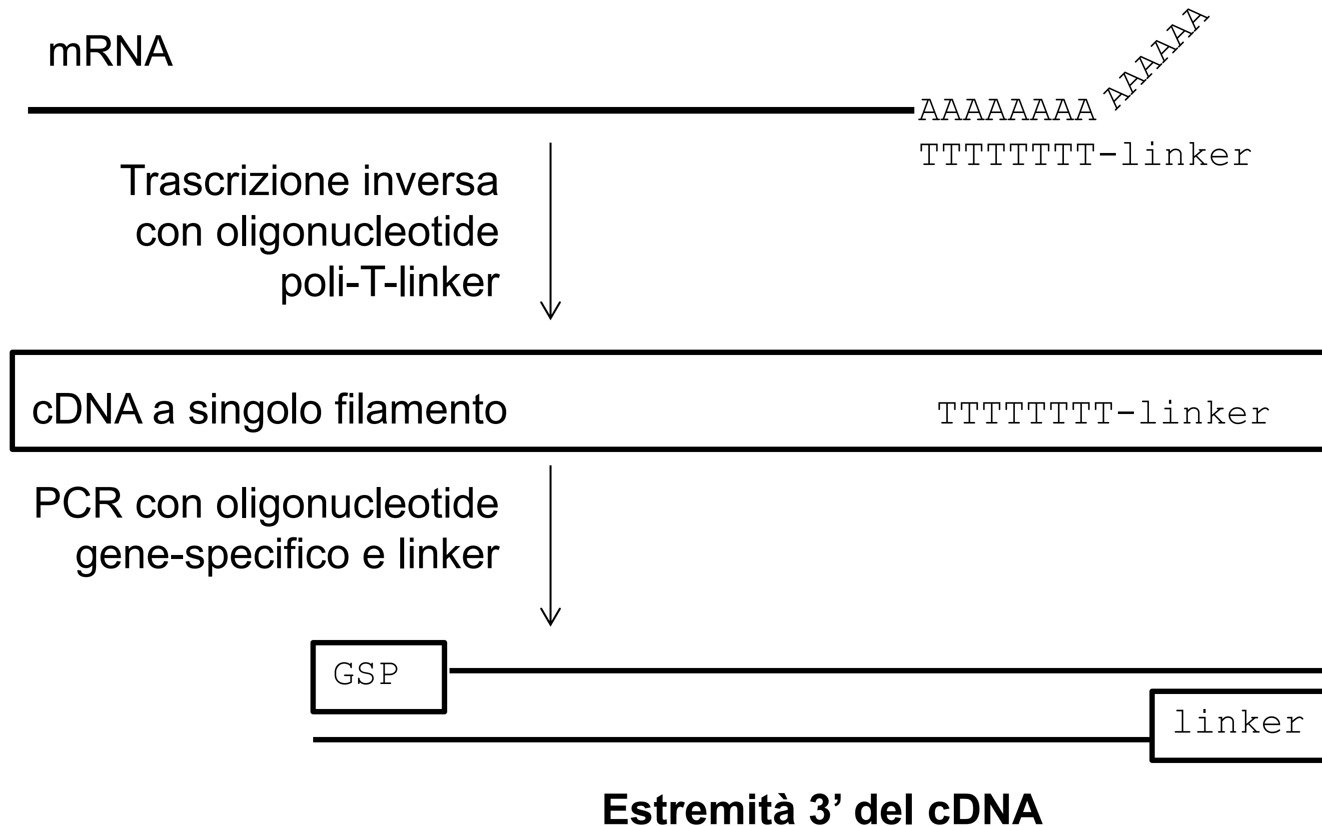
RACE (Rapid Amplification of cDNA Ends)

mRNA is transcribed to cDNA by reverse transcriptase using a generic primer (poli-T+adapter, 3' RACE) or a gene-specific primer (5' RACE). The cDNA is amplified with the adapter primer and a gene-specific primer.

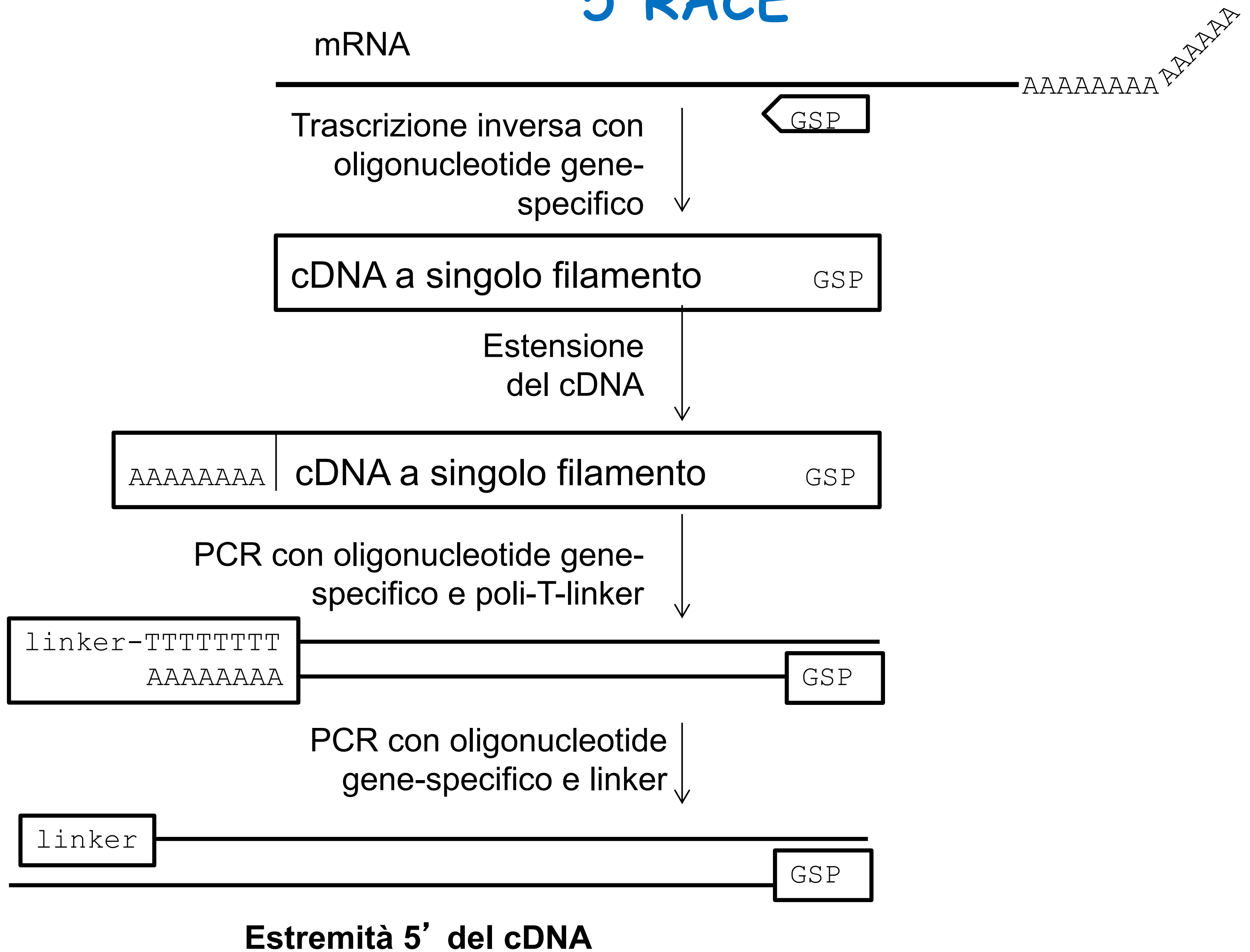
Some information on the aminoacid or nucleotide target sequence is required to design the gene-specific primer(s)

Information can be obtained directly on the protein (Edman degradation or mass spectrometry) or by bioinformatic analyses (sequence alignments and identification of conserved sequences)

3' RACE



If the gene-specific primer anneals at the starting ATG codon the full-length coding sequence can be obtained


5' RACE
mRNA
AAAAAAAA AAAAAA
GSP
Trascrizione inversa con oligonucleotide gene-specifico
cDNA a singolo filamento
GSP
Estensione del cDNA
AAAAAAAA
cDNA a singolo filamento
GSP
PCR con oligonucleotide gene-specifico e poli-T-linker
linker-TTTTTTTT
AAAAAAAA
GSP
PCR con oligonucleotide gene-specifico e linker
linker
GSP
Estremità 5' del cDNA

How to design a primer starting from the amino acid sequence?

The genetic code is degenerated!

Prima base del codone		Seconda base del codone				Terza base del codone
		T	C	A	G	
T		TTT → Phe **F**	TCT → Ser **S**	TAT → Tyr **Y**	TGT → Cys **C**	T
		TTC → Phe **F**	TCC → Ser **S**	TAC → Tyr **Y**	TGC → Cys **C**	C
		TTA → Leu **L**	TCA → Ser **S**	TAA → **STOP**	TGA → **STOP**	A
		TTG → Leu **L**	TCG → Ser **S**	TAG → **STOP**	TGG → Trp **W**	G
C		CTT → Leu **L**	CCT → Pro **P**	CAT → His **H**	CGT → Arg **R**	T
		CTC→ Leu **L**	CCC → Pro **P**	CAC → His **H**	CGC → Arg **R**	C
		CTA → Leu **L**	CCA → Pro **P**	CAA → Gln **Q**	CGA → Arg **R**	A
		CTG → Leu **L**	CCG → Pro **P**	CAG → Gln **Q**	CGG → Arg **R**	G
A		ATT → Ile **I**	ACT → Thr **T**	AAT → Asn **N**	AGT → Ser **S**	T
		ATC → Ile **I**	ACC → Thr **T**	AAC → Asn **N**	AGC → Ser **S**	C
		ATA → Ile **I**	ACA → Thr **T**	AAA → Lys **K**	AGA → Arg **R**	A
		ATG → Met **M**	ACG → Thr **T**	AAG → Lys **K**	AGG → Arg **R**	G
G		GTT → Val **V**	GCT → Ala **A**	GAT → Asp **D**	GGT → Gly **G**	T
		GTC → Val **V**	GCC → Ala **A**	GAC → Asp **D**	GGC → Gly **G**	C
		GTA → Val **V**	GCA → Ala **A**	GAA → Glu **E**	GGA → Gly **G**	A
		GTG → Val **V**	GCG → Ala **A**	GAG → Glu **E**	GGG → Gly **G**	G

How to design a primer starting from the amino acid sequence?

The genetic code is degenerated!

Avoid amino acids encoded by six codons (Leu, Ser, Arg)
Exclude the third base of the last codon
Choose the least degenerated amino acid sequence

Met-Leu-Pro-Gln-Lys-Trp-Asp-Gly-Ser-Met-Asn-Gly-Arg
1 6 4 2 2 1 2 4 6 1 2 4 6

Pro-Gln-Lys-Trp-Asp-Gly
4 2 2 1 2 4
CCN-CAA/G-AAA/G-TGG-GAT/C-GGN 18 bases degeneration 128
17 bases degeneration 32

Met-Leu-Pro-Gln-Lys-Trp
1 **6** 4 2 2 1
ATG-**TTA/G**-CCN-CAA/G-AAA/G-TGG 18 bases degeneration 32 (96)

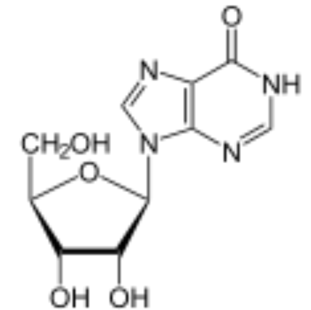
How to design a primer starting from the amino acid sequence?

Degeneration can be further reduced by:

1) Introduction of unusual bases

deoxyinosine (I) N→I

I→A	I→G
I→C	
I→T	
ok	no



es. Arg CGN

AGA/G

5'-IGN-3'

2) Exploitation of 'codon usage'

- Synonimous codons for a single amino acid are generally used with different frequencies in an organism/cell. By choosing the most used codon for each amino acid in the target sequence it is possible to reduce the total degeneration of the primer; however, also the probability of perfect annealing is reduced.

Codon usage frequencies in *Saccharomyces cerevisiae*

Saccharomyces cerevisiae [gbpln]: 11937 CDS's (5695112 codons)
fields: [triplet] [amino acid] [fraction] [frequency: per thousand] ([number])

Triplet	AA	Fraction	Freq	Number	Triplet	AA	Fraction	Freq	Number	Triplet	AA	Fraction	Freq	Number	Triplet	AA	Fraction	Freq	Number
UUU	F	0.59	26.1	(148716)	UCU	S	0.26	23.5	(134028)	UAU	Y	0.56	18.9	(107653)	UGU	C	0.63	8.0	(45462)
UUC	F	0.41	18.2	(103686)	UCC	S	0.16	14.2	(80799)	UAC	Y	0.44	14.6	(83397)	UGC	C	0.37	4.7	(26534)
UUA	L	0.28	26.7	(151850)	UCA	S	0.21	18.8	(106932)	UAA	*	0.47	1.0	(5692)	UGA	*	0.31	0.7	(3804)
UUG	L	0.28	27.0	(153619)	UCG	S	0.10	8.5	(48582)	UAG	*	0.22	0.5	(2682)	UGG	W	1.00	10.3	(58581)
CUU	L	0.13	12.1	(69114)	CCU	P	0.31	13.6	(77417)	CAU	H	0.64	13.7	(78282)	CGU	R	0.15	6.5	(36777)
CUC	L	0.06	5.4	(30538)	CCC	P	0.15	6.8	(38522)	CAC	H	0.36	7.7	(44112)	CGC	R	0.06	2.6	(14638)
CUA	L	0.14	13.3	(75972)	CCA	P	0.42	18.2	(103680)	CAA	Q	0.69	27.5	(156586)	CGA	R	0.07	3.0	(17099)
CUG	L	0.11	10.4	(59114)	CCG	P	0.12	5.3	(29915)	CAG	Q	0.31	12.1	(69006)	CGG	R	0.04	1.7	(9848)
AUU	I	0.47	30.4	(173137)	ACU	T	0.35	20.3	(115551)	AAU	N	0.59	36.3	(206516)	AGU	S	0.16	14.2	(80728)
AUC	I	0.26	17.0	(97055)	ACC	T	0.21	12.6	(71608)	AAC	N	0.41	24.9	(141737)	AGC	S	0.11	9.7	(54958)
AUA	I	0.27	17.8	(101628)	ACA	T	0.30	17.8	(101336)	AAA	K	0.58	42.2	(240515)	AGA	R	0.48	21.3	(121292)
AUG	M	1.00	20.9	(119008)	ACG	T	0.14	7.9	(45120)	AAG	K	0.42	30.7	(174597)	AGG	R	0.21	9.2	(52441)
GUU	V	0.39	22.0	(125255)	GCU	A	0.38	21.1	(120208)	GAU	D	0.65	37.8	(215249)	GGU	G	0.47	24.0	(136434)
GUC	V	0.21	11.6	(65876)	GCC	A	0.22	12.5	(71233)	GAC	D	0.35	20.2	(115047)	GGC	G	0.19	9.7	(55103)
GUA	V	0.21	11.8	(67408)	GCA	A	0.29	16.1	(91967)	GAA	E	0.71	45.8	(261115)	GGA	G	0.22	10.9	(62176)
GUG	V	0.19	10.6	(60537)	GCG	A	0.11	6.1	(34942)	GAG	E	0.29	19.1	(108666)	GGG	G	0.12	6.0	(34032)

Coding GC 39.64% 1st letter GC 44.53% 2nd letter GC 36.55% 3rd letter GC 37.85%
Genetic code 1: Standard

Designing the primer for cloning of Ftr1 from the yeast *Pichia pastoris*

Sequence alignment of Ftr1 from different yeasts to identify conserved regions

```
FTR1_YEAST      MPNKVFNVAVFFVVFRECLEAVIVISVLLSFLKQAIGEHDR-----ALYRKLRIQVWVGV 55
FTR2_CANAL      -MVDVFNVQVFFVVFREALEAVIVVSVLLAFVKQSMGSSANG---PELKKKLYRQIWLGA 56
FTR1_CANAL      -MVDVFNVQIFFIVFRESLEAIIVVSVLLAFVKQSMGGSSD----PQLKKRLYRQIWLGA 55
YAK7_SCHPO      MAKDVFSVAIFFIVLRETLEASIIVSVLMSFISQTLMDKDGNVTDPKLKRKFMLQVWIGS 60
                   .**.* :**:*:** *** *::***::*:.*::           * :::  *:*:*

FTR1_YEAST      LLGFIICLAIGAGFIGAYYSLQKDIFG-SAEDLWEGIFCMIATIMISMMGIPMLRMNKMQ 114
FTR2_CANAL      GLGVLICIIIGGAFIGTFYGLGKDIWG-KSEDLWEGIFCIIATVLITAMGIPMLRINKMK 115
FTR1_CANAL      GLGVLVCLY-GVLSIGASYGLGKDIFGVISEDLWEGIFCIIATVLITAMGIPMLRINKMK 114
YAK7_SCHPO      FTALFICLAIGGGFIGAFYALDKDIWS-GSEEIWEGVFSLIAVVLITVMGFAMLRVSHLQ 119
                  ..::*:  *   **: *.* ***:.  :*::***:*.:**.::*: **:.***:.:::

FTR1_YEAST      SKWRVKIARSLVEIPHRKRDYFKIGFLSRRYAMFLLPFITVLREGLEAVVFVAGAGITTQ 174
FTR2_CANAL      EKWRVKLAQALIKSPENKKNRFKLGYLGKKYALFILPFITCLREGLEAVVFVGGVGITSP 175
FTR1_CANAL      EKWRVKLAQALIKSPTNKKDRFKLGYLGKKYALFILPFLQVLREGLEAVVFVGGVGLNSP 174
YAK7_SCHPO      EKWRKKLMKSIAN--RKAKG---ISNWGKKYSMFLLPFFTVLREGLEVVVFVGGVGLETP 174
                .*** *: ::: :   . :.   :.  .::*::*:***:  ******.****.*.*: :

FTR1_YEAST      GSHASAYPLPVVVGLICGGLVGYLLYYGASKSSLQIFLILSTSILYLISAGLFSRGAWYF 234
FTR2_CANAL      ---ASSFPIPVIVGIICGLAVGALLYYFGSNMSMQIFLIISTCILYLIAAGLFSRGVWFF 232
FTR1_CANAL      ---ATSFPIPVIVGLIAGIVVGALLYYFGSSMSMQIFLIISTCILYLIAAGLFSRGIWYF 231
YAK7_SCHPO      ---ATAFPLPVICGLIVGCLIGYFIYRGGNVMNLQWFLIASTCILYLISAGLMSKATFYF 231
                   *:::*:**: *:* *  :* ::*  ..  .:* *** **.*****:***:*:. ::*

FTR1_YEAST      ENYRFNLASGGDASEGGDGNGSYNIRKAVYHVNCCNPELDN--GWDIFNALLGWQNTGYL 292
FTR2_CANAL      ESYQYNKKTGGDAAENGSGPGTYDISKSVWHVNCCNPLTDN--GWDIFNAILGWQNSATY 290
FTR1_CANAL      ETNTYNKKTGGDASENGSGPGTYDISKSVWHVNCRNPETDN--GWDIFNAILGWQNSATY 289
YAK7_SCHPO      EMNKWNHQTGGDAGELGDGPGSYPFKSAVWHVNYGNPEMNSNGGYMIFNAILGWNNTGTY 291
                *   :*  :****.* *.* *:* : .:*:***  **  :.  *: ****:***:*:.

FTR1_YEAST      SSMLCYNIYWLVLIIVLSLMIFEERRGHLPFTKNLQLKHLNPGYWIKNKKKQELTEEQKR 352
FTR2_CANAL      GSVISYNVYWIFIISVLLLMVYEEKHGHLPLTKNLTLVQLNPMYHIKGKKKLELNKAEKE 350
FTR1_CANAL      GSVISYNIYWLFIICVLLLMVYEEKHGHLPFTKNLTLVQLNPMYHIKGKKKLELNKAEKD 349
YAK7_SCHPO      GSILSYIIYWLFVAFIMFLMWYKERRAARLLIAKLGDKVVDLEAASSHTPVQSSSSEDEF 351
                .*::.* :**:.:  :: ** ::*::.   :  :*    ::     . .   . .. ::

FTR1_YEAST      QLFAKMENINFNEDGEINVQENYELPEQTTSHSSSQNVATDKEVLHVKADSL 404
FTR2_CANAL      ELFTKLQKQKFGDVQEIDETSSNKLVETQENK-------------------- 382
FTR1_CANAL      ELFTKLQQQNFGQAAEVDETSSNKWMDSQENS-------------------- 381
YAK7_SCHPO      KINSPTDDKGDKAIDIVTEVRESSSPVEEHKDDKTVDVINEIRESH------ 397
                :: :  :.        :    . .      .
```

Designing the primer for cloning of Ftr1 from the yeast *Pichia pastoris*

Sequence: Leu-Arg-Glu-Gly-Leu-Glu-Ala-Val-Val

- conserved in all Ftr1 proteins
- the two Glu residues are essential for Ftr1 function

Sequence: Leu-Arg-Glu-Gly-Leu-Glu-Ala-Val-Val

Degeneration: 6 6 2 4 6 2 4 4 4

Codon usage					
	Leu	TTA 0.16	TTG 0.33		
		CTT 0.17	CTC 0.08	CTA 0.12	CTG 0.16
	Arg	AGA 0.47	AGG 0.16		
		CGT 0.16	CGC 0.05	CGA 0.11	CGG 0.05
	Glu	GAA 0.58	GAG 0.42		
	Gly	GGT 0.43	GGC 0.14	GGA 0.32	GGG 0.10
	Ala	GCT 0.45	GCC 0.25	GCA 0.24	GCG 0.06
	Val	GTT 0.42	GTC 0.23	GTA 0.16	GTG 0.20

Primer: TT(AG) AGA GA(AG) GG(AT) (CT)T(AG) GA(AG) GC(ACT) GT(AGCT) GT

Degeneration 768 instead of 221184

Cloning of *Pichia pastoris* Ftr1 cDNA

3' RACE with Ftr1 primer and adaptor primer (GACTCGAGTCGACATCG) on *Pichia pastoris* mRNA retrotranscribed with poly-dT-adaptor

Result obtained: PCR product of about 900 bp

TTAAGAGAGGGTTTAGAGGCTGTAGTGTTCATAGGTGGTGTTGGCCTGGGATCTCCAGCTACATCTTTTCCCATCCCTG
TGATTGTTGGTCTTATTGCTGGTATTTCTGTCGGTGTGCTTTTGTACTATTCAGGATCAACTCTTTCTTTGCAAGTTTT
CTTATGTATATCGACGGCTATTTTGTACTTGATTGCTGCTGGGTTGTTTTCTAGGGCTATCTGGTTCTTTGAAACCTAC
AAATACAACCAAAAGACTGGAGGAGATGCTTCTGAAAATGGTTCGGGACCCGGTACTTATGACATTAAGACTTCTGTTT
GGCATGTGAACTGCTGTAACCCTGAAACTGATAATGGTTGGGATATTTTTAACGCGTTGTTGGGATGGCAAAACTCTGC
CACTTATGGCTCAGTCATAGGCTACAACATATACTGGCTGGCGGTAATGATCACTTTGTATTTACTCTGGTTCGAAGAA
AAGAATAACCACTTGCCATTTATGAAAAATCTGAAGTTGAGACAACTGAACCCATTATACTGGATGAAGGGTAAAAACA
AGAAGGAGGTCTCCAAGGAAGATCAGGAAAAATTATTTGAGCAATTGAAGAGTAAGGAGTTTGCCAATAAACTTGCGGA
GGAA**TAA**TTAGATTGACCTAGCTGAATCTTGTTAGCAACTGTATGCACACACACACACACACACACACACACATACAAG
GGCTAACAGAGGATGATACGTTTTTCAGTCACAAACACATACACATGCACATGTACCCTAACAAGAACAAGACATTTTG
CTTGAATTAGAACTCAGGACTGCAAAAGGTTTACCGCATACTACGTACGCACATACACCATCTAATACACACGTTTTGT
ATAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAATCGATGTCGACTCGAGTC

Primer:	TT(AG)	AGA	GA(AG)	GG(AT)	(CT)T(AG)	GA(AG)	GC(ACT)	GT(AGCT)	GT
PCR product:	TT A	AGA	GA G	GG T	T T A	GA G	GC T	GT A	GT

Alignment of the deduced aminoacid sequence of *Pichia pastoris* Ftr1 with other Ftr1 permeases

```
PpaFTR1         LREGLEAVVFIGGVGLGSP---ATSFPIPVIVGLIAGISVGVLLYYSGSTLSLQVFLCI
CaFTR2          LREGLEAVVFVGGVGITSP---ASSFPIPVIVGIICGLAVGALLYYFGSNMSMQIFLII 212
CaFTR1          LREGLEAVVFVGGVGLNSP---ATSFPIPVIVGLIAGIVVGALLYYFGSSMSMQIFLII 211
FTR1_YEAST      LREGLEAVVFVAGAGITTQGSHASAYPLPVVVGLICGGLVGYLLYYGASKSSLQIFLIL 214
YAK7_SCHPO      LREGLEVVVFVGGVGLETP---ATAFPLPVICGLIVGCLIGYFIYRGGNVMNLQWFLIA 211
                ******.***:.* *:      *:::*:.:: *:* *  .* ::*  ..  .:* **
PpaFTR1         STAILYLIAAGLFSRAIWFFETYKYNQKTGGDASENGSGPGTYDIKTSVWHVNCCNPET
CaFTR2          STCILYLIAAGLFSRGVWFFESYQYNKKTGGDAAENGSGPGTYDISKSVWHVNCCNPLT 271
CaFTR1          STCILYLIAAGLFSRGIWYFETNTYNKKTGGDASENGSGPGTYDISKSVWHVNCRNPET 270
FTR1_YEAST      STSILYLISAGLFSRGAWYFENYRFNLASGGDASEGGDGNGSYNIRKAVYHVNCCNPEL 273
YAK7_SCHPO      STCILYLISAGLMSKATFYFEMNKWNHQTGGDAGELGDGPGSYPFKSAVWHVNYGNPEM 270
                **.*****:***:*:. ::**   :*  .****.* *.* *:* :  :*:***  **
PpaFTR1         DN--GWDIFNALLGWQNSATYGSVIGYNIYWLAVMITLYLLWFEEKNNHLPFMKNLKLRQ
CaFTR2          DN--GWDIFNAILGWQNSATYGSVISYNVYWIFIISVLLLMVYEEKHGHLPLTKNLTLVQ 329
CaFTR1          DN--GWDIFNAILGWQNSATYGSVISYNIYWLFIICVLLLMVYEEKHGHLPFTKNLTLVQ 328
FTR1_YEAST      DN--GWDIFNALLGWQNTGYLSSMLCYNIYWLVLIIVLSLMIFEERRGHLPFTKNLQLKH 331
YAK7_SCHPO      NSNGGYMIFNAILGWNNTGTYGSILSYIIYWLFVAFIMFLMWYKERRAARLLIAKLGDKV 330
                :.  *: ****:*** *:.  .*:: *  **: :   : *:   :*:         *
PpaFTR1         LNPLYWMKGKNKKEVSKEDQEKLFEQLKSKEF----------ANKLAEE-----------
CaFTR2          LNPMYHIKGKKKLELNKAEKEELFTKLQKQKFGDVQEIDETSSNKLVETQENK------- 382
CaFTR1          LNPMYHIKGKKKLELNKAEKDELFTKLQQQNFGQAAEVDETSSNKWMDSQENS------- 381
FTR1_YEAST      LNPGYWIKNKKKQELTEEQKRQLFAKMENINFNEDGEINVQENYELPEQTTSHSSSQNVA 391
YAK7_SCHPO      VDLEAASSHTPVQSSSSEDEFKINSPTDDKGDKAIDIVTEVRESSSPVEEHKDDKTVDVI 390
                       .        . .  ::    ..               .
FTR1_YEAST      TDKEVLHVKADSL 404
YAK7_SCHPO      NEIRESH------ 397
```

To complete the coding sequence of *Pichia pastoris* Ftr1 this fragment has been used to screen a cDNA library

Modification of the coding sequence. PCR-based mutagenesis strategies

Site-specific mutagenesis

to evaluate the structural/functional role of specific amino acid residues: a low number of variants is produced on the basis of predefined hypotheses

Random mutagenesis (in vitro directed evolution)

to modify the properties of a protein (thermostability, substrate specificity): a very large number of variants is produced that must be analyzed to identify the improved ones ➔ an efficient screening method is required

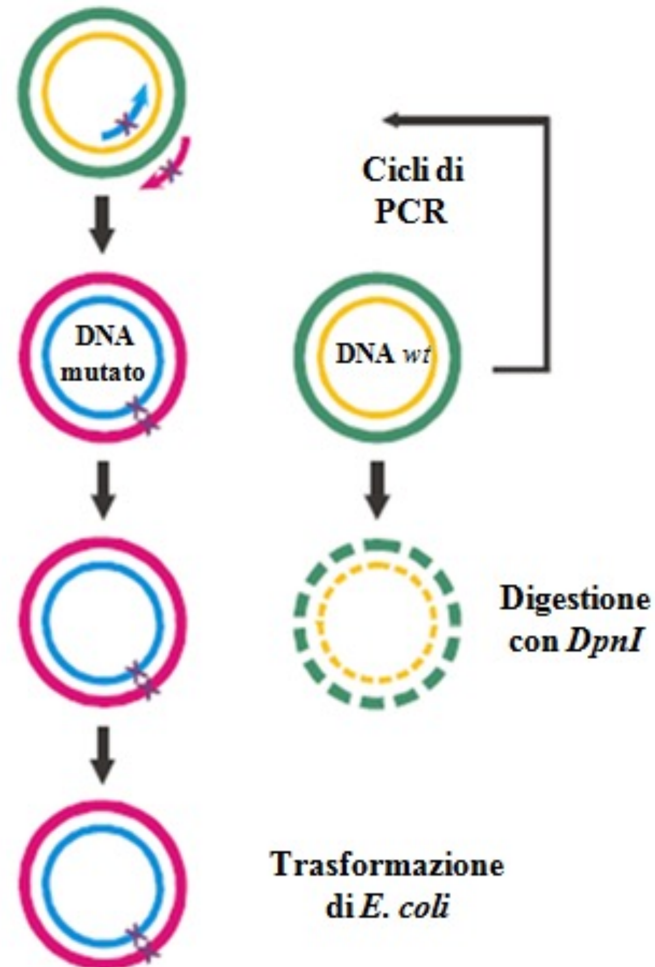
error-prone PCR

DNA shuffling

saturation mutagenesis

Site-specific mutagenesis QuikChange:

1. PCR to amplify the whole plasmid with two complementary mutagenic primers
2. Digestion with DpnI to degrade the wild type methylated template
3. Transformation of E. coli
4. Verification by DNA sequencing



Site-specific mutagenesis: simultaneous introduction of multiple mutations

Overlap extension PCR

Mutagenic primers are employed with flanking primers to amplify overlapping regions of the coding sequence. These PCR products are used as template in a second round of PCR that reassembles the full coding sequence

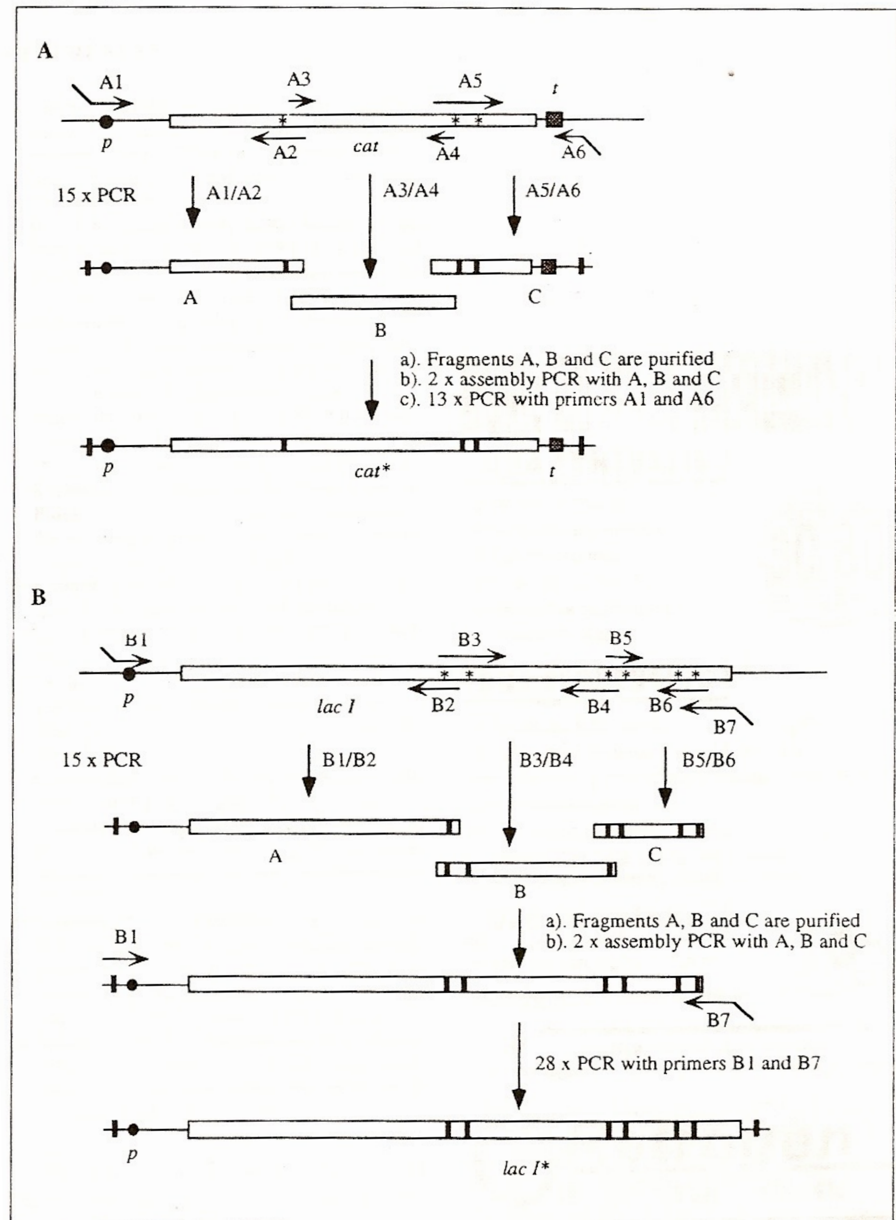
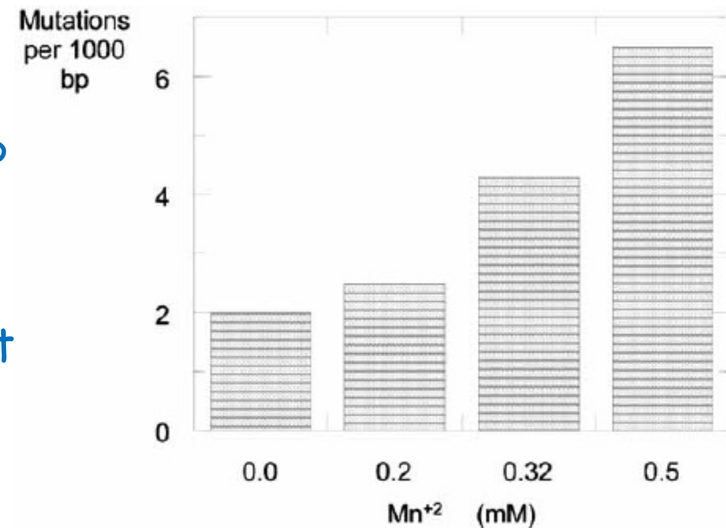


Figure 1. Outline of the simultaneous multisite mutagenesis procedure. The *cat* (A) and *lac*I (B) gene segments were amplified, purified and assembled to produce complete genes with introduced mutations. *P*, promoter; *t*, terminator; *, restriction site that is to be changed; and solid bar, introduced mutation.

In vitro directed evolution: error-prone PCR

- PCR in conditions that increase the error frequency of Taq polymerase:
  - Addition of Mn^{2+} (reduces specificity of base pairing),
  - Unbalanced concentration of deoxyribonucleotides (if one of the dNTP becomes limiting, it could be replaced by a different dNTP),
  - Increase of Mg^{2+} (may stabilize uncorrect base pairing),
  - Increase Taq concentration (increases the probability of extension from uncorrect primer annealing)



In vitro directed evolution: DNA shuffling

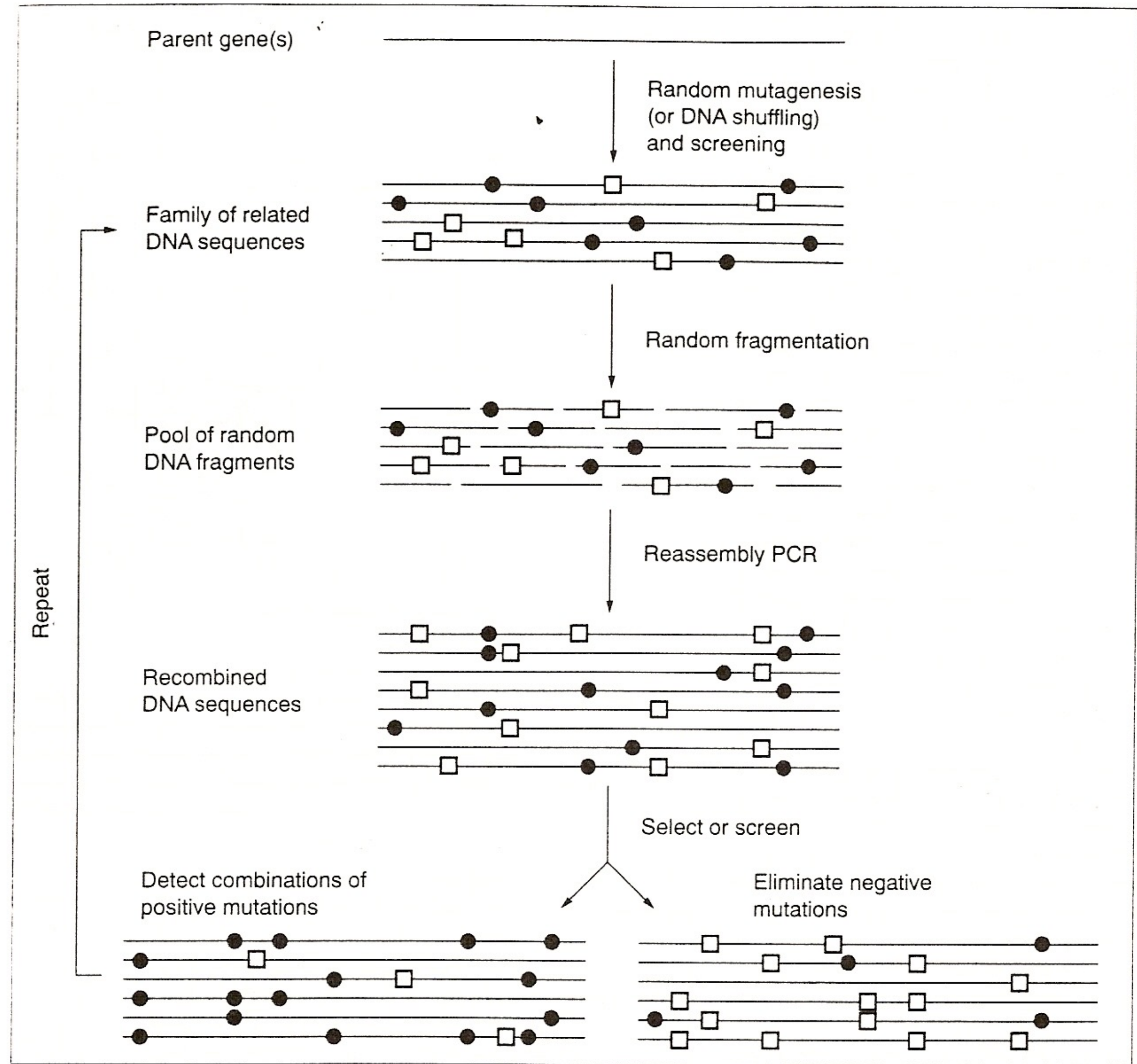


Figure 2

In vitro recombination by DNA shuffling. Random fragments of a single parent gene or pool of genes containing different mutations are reassembled in a PCR reaction, during which they are allowed to act as primers for one another. The fragments elongate, producing a family of DNA sequences containing different combinations of the mutations contained in the parent pool. Beneficial mutations are accumulated and deleterious mutations eliminated during recursive cycles of shuffling and screening or selection (adapted from Ref. 11).

In vitro directed evolution
Screening methods

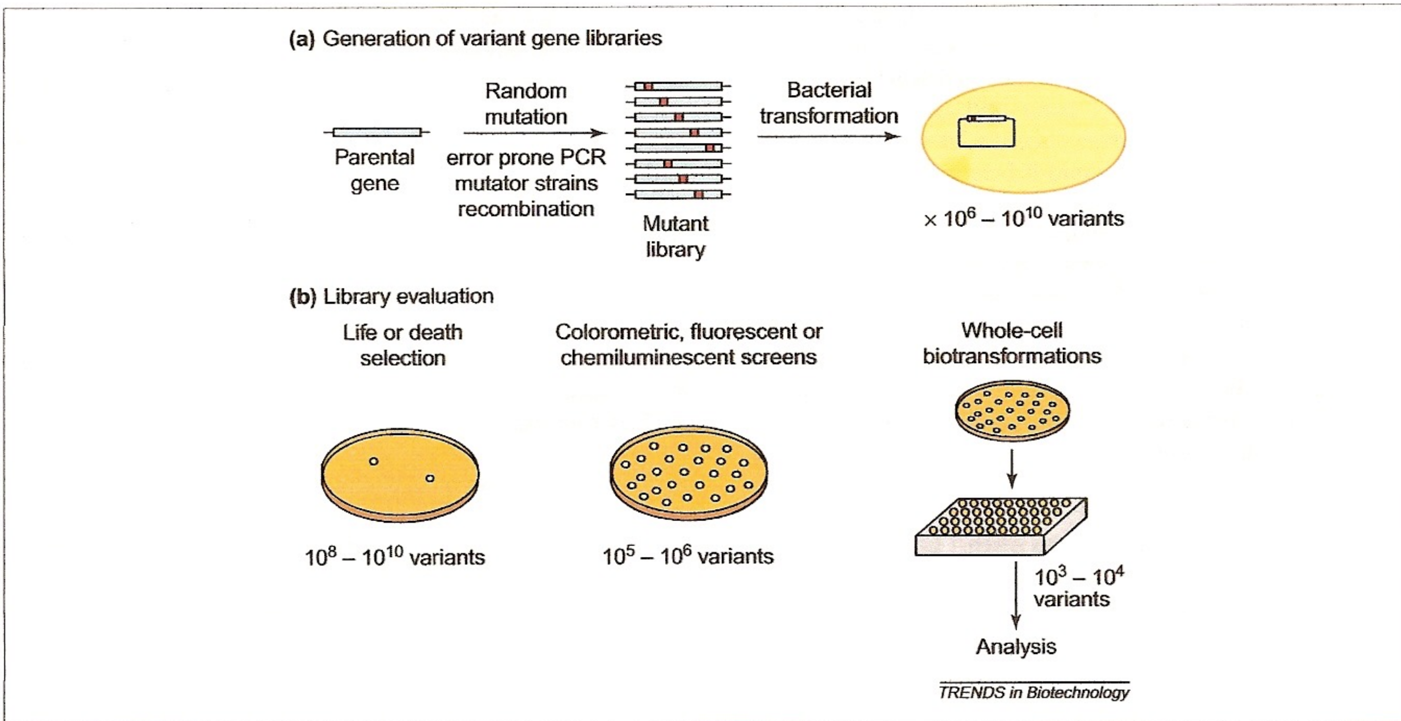
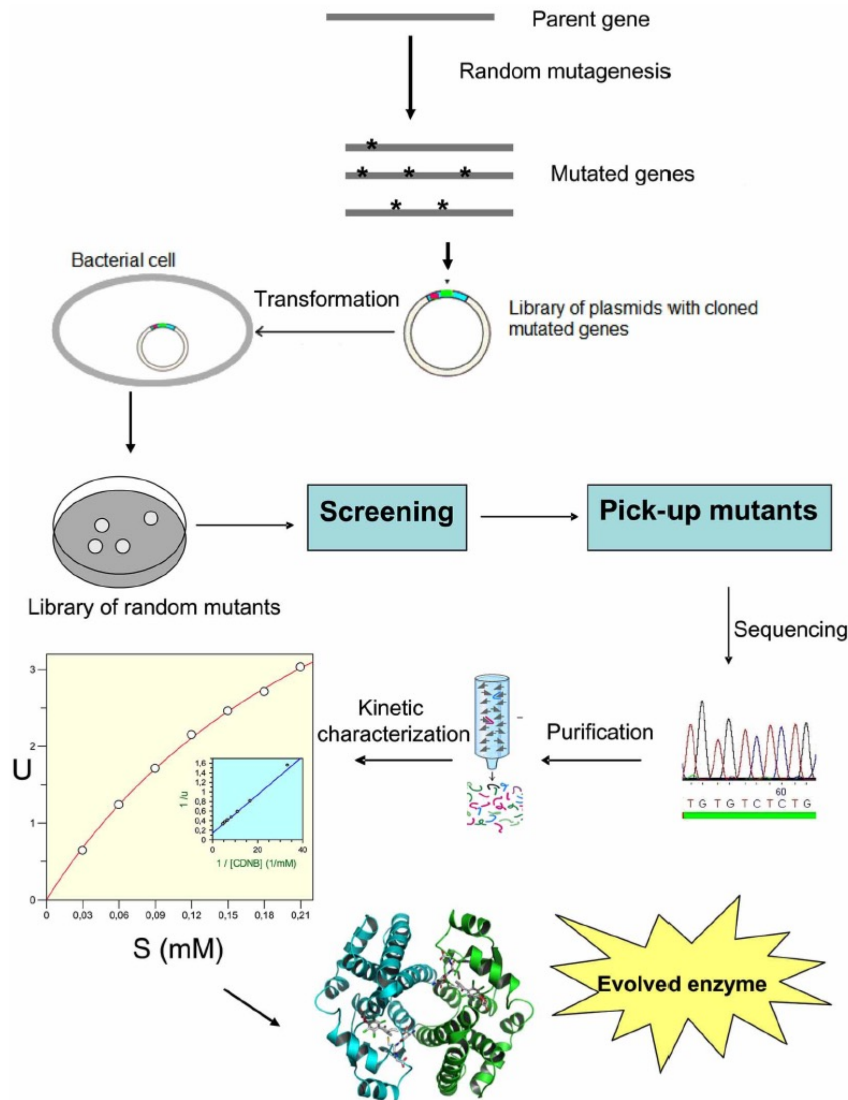
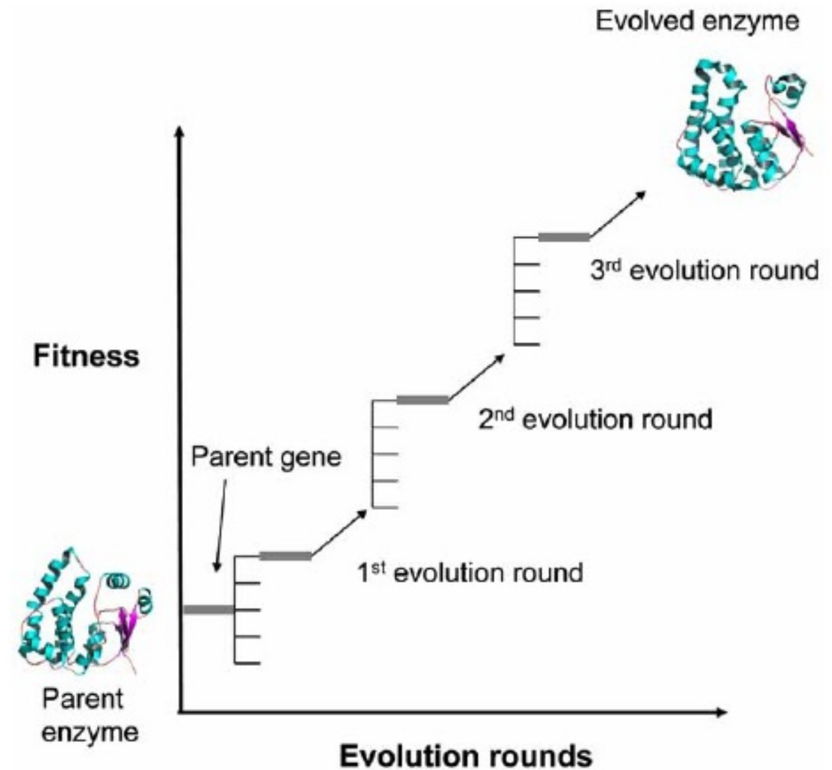


Figure 1. Strategies for the directed evolution of enzymes involving (a) generation of libraries of variant genes and (b) high-throughput screening of the libraries using different analytical methods.

In vitro directed evolution

Strategy

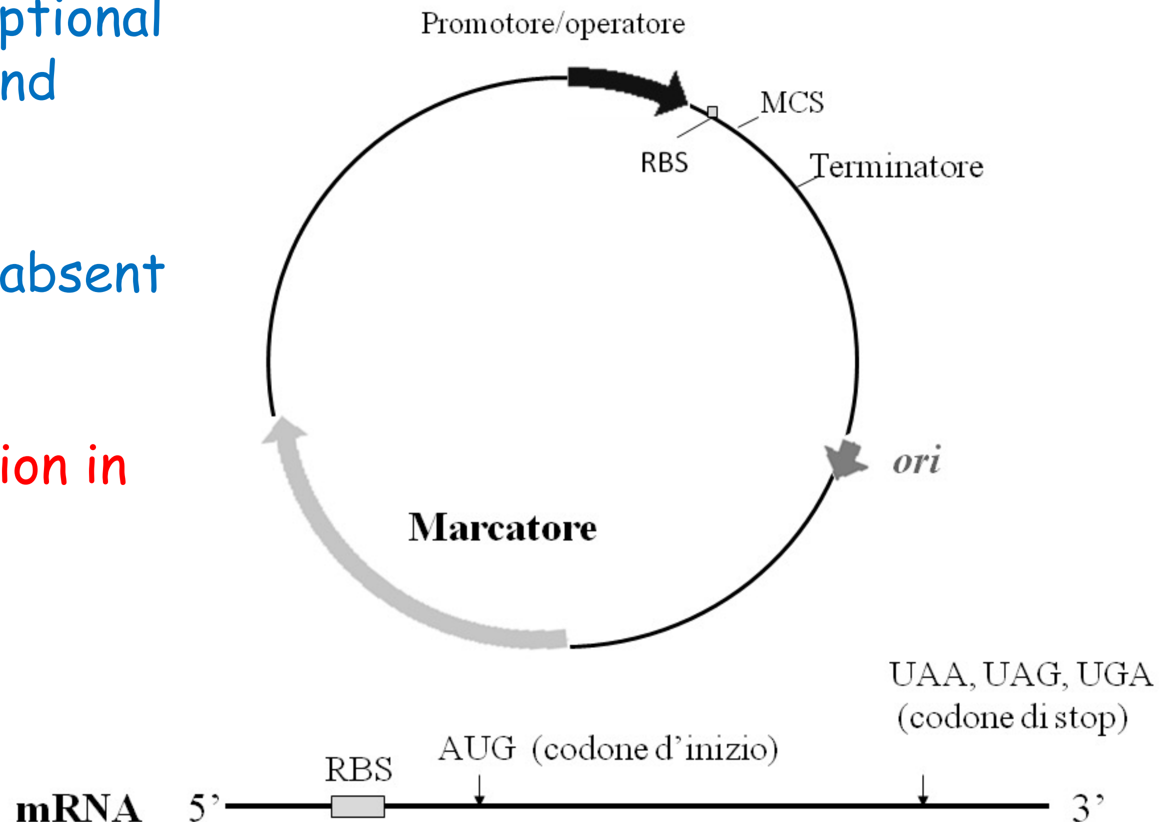




Expression vectors

Host-specific elements:
- Elements for transcriptional regulation (promoter and terminator)
- Selection Marker
- Origin of replication (absent in integrative vectors)

Elements for propagation in E. coli:
- Selection Marker
- Origin of replication

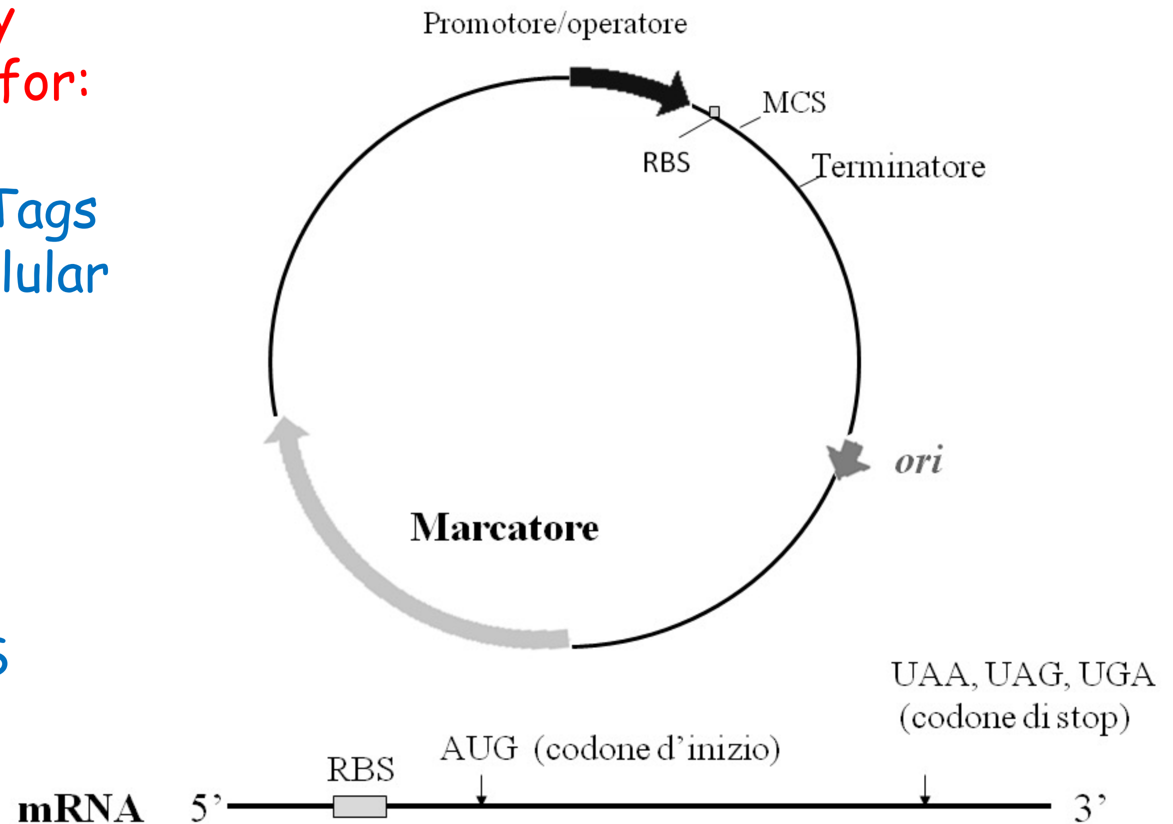


Expression vectors

Elements for modification of the coding sequence by addition of sequences for:

- Peptide or protein Tags
- Secretion or subcellular localization signals

FUSION PROTEINS



Host organism

- Prokaryotes
  - Bacteria (*Escherichia coli, Bacillus sp.*)
- Eukaryotes
  - Yeast (*Saccharomyces cerevisiae, Pichia pastoris, Kluyveromyces lactis*)
  - Mammalian cells (CHO, BHK, HEK-293)
  - Insect cells (Sf9)
  - Transgenic plants and animals

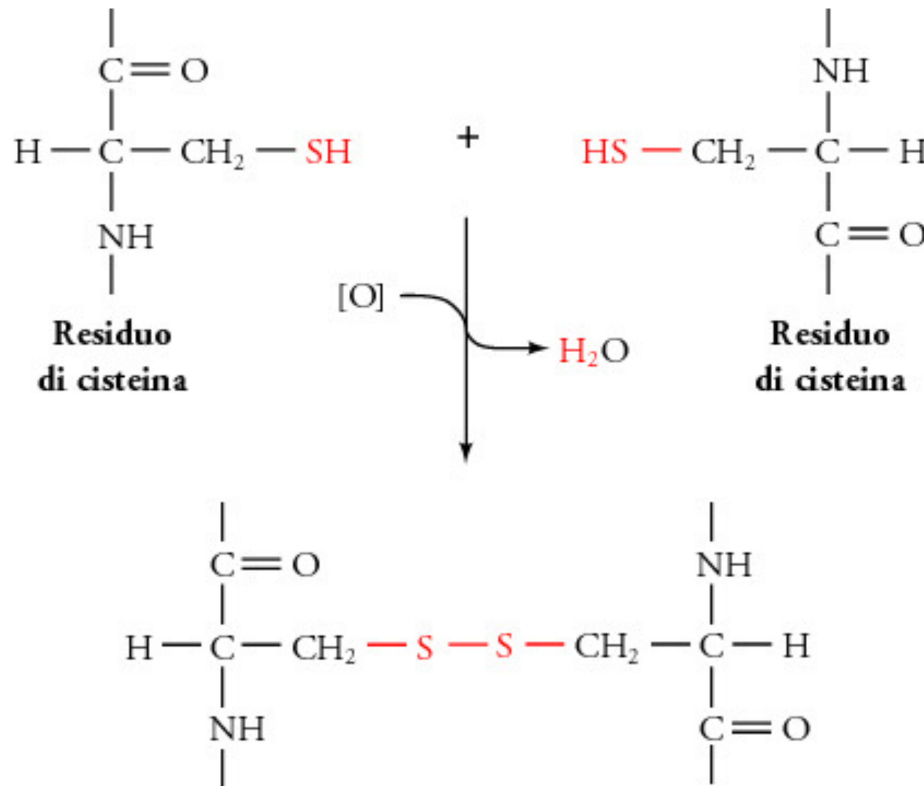
Choice of expression system

- Features of the protein, including any post-translational modifications necessary for folding and biological activity
- Production time
- Production cost
- Regulatory issues (FDA, EMA)
- Royalties

The main post-translational modifications of proteins

- Disulfide bridges
- Glycosylation
- Proteolytic processing
- γ-carboxylation, β-hydroxylation
- Sulfation
- Phosphorylation
- Acetylation, methylation, ADP-ribosylation, amidation

Cysteine and disulfide bridges



The formation of disulfide bridges takes place in the periplasm in bacteria and in the endoplasmic reticulum in eukaryotes

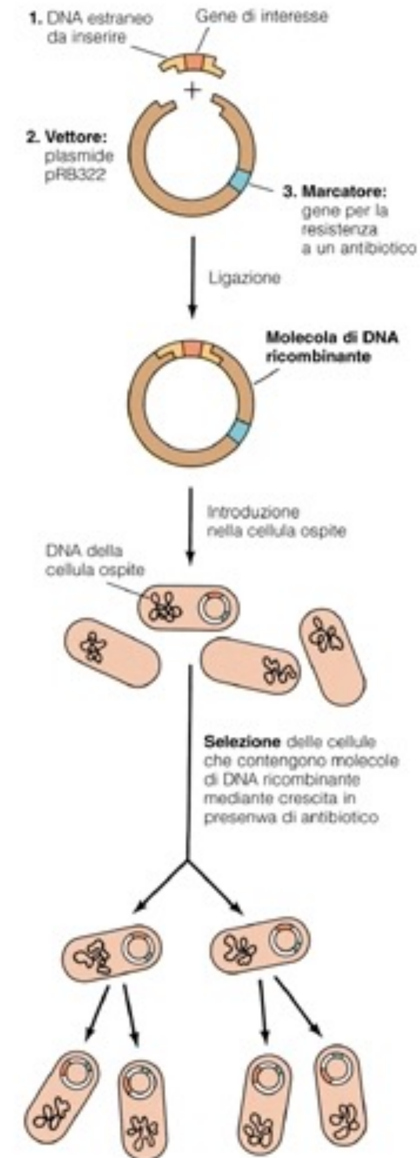
It requires oxidizing conditions

Glycosylation of proteins. Role of the glycocomponent

- Glycosylation may assist folding of the polypeptide chain by affecting secondary structure
- Oligosaccharide chains may contribute to protein stability by increasing solubility, shielding hydrophobic regions, protecting from proteases or favoring intra-chain interactions
- Glycosylation may affect trafficking/sorting of a protein to its final destination
- Oligosaccharide chains may participate to recognition/binding to ligands or receptors or to the biological activity of a protein
- Sialic acid that is found at the terminus of oligosaccharide chains increases the half-life of proteins in plasma
- Oligosaccharide chains may contribute to immunogenicity of heterologous proteins

Escherichia coli

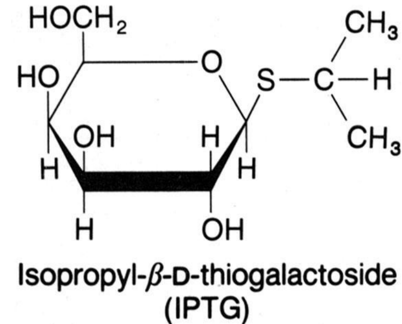
- Rapid growth (ca 30 min doubling time)
- Low cost platform system
- Detailed understanding of transcriptional and translational molecular mechanisms
- Availability of mutant strains
- Easy transformation ($CaCl_2$, electroporation)
- Elevated expression levels from multicopy plasmids with strong inducible (lac, tac) and constitutive (T7) promoters
- Antibiotic resistance selection
- Preferential intracellular expression
- Post-translational modifications?



Ceppo *E. coli*	**Caratteristiche**
BL21(DE3)	Esprime la RNA polimerasi del fago T7 a seguito di induzione con IPTG. È difettivo delle proteasi *lon* e *ompT*
BL21(DE3)-pLysS	Come le BL21(DE3) esprime la T7 RNA polimerasi. Esprime dal plasmide pLysS il lisozima T7. Adatto per l'espressione di proteine tossiche
Origami	Ha mutazioni sui geni *trxB* e *gor*. Facilita la formazione dei ponti disolfuro citoplasmatici
Origami-pLysS	Come il precedente con l'aggiunta del sistema pLysS per il controllo trascrizionale
Rosetta	Sovraesprime i tRNA per i codoni rari. Facilita l'espressione di proteine eucariotiche
Rosetta-gami-pLysS	Come il precedente con l'aggiunta del sistema pLysS per il controllo trascrizionale
BL21-codon plus	Sovraesprime i geni dei tRNA più rari per arginina (AGA/AGG), isoleucina (AUA) e leucina (CUA)
C41 e C43 (DE3)	Adatti per l'espressione di proteine tossiche di ogni origine

E. coli: strong and inducible promoters

- Lac promoter (inducible with IPTG, a galactose analogue)



Isopropyl-β-D-thiogalactoside (IPTG)

- Trp promoter (inducible in the absence of tryptophan)
- Tac promoter: regions -35 form trp promoter
  -10 from lac promoter
- Phage T7 constitutive promoter (the strain MUST contain T7 RNA polymerase)

Yeast

Table 1. Overview of the studied yeast genomes and proteomes

Species	Strain	Genome size (Mbp)*	Number of predicted proteins*	Predicted secretome size†	Interest	Comments
Saccharomyces cerevisiae	S288C	12.16	5907	156	Model yeast, protein production	Whole genome duplication
Candida glabrata	CBS 138	12.34	5213	121	Opportunistic pathogen	Whole genome duplication
Kluyveromyces lactis	NRRL Y-1140	10.73	5085	113	Model yeast, protein production	
Candida albicans	WO-1	14.47	5752	449‡	Pathogen	Most strains are diploid; for clarity data of a haploid strain was analyzed
Hansenula polymorpha	DL-1	8.86	4156	–	Protein production, model yeast	
Pichia pastoris	GS115	9.44	5033	105	Protein production, model yeast	
Yarrowia lipolytica	CLIB122	20.55	6472	299	Protein production	
Schizosaccharomyces pombe	972 h-	12.59	5020	112	Model yeast, protein production	Fission yeast

*Data taken from http://www.ncbi.nlm.nih.gov/genome.

†Data taken from Lum & Min (2011). The secretomes defined in that study include all proteins predicted to have a secretion signal peptide with a subcellular location predicted as extracellular, but not having a transmembrane domain or an ER targeting signal, as well as all manually curated secreted proteins. Data for *H. polymorpha* are not available.

‡Data derived from another, diploid strain of *C. albicans*.

Saccharomyces cerevisiae

- Rapid growth (ca 90 min doubling time)
- Low cost platform system
- Detailed understanding of transcriptional and translational molecular mechanisms
- Availability of mutant strains (auxotrophies, proteases)
- Easy transformation ($LiCl_2$, electroporation)
- Episomal or integrative plasmids with strong inducible (GAL1, PHO5) and constitutive (GAP, TPI, PGK) promoters
- Auxotrophic complementation selection
- Intracellular or extracellular expression (α-factor signal sequence)
- Post-translational modifications?

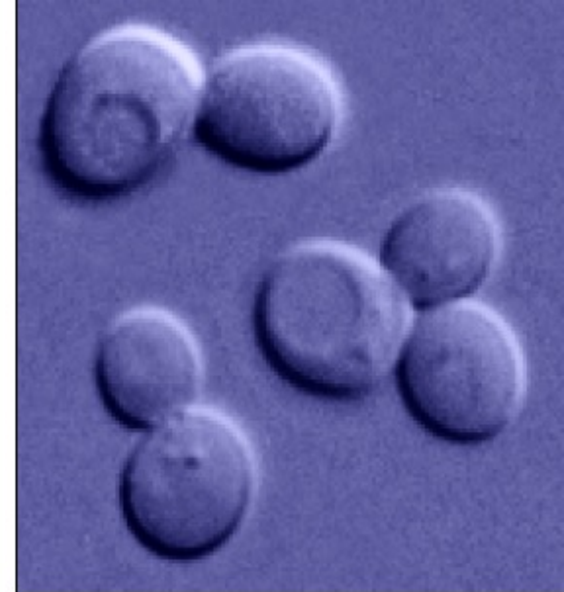
S. cerevisiae si trasforma con un frequenza di almeno di 3 ordini di grandezza inferiore a quella di *E. coli*

può mantenere plasmidi che si replicano ma il numero di copie è molto inferiore a quello che si può ottenere in *E. coli* (max. 50)

Le procedure di clonaggio possono essere difficoltose, quindi *E. coli* viene utilizzato come intermedio nella produzione dei plasmidi ricombinanti, attraverso l'uso di vettori shuttle

- I plasmidi vengono assemblati *in vitro*
- trasformati, manipolati, modificati e propagati in *E. coli*
- e poi trasformati in lievito

MARCATORI DI SELEZIONE



I lieviti sono naturalmente resistenti agli antibiotici

i geni di resistenza non sono marcatori idonei
per la selezione dei trasformanti

In lievito si usano marcatori genici
basati su auxotrofie

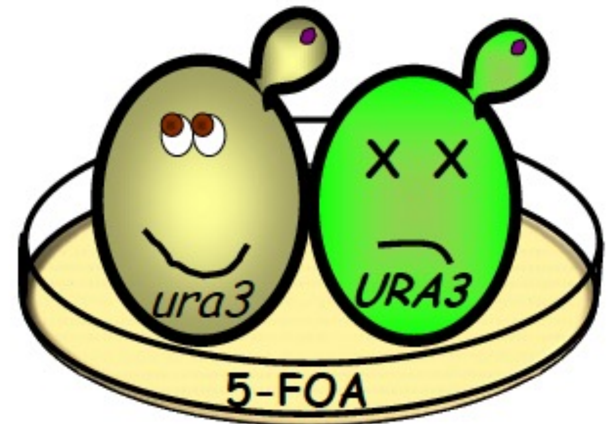
HIS3, URA3, TRP1, LEU2, LYS2, ADE2

Alcuni di questi marcatori (URA3, LYS2)
hanno il vantaggio di poter essere usati per

SELEZIONI POSITIVE	basate sulla complementazione auxotrofica delle corrispondenti mutazioni
SELEZIONI NEGATIVE	Basate su inibitori specifici che prevengono la crescita dei ceppi prototrofici ma non quella dei mutanti auxotrofi

SELEZIONE NEGATIVA URA3

URA3 codifica la orotidina-5'fosfato decarbossilasi, un enzima richiesto per la biosintesi dell'uracile



I mutanti *ura3-* (o *ura5-*) possono essere selezionati su terreni contenenti 5-FOA (acido 5-fluoroorotico)

Il prodotto di URA3 converte il 5-FOA in un prodotto tossico che uccide le cellule wild type che lo esprimono

5-FOA è molto discriminante e può essere usato

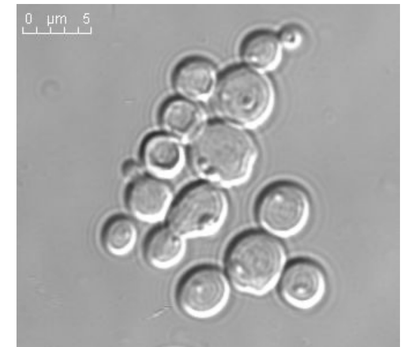
- per produrre marcatori *ura3* per mutazione in ceppi aploidi
- Per eliminare plasmidi con URA-3 (YCp, YEp e YIp)

Le piccole dimensioni del gene e la facilità della selezione negativa hanno fatto di URA-3 uno dei marcatori in assoluto più popolari per i lieviti

Nell'allele impiegato *ura3-52* è inserito un trasposone Ty1, questo evita l'integrazione di plasmidi Yip-URA3 nel locus *URA3* nella maggior parte dei ceppi di lievito

Pichia pastoris and Kluyveromyces lactis

- High density growth
- P. pastoris and K. lactis are considered GRAS (Generally Regarded As Safe)
- Protein export systems more efficient than S. cerevisiae
- Glycosylation more similar to mammals
- Inducible promoters
  - P. pastoris AOX1 (methanol), FLD (methanol or methylamine), CUP1 (copper)
  - K. lactis LAC4 (galactose, lactose), PHO5 (phosphate)
- Constitutive promoters
  - P. pastoris GAP, YPT1
  - K. lactis PGK
- More limited availability of strains and vectors



Methanol metabolism in *P. pastoris*

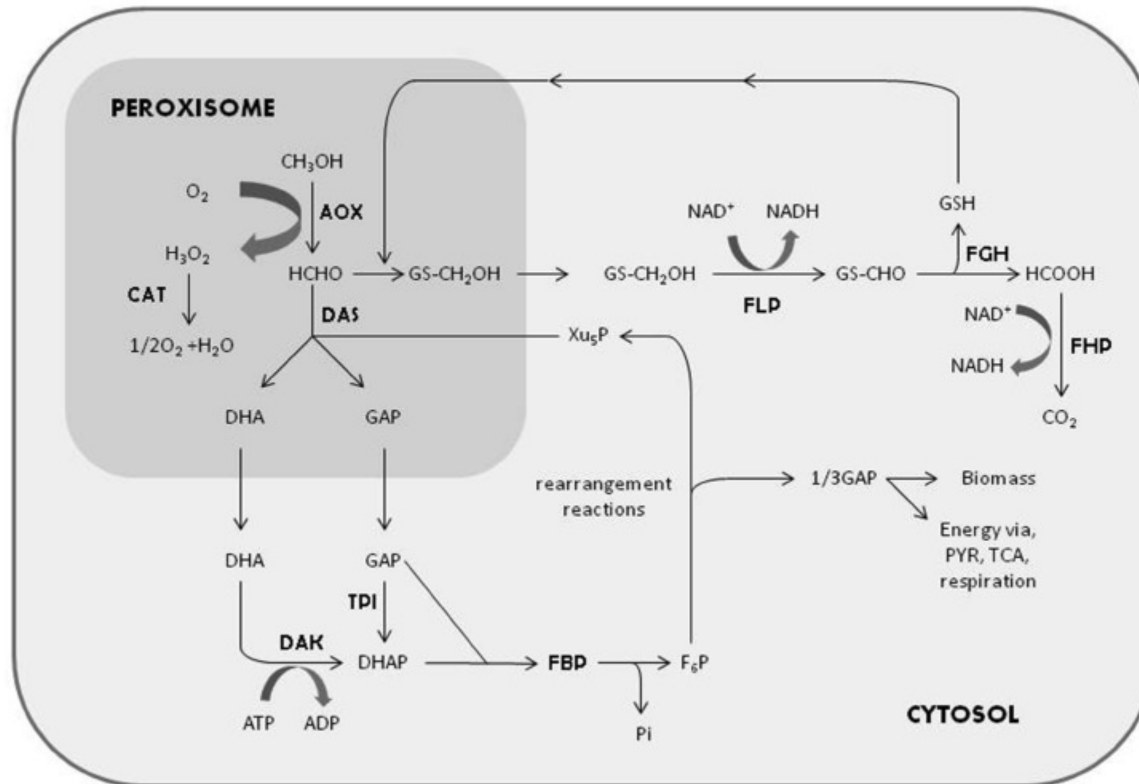


Fig. 2. Methanol metabolic pathway can be split into three steps.
1. Methanol is oxidized to formaldehyde by alcohol oxidase, alcohol dehydrogenase, or the catalase peroxidative pathway. 2. Formaldehyde is rapidly oxidized to formic acid mainly by formaldehyde dehydrogenase. Catalase and aldehyde dehydrogenase can also mediate this step. 3. Formic acid, which is the principal toxic agent in methanol poisoning, is metabolized by formate dehydrogenase. The catalase peroxidative pathway can also oxidize formic acid to CO_2.

Genomic integration of the vector in P. pastoris by homologous recombination

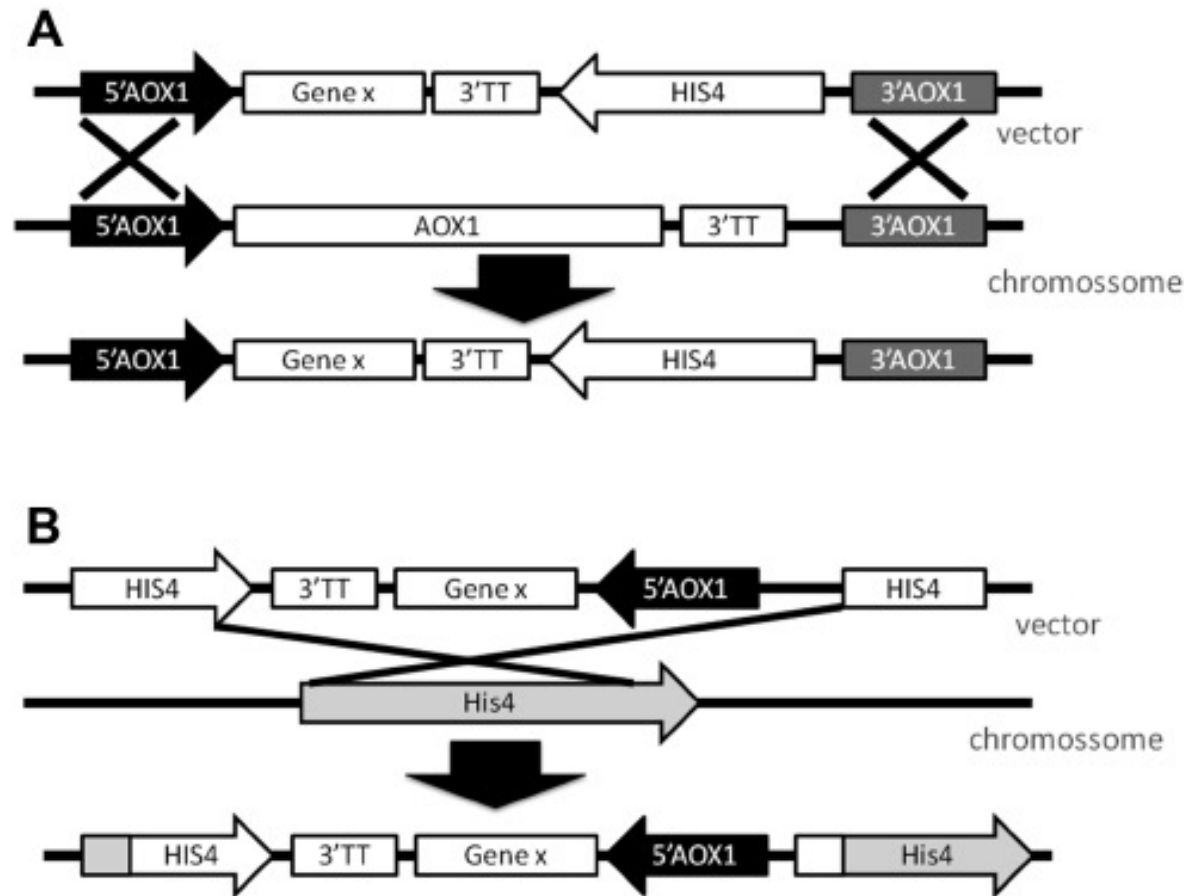
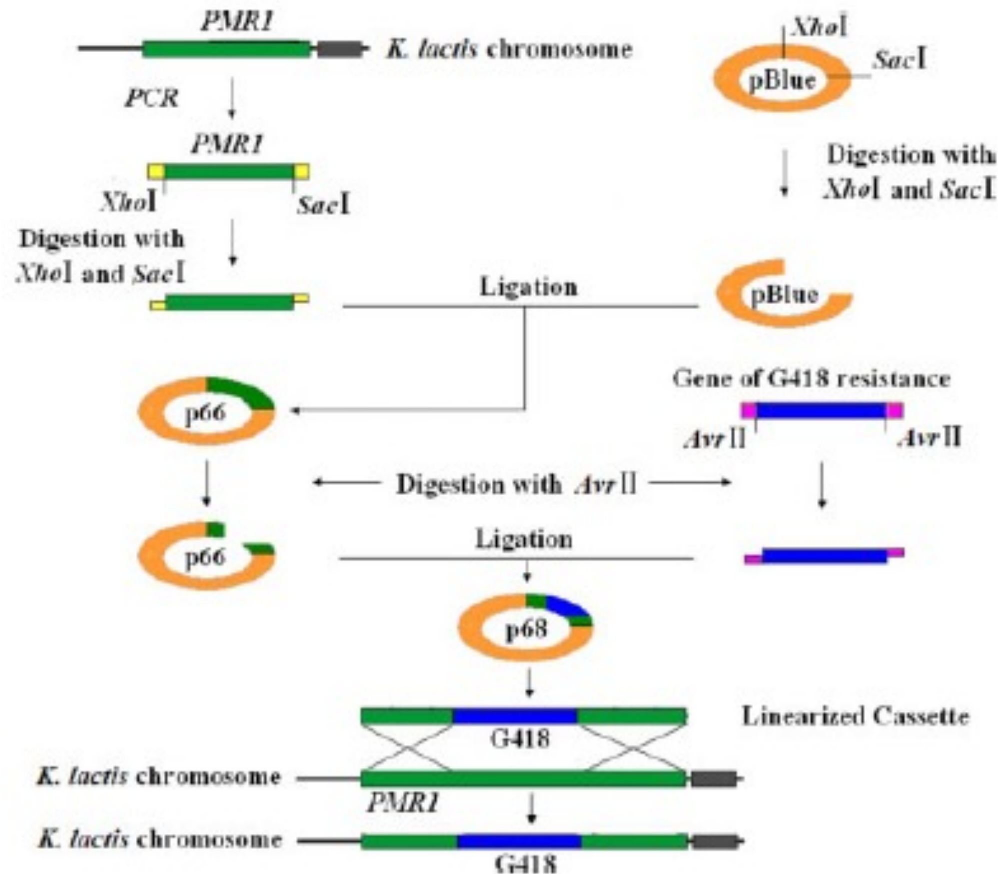


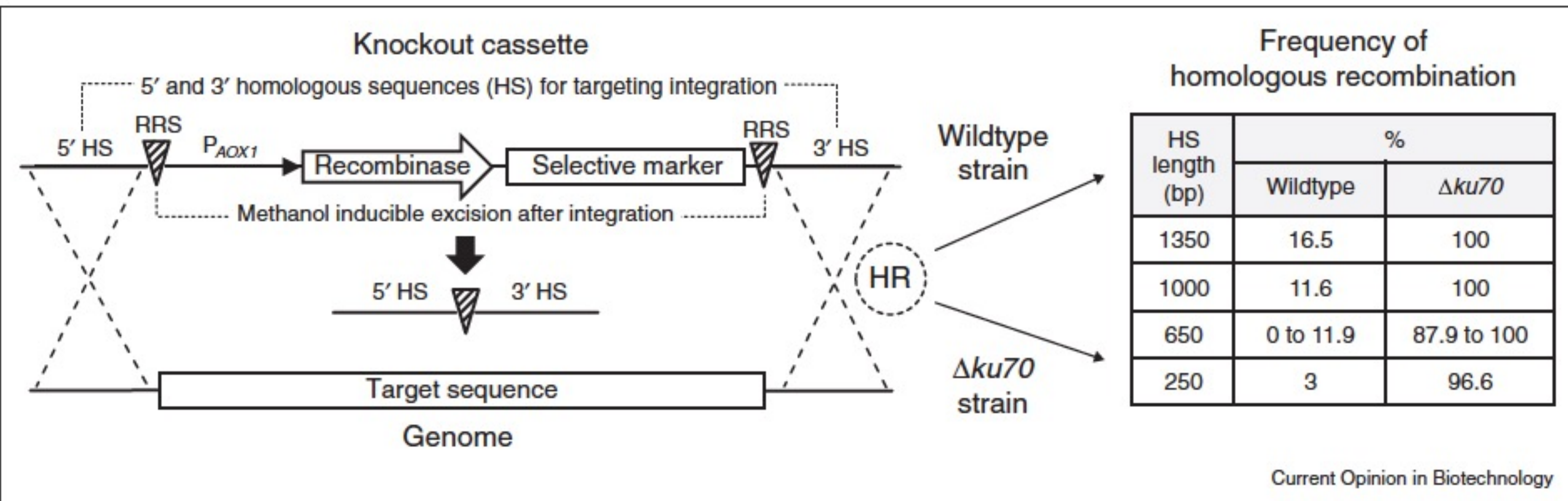
Fig. 1. Integration into *P. pastoris* genome by (**A**) gene replacement, and (**B**) gene insertion at his4.

Construction of mutant yeast strains
How can you inactivate a gene in yeast?

Inactivation of the gene PMR1 in K. lactis by insertion of a G418-resistance



Construction of mutant yeast strains
Cassettes for selection marker regeneration after gene inactivation



HS length (bp)	%	
	Wildtype	*Δku70*
1350	16.5	100
1000	11.6	100
650	0 to 11.9	87.9 to 100
250	3	96.6

Recombinase based self-excisable knockout cassettes for marker regeneration (left side). Increased rates of homologous recombination in a *P. pastoris Δku70* strain (right side). The knockout cassettes consist of a recombinase (Cre or FLP [48,49,50•]) and a marker gene flanked by the respective recombinase recognition sites and are directed to the genome via the 5′ and 3′ homologous sequences to delete the desired target sequence. After integration via a double cross-over event, self-excision of the recombinase and the marker gene can be initiated by the expression of the recombinase from the methanol inducible *AOX1* promoter (P_{AOX1}), leaving only the recombinase recognition site in the genome (notably Marx *et al.* [49] provided the recombinase transiently on a CEN/ARS plasmid). The initial integration in the genome is dependent on homologous recombination (HR). Exemplary frequencies of homologous recombination (in %) of the wildtype compared to the *Δku70* strain are shown (right side). The length of the homologous sequence indicates the number of base pairs (bp) added on both sides of the cassette [50•]. For 650 bp two different integration loci were tested, therefore two % values are given.

Mammalian cell lines

- Slow growth (ca 24 hours doubling time)
- Expensive and complex platform systems
- Episomal and integrative plasmids with strong inducible (fitoestrogens, TetOn) or constitutive (CMV) promoters
- Transient or stable transfection (lipocomplexes, electroporation)
- Transduction with viral-derived vectors
- Selection by antibiotic resistance or selective advantage (dihydrofolate reductase, glutamine synthase)
- Folding and post-translational modifications are generally correct

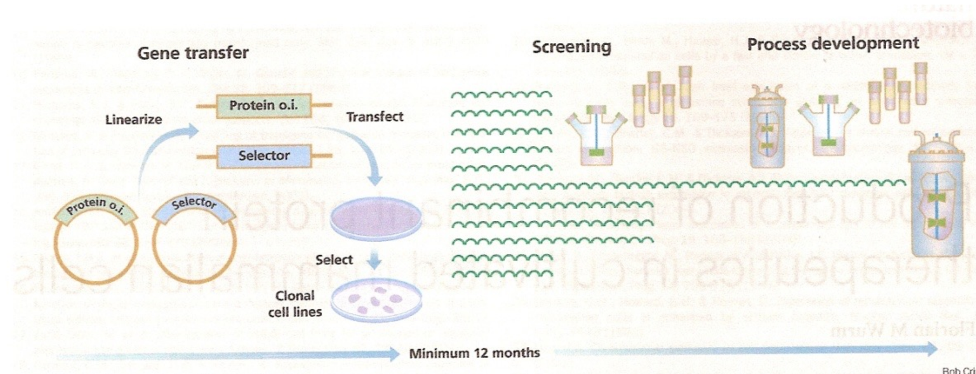
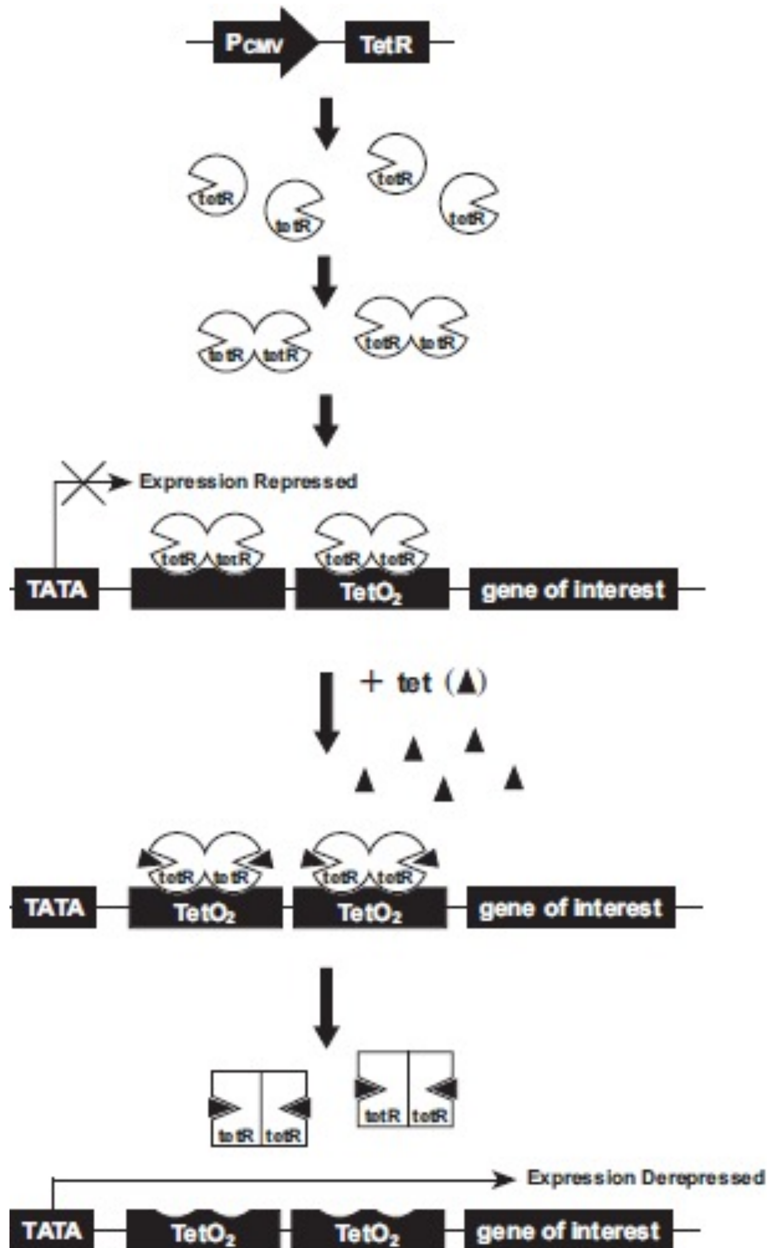


Figure 1 Cell line generation and development for cell culture processes for the generation of recombinant proteins of interest (o.i.). The wavy lines indicate subcultivations of individual cell lines that are in a screening program to obtain the final producer. Vials indicate banks of cells frozen in liquid nitrogen. Spinner flasks represent scale-down systems for process optimization, and bioreactors represent large-scale production processes. (Artwork after an original design by Danièle Fraboulet, Lyon, France.)

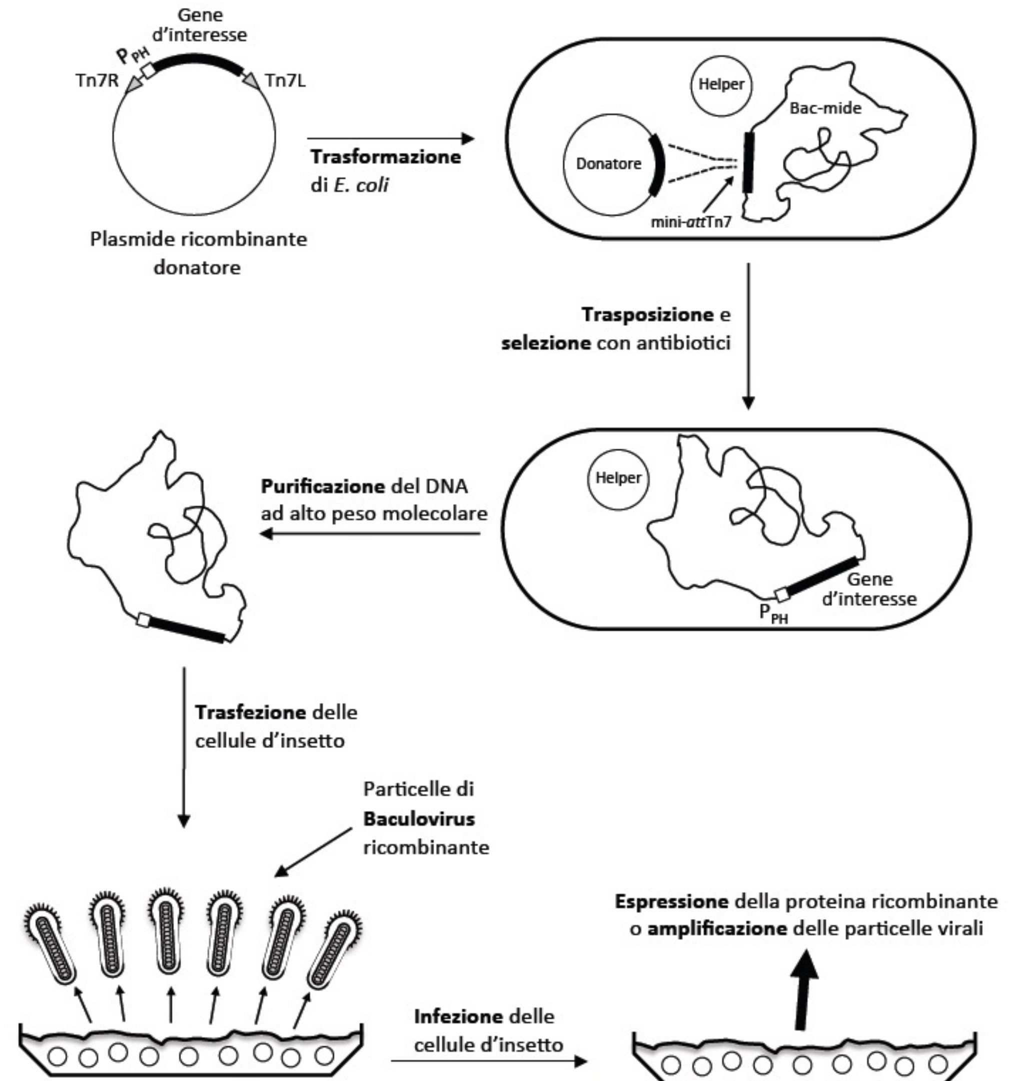


TetOn system

The bacterial repressor TetR binds to specific DNA sequences in the absence of tetracycline.
The gene of interest is under control of a hybrid tetO-CMV promoter and it is inserted in a cell line that constitutively expresses TetR.
In the presence of tetracycline or its analogue doxycycline, TetR does not bind to DNA and transcription of the gene of interest is allowed.

Insect cells/baculovirus

- Sf9 insect cells infected with recombinant baculovirus
- Strong polyhedrin promoter
- Slow growth (ca 24 hours/duplication)
- Still quite complex and expensive system
- Folding and post-translational modifications are generally correct



Purification of recombinant proteins

- Conventional methods: ammonium sulfate fractionation, ion exchange chromatography, gel-filtration or hydrophobic interaction chromatography
- Affinity chromatography exploiting peptide or protein tags fused to the protein (N- o C-terminal). Immobilization of antibodies against peptide tags or substrates of protein tags

Peptide tags: 6xHis, flag, c-myc, V5, HA

Protein tags: GST (glutathione-S-transferase), MBP (maltose binding protein), TRX (thioredoxin)

Tag removal: cleavage sites for proteases in the linker sequence between the protein and the tag

Peptide and protein tags

Tag	Residues	Sequence	Size (kDa)
Poly-Arg	5–6 (usually 5)	RRRRR	0.80
Poly-His	2–10 (usually 6)	HHHHHH	0.84
FLAG	8	DYKDDDDK	1.01
Strep-tag II	8	WSHPQFEK	1.06
c-myc	11	EQKLISEEDL	1.20
S-	15	KETAAAKFERQHMDS	1.75
HAT-	19	KDHLIHNVHKEFHAHAHNK	2.31
3x FLAG	22	DYKDHDGDYKDHDIDYKDDDDK	2.73
Calmodulin-binding peptide	26	KRRWKKNFIAVSAANRFKKISSSGAL	2.96
Cellulose-binding domains	27–189	Domains	3.00–20.00
SBP	38	MDEKTTGWRGGHVVEGLAGELEQLRARLEHHPQGQREP	4.03
Chitin-binding domain	51	TNPGVSAWQVNTAYTAGQLVTYNGKTYKCLQPHTSLAGWEPSNVPALWQLQ	5.59
Glutathione S-transferase	211	Protein	26.00
Maltose-binding protein	396	Protein	40.00

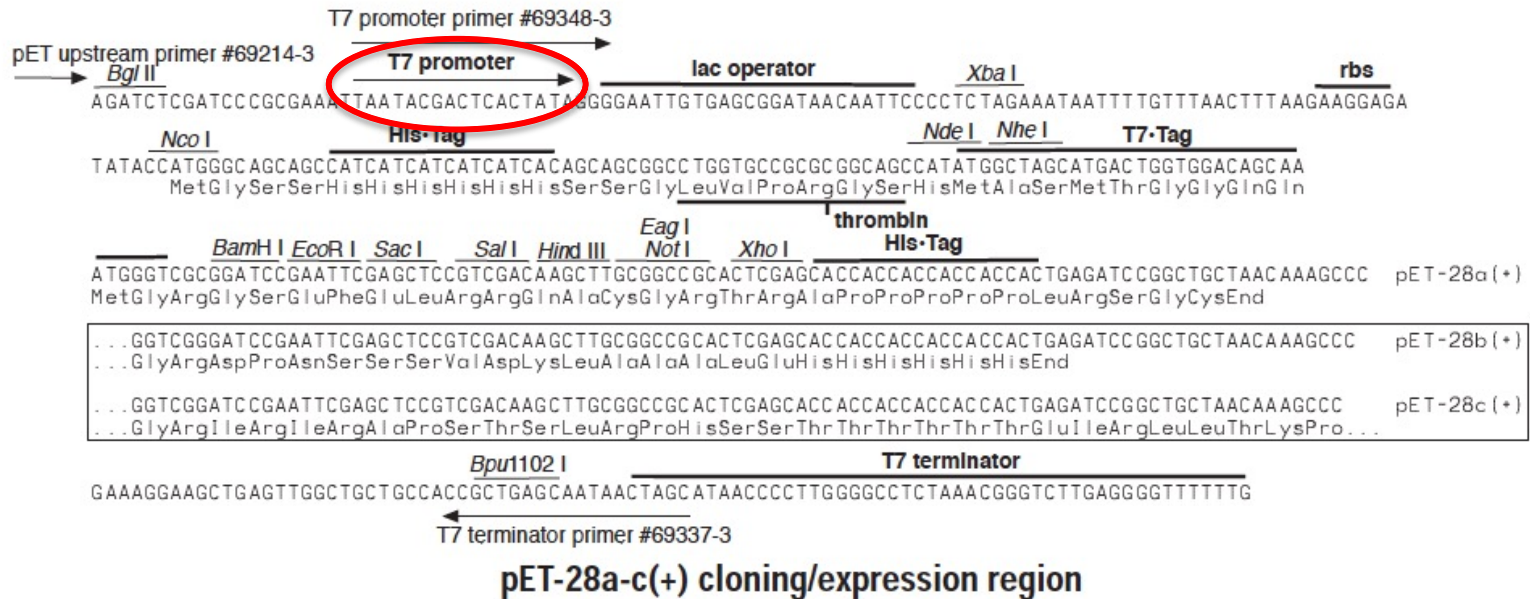
Peptide tags help purification and detection

Protein tags increase expression levels and solubility, help purification and detection

Affinity chromatography for purification of tagged recombinant proteins

Affinity tag	Matrix	Elution condition
Poly-Arg	Cation-exchange resin	NaCl linear gradient from 0 to 400 mM at alkaline pH>8.0
Poly-His	Ni^{2+}-NTA, Co^{2+}-CMA (Talon)	Imidazole 20–250 mM or low pH
FLAG	Anti-FLAG monoclonal antibody	pH 3.0 or 2–5 mM EDTA
Strep-tag II	Strep-Tactin (modified streptavidin)	2.5 mM desthiobiotin
c-myc	Monoclonal antibody	Low pH
S	S-fragment of RNaseA	3 M guanidine thiocyanate, 0.2 M citrate pH 2, 3 M magnesium chloride
HAT (natural histidine affinity tag)	Co^{2+}-CMA (Talon)	150 mM imidazole or low pH
Calmodulin-binding peptide	Calmodulin	EGTA or EGTA with 1 M NaCl
Cellulose-binding domain	Cellulose	Family I: guanidine HCl or urea>4 M Family II/III: ethylene glycol
SBP	Streptavidin	2 mM Biotin
Chitin-binding domain	Chitin	Fused with intein: 30–50 mM dithiothreitol, β-mercaptoethanol or cysteine
Glutathione S-transferase	Glutathione	5–10 mM reduced glutathione
Maltose-binding protein	Cross-linked amylose	10 mM maltose

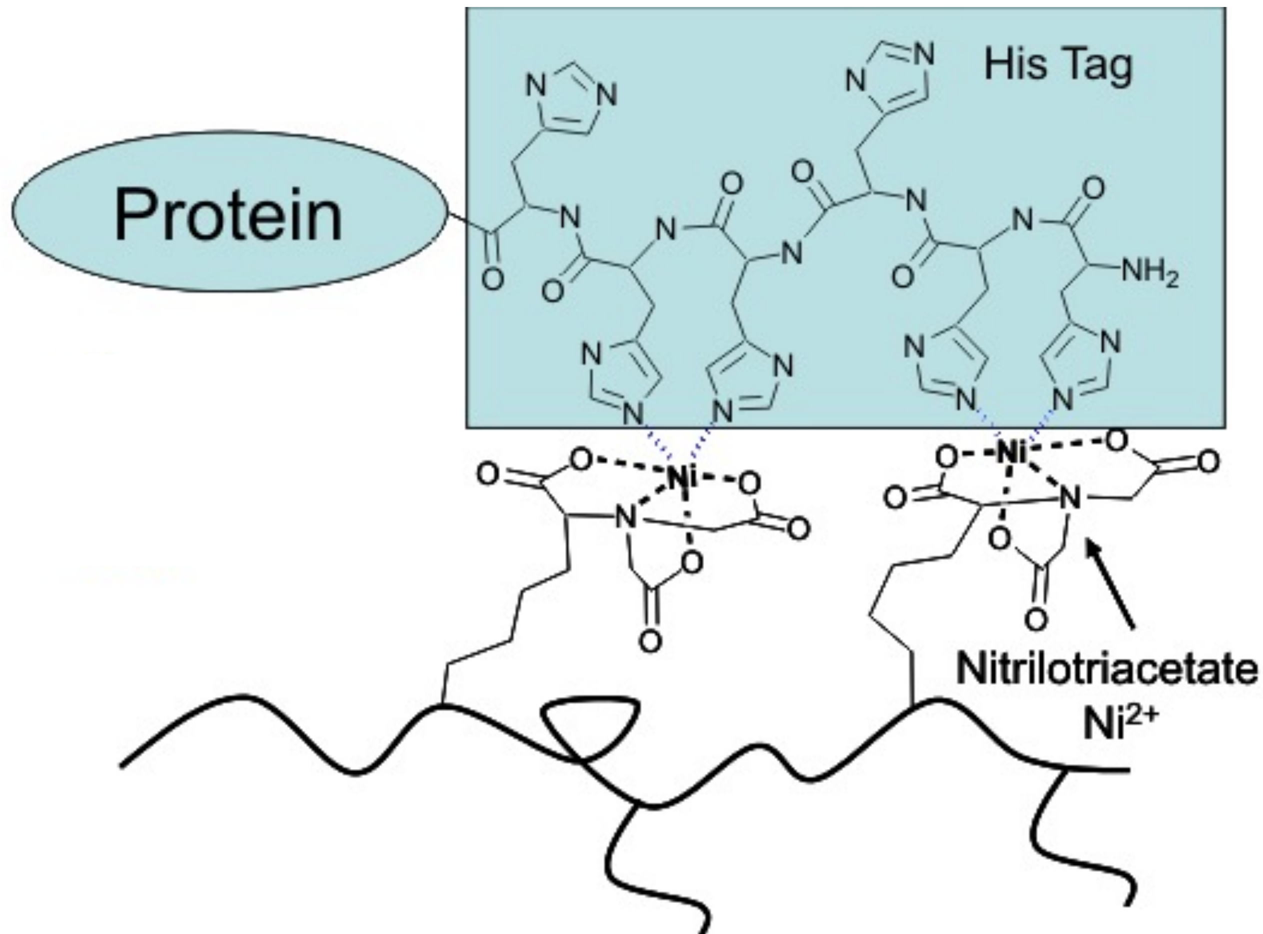
pET system for production of His-tagged proteins in E. coli



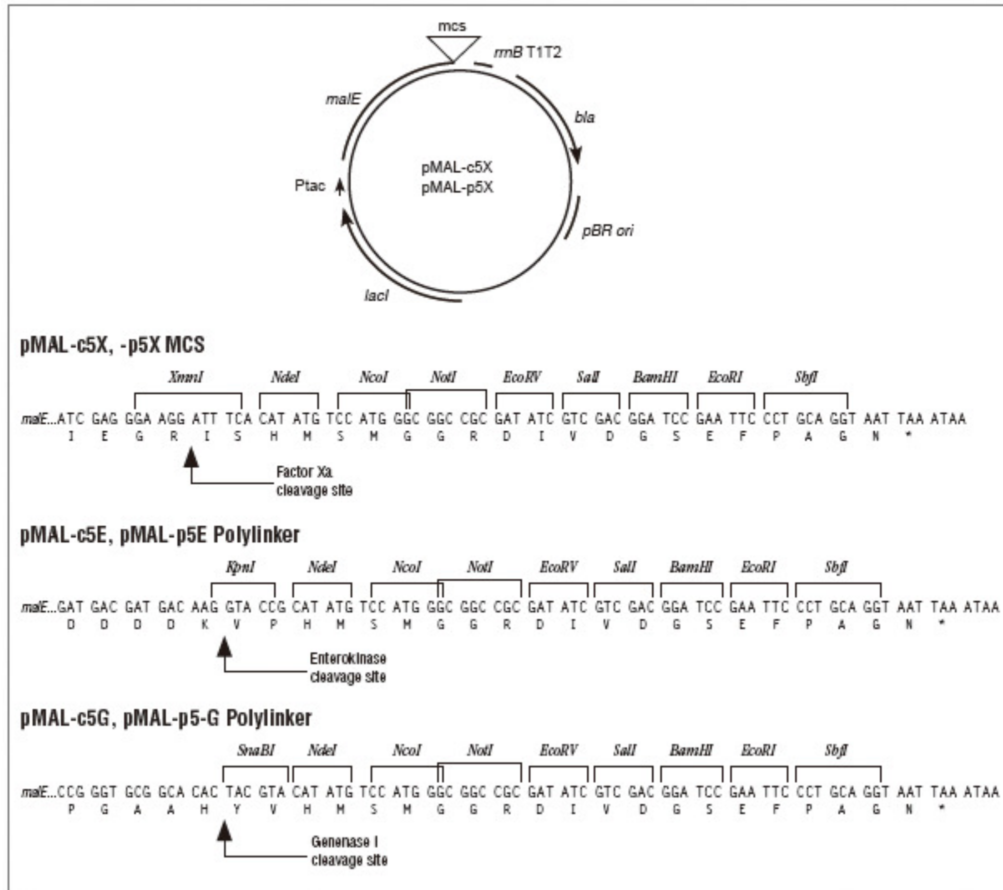
pET-28a-c(+) cloning/expression region

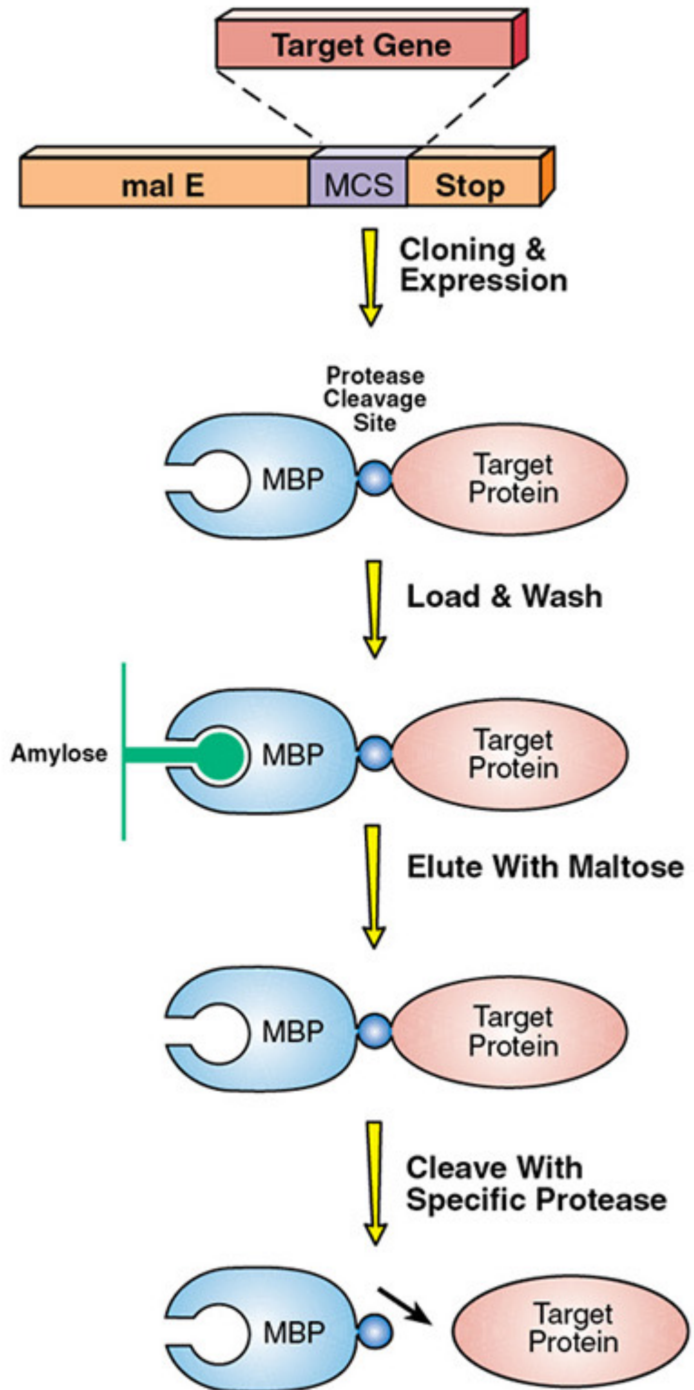
- T7 promoter
- Cloning site for addition of N- or C-terminal His-Tag
- Thrombin cleavage site for removal of N-terminal His-Tag
- The E. coli strain must contain T7 RNA polymerase

His-tag allows binding to a matrix with immobilized divalent metals: Ni^{2+} or Co^{2+}



System for production of MBP fusion proteins in E. coli





How to increase expression levels of recombinant proteins?

- Increasing translation
  - Ribosome binding sequences
  - mRNA stability
  - Codon usage
- Increasing folding
  - Changing growth conditions (temperature, medium composition)
  - Coexpression of folding machinery (molecular chaperones)
  - Fusion proteins
- Increasing stability
  - Secretion (requires N-terminal signal sequence)
  - Fusion proteins
  - Post-translational modifications

Molecular chaperones that assist protein folding in *E. coli*

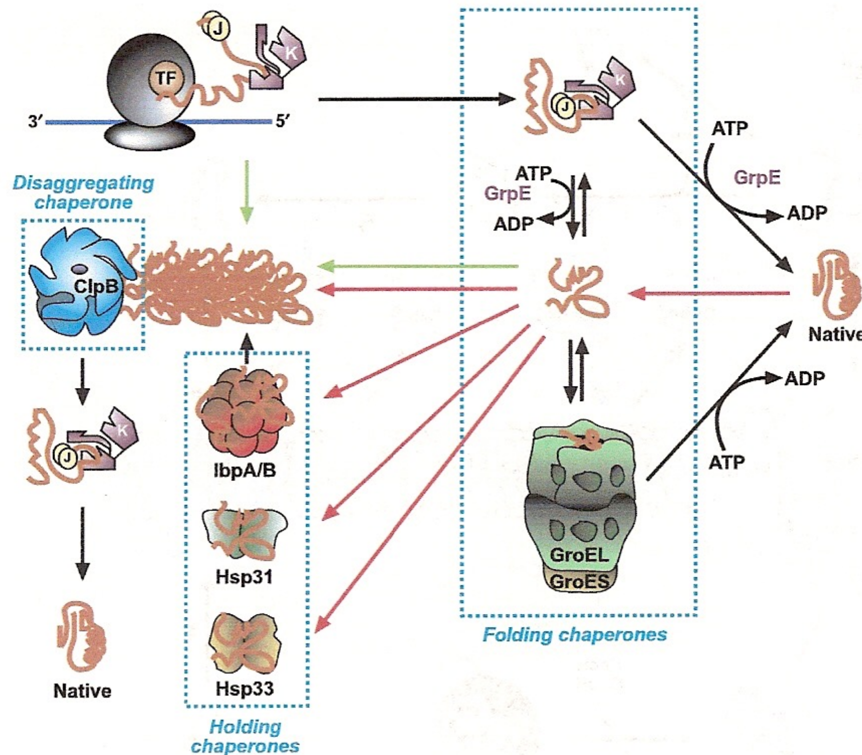
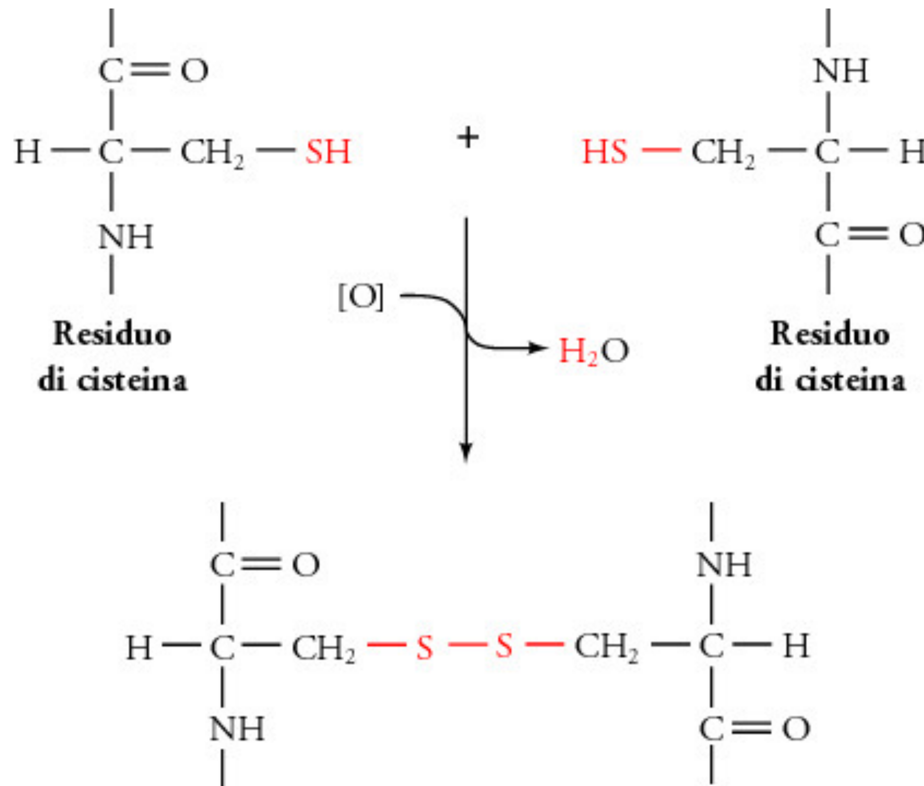


Figure 1 Chaperone-assisted protein folding in the cytoplasm of *E. coli*. Nascent polypeptides requiring the assistance of molecular chaperones first encounter TF or DnaK-DnaJ. Both chaperones engage solvent-exposed stretches of hydrophobic amino acids, shielding them from the solvent and each other. After undocking from TF- or GrpE-mediated release from DnaK, folding intermediate may reach a native conformation, cycle back to DnaK-DnaJ or be transferred to the central chamber of GroEL for folding at infinite dilution upon GroES capping. In times of stress (red arrows), thermolabile proteins unfold and aggregate. IbpB binds partially folded proteins on its surface to serve as a reservoir of unfolded intermediates until folding chaperones become available and intercalates within large aggregates. The holding chaperones Hsp33 and Hsp31 become important under oxidative and severe thermal stress, respectively. ClpB promotes the shearing and disaggregation of thermally unfolded host proteins and cooperates with DnaK-DnaJ-GrpE to reactivate them once stress has abated. Recombinant proteins that miss an early interaction with TF or DnaK/DnaJ, that undergo multiple cycles of abortive interactions with folding chaperones or titrate them out, accumulate in inclusion bodies (green arrows).

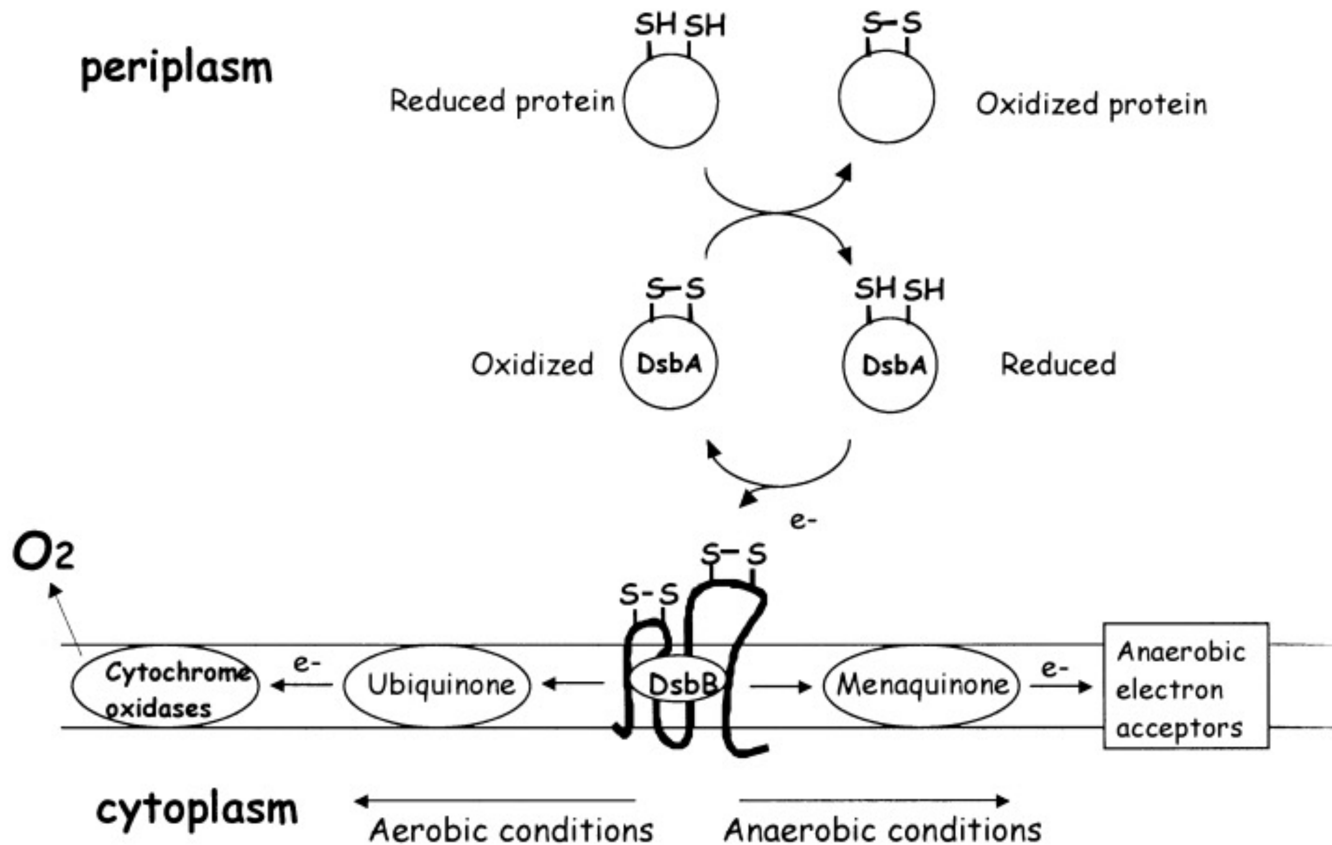
Cysteine and disulfide bridges



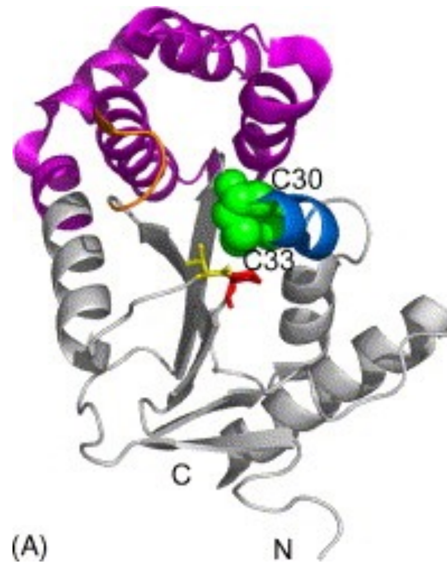
The formation of disulfide bridges takes place in the periplasm in bacteria and in the endoplasmic reticulum in eukaryotes

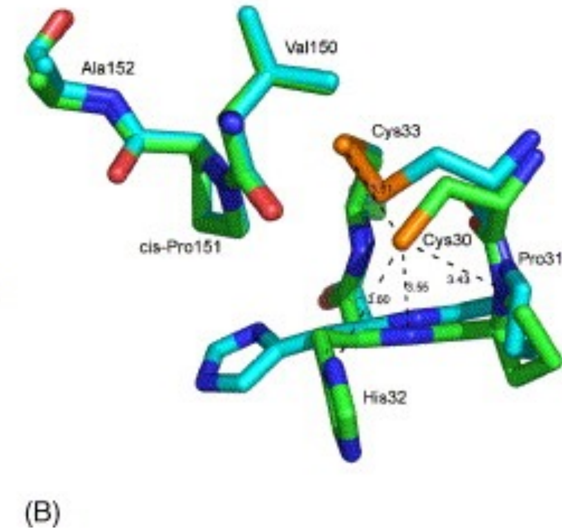
It requires oxidizing conditions

Disulfide bridge formation in the periplasm of *E. coli*



Mechanism of DsbA





Active site Cys_{30} Pro His Cys_{33}

Higher stability in the reduced state

High reactivity in the oxidized state

Stabilization of the reduced state determines Cys30 with low pKa, in the reduced state formation of:

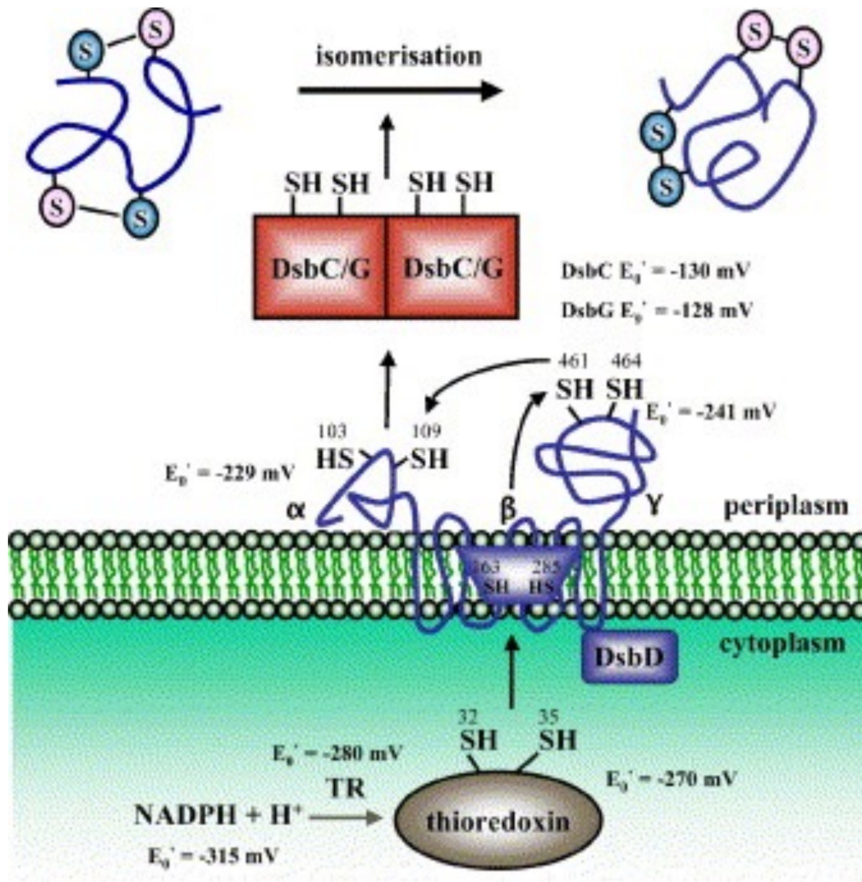
- Electrostatic bond at the positive extremity of an α-helix
- Hydrogen bond with His32.

Oxidized DsbA has a flexible structure that allows binding of a great number of substrates

Reduced DsbA has a rigid structure that allows binding of DsbB only

Isomerase: DsbC /DsbG

NADPH
Thioredoxin reductase
Thioredoxin
DsbDα-DsbDβ- DsbDγ
DsbC/DsbG
Reshuffling of incorrect disulfides



Folding and secretion mechanisms in *E. coli*

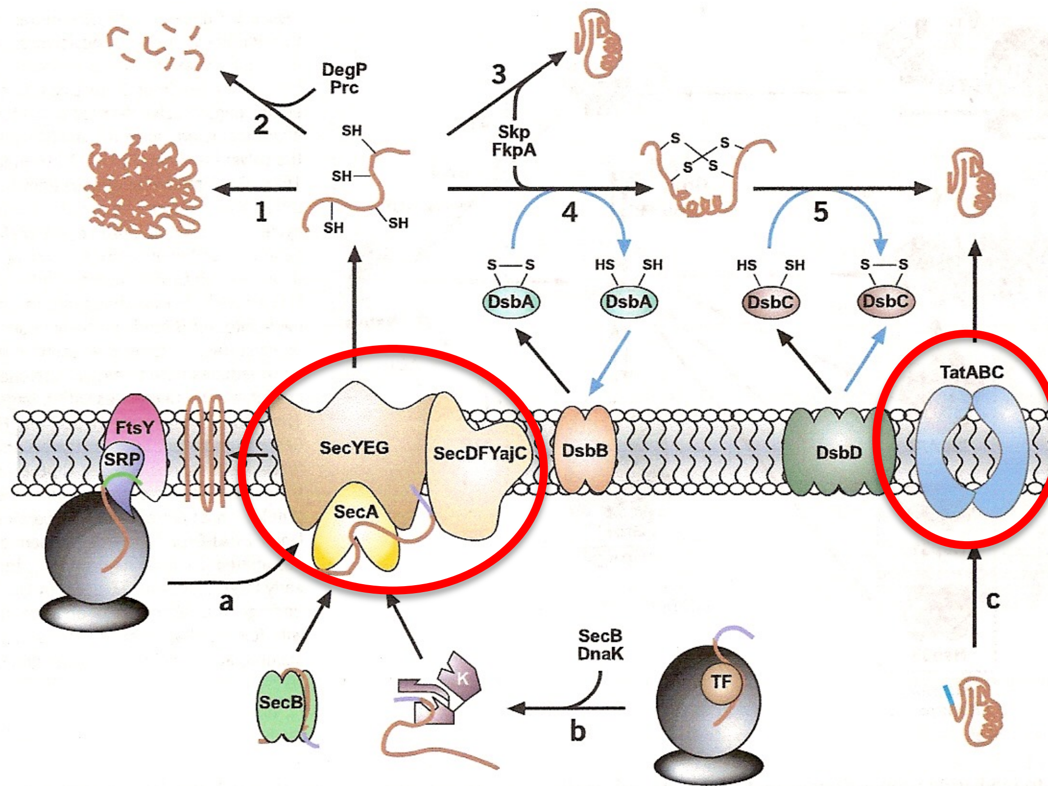


Figure 2 Export and periplasmic folding pathways. Proteins destined for export can be translocated across the inner membrane in three different fashions. **(a)** Preproteins with highly nonpolar signal sequences (green) or transmembrane segments of inner membrane proteins are recognized by SRP which, along with TF, scans nascent chains. SRP-dependent export involves delivery of the ribosome-nascent chain complex to FtsY and subsequent translocation through the SecYEG-SecDFYajC translocon. **(b)** The vast majority of preproteins have less hydrophobic signal sequence (lavender) and undergo Sec-dependent export. TF associates with the nascent polypeptide, halting cotranslational folding. As the chain grows, TF dissociates and the polypeptide is transferred to SecB or DnaK that maintain it in an extended conformation. Delivery to SecA and ATP-dependent translocation through SecYEG completes the process. **(c)** Preproteins with signal sequences containing the twin-arginine motif (cyan) are exported via the Tat-dependent pathway in a folded form. After cleavage of the signal sequence, partially folded periplasmic proteins may aggregate (1), undergo proteolysis (2) or reach a native conformation, possibly with the assistance of folding modulators (3). Cysteine pairs in proteins containing disulfide bonds are oxidized by DsbA (4) whereas incorrect disulfides are isomerized by DsbC (5). These oxidoreductases are reactivated by DsbB and DsbD, respectively. Black arrows show products obtained after each step, whereas blue arrows represent electron flow.

Folding and secretion mechanisms in eukaryotic cells

- Endoplasmic Reticulum
  - Removal of the signal peptide
  - First steps of glycosylation
  - Formation of disulfide bridges and folding
- Golgi Network
  - Maturation of oligosaccharides
- Secretion Vesicles
  - Transport to the plasma membrane and secretion

Secretion pathway in yeast

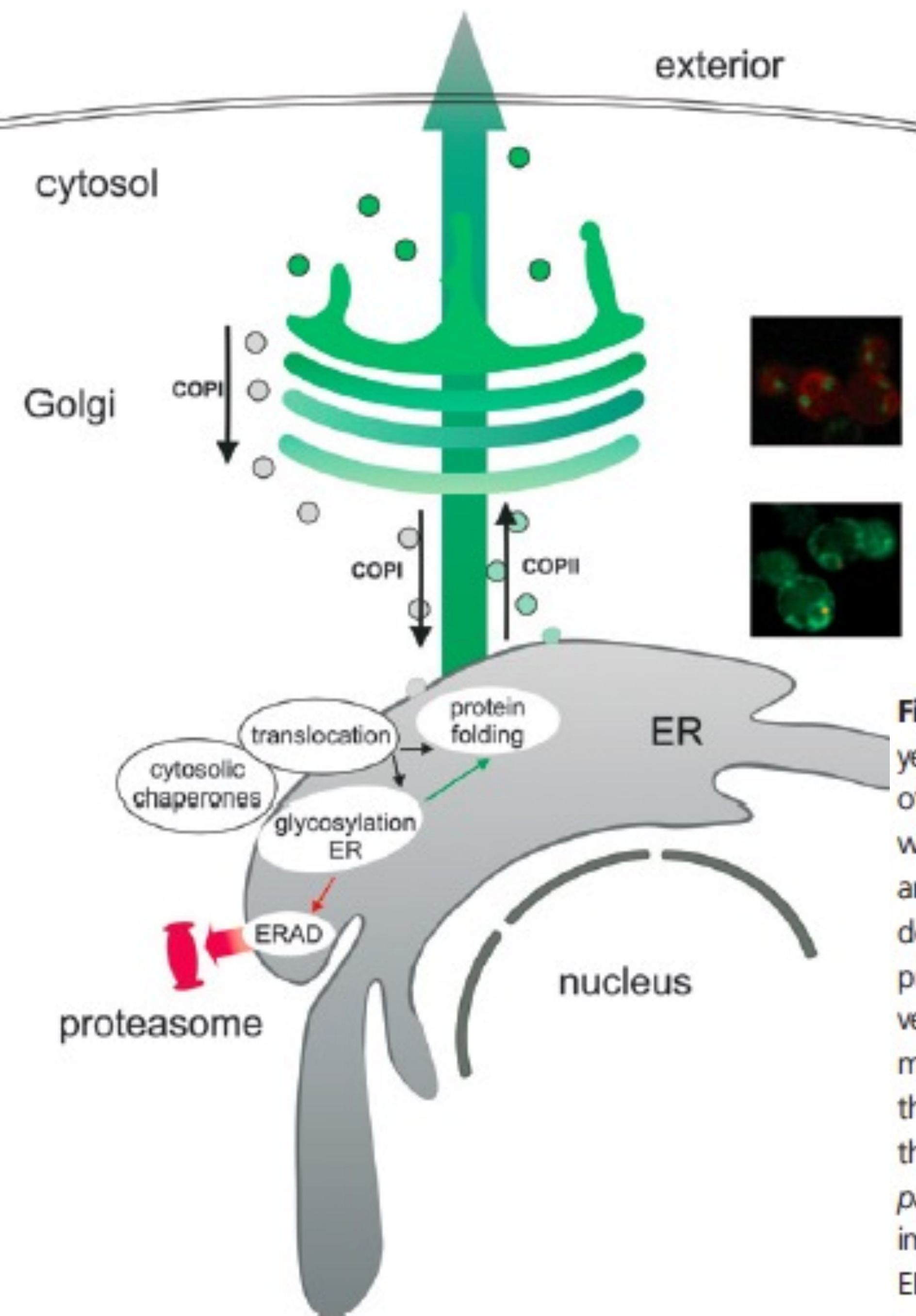
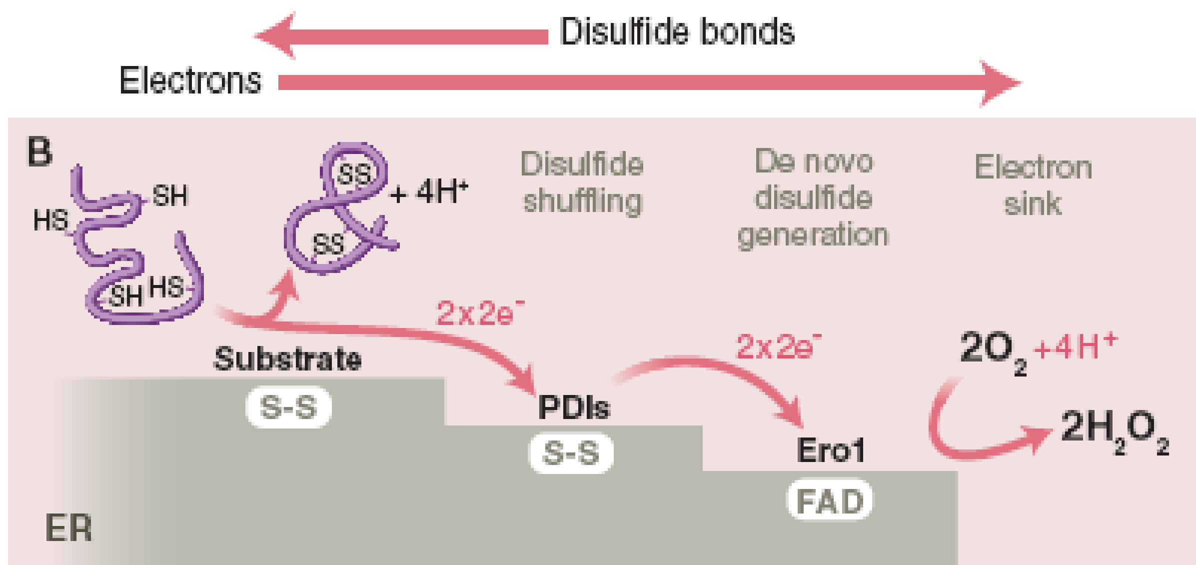
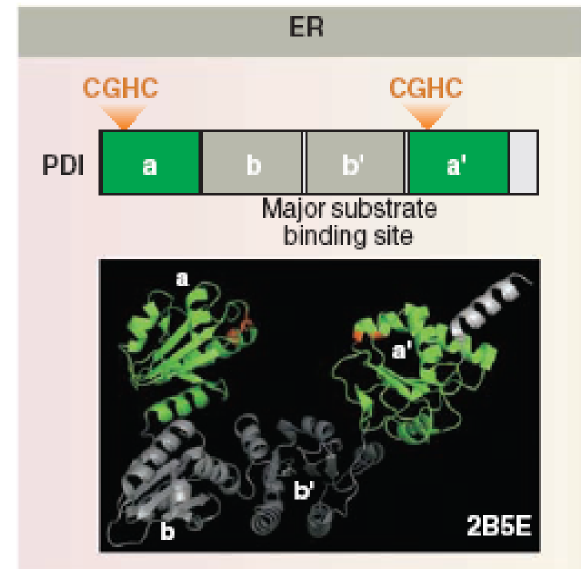
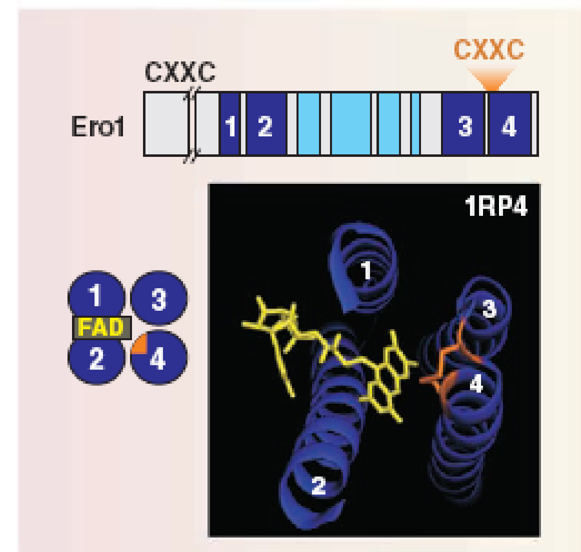


Fig. 2. Overview of the canonical protein secretion pathway in yeasts. During or after synthesis in the cytosol and eventual binding of cytosolic chaperones proteins are translocated to the ER lumen where they may by glycosylated. Folding is assisted by chaperones and enzymes. Misfolded proteins become target of the ER-associated degradation pathway which ends in proteasomal decay. Folded proteins are transported by COPII vesicles to the Golgi while COPI vesicles are responsible for retro-transport to *cis*-Golgi and ER. While most proteins that enter the secretion pathway remain in the cell, those destined for the exterior are transported by secretory vesicles to the cell membrane. Inserts: Fluorescence micrographs of *Pichia pastoris* cells illustrate the native relations inside yeast cells. Lower image: ER colored in green, and COPII vesicles in red. Upper image: ER in red, and Golgi in green.

Formation of disulfide bridges in the endoplasmic reticulum
PDI: protein disulfide isomerase
Ero1: sulfhydryl-oxidase







Glycosylation of proteins. Role of the glycocomponent

- Glycosylation may assist folding of the polypeptide chain by affecting secondary structure
- Oligosaccharide chains may contribute to protein stability by increasing solubility, shielding hydrophobic regions, protecting from proteases or favoring intra-chain interactions
- Glycosylation may affect trafficking/sorting of a protein to its final destination
- Oligosaccharide chains may participate to recognition/binding to ligands or receptors or to the biological activity of a protein
- Sialic acid that is found at the terminus of oligosaccharide chains increases the half-life of proteins in plasma
- Oligosaccharide chains may contribute to immunogenicity of heterologous proteins

Protein N-glycosylation at N-X-T/S sequence motifs

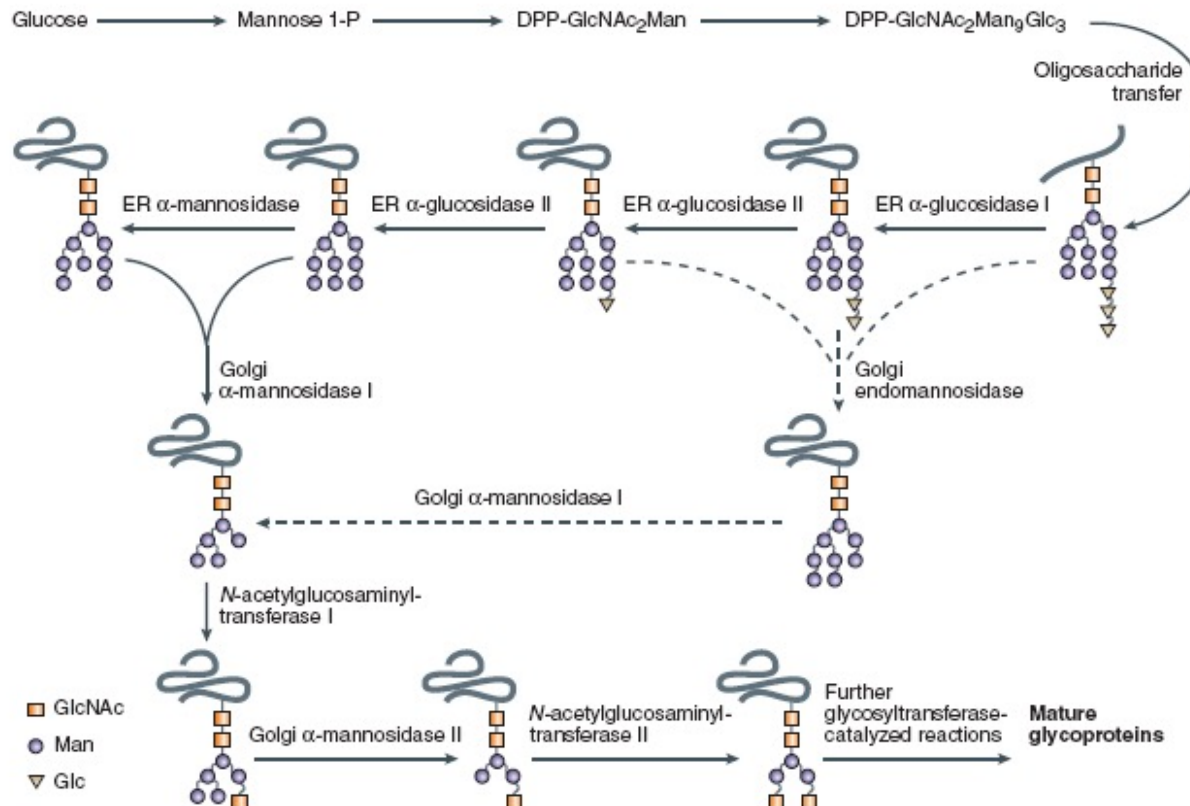
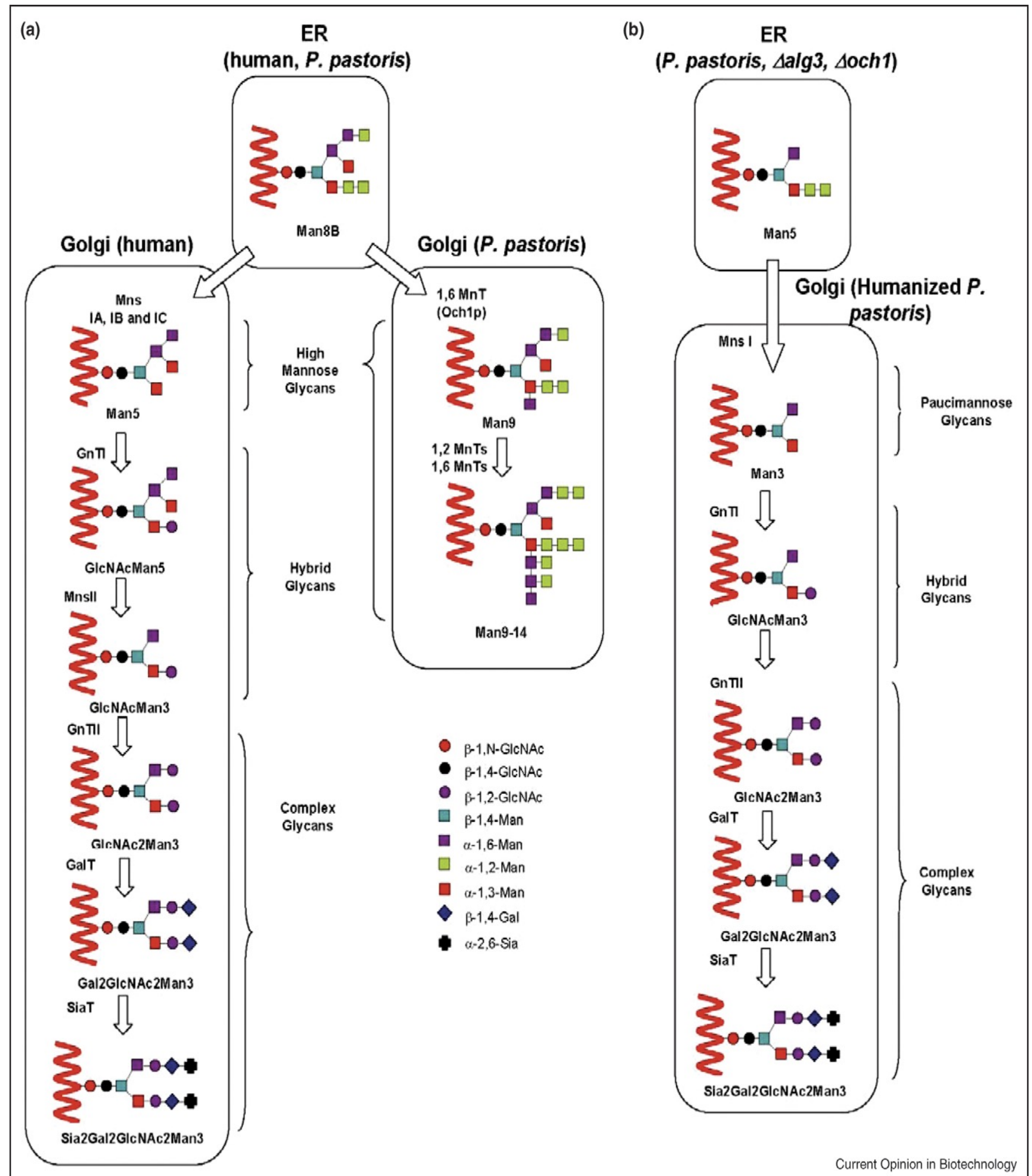


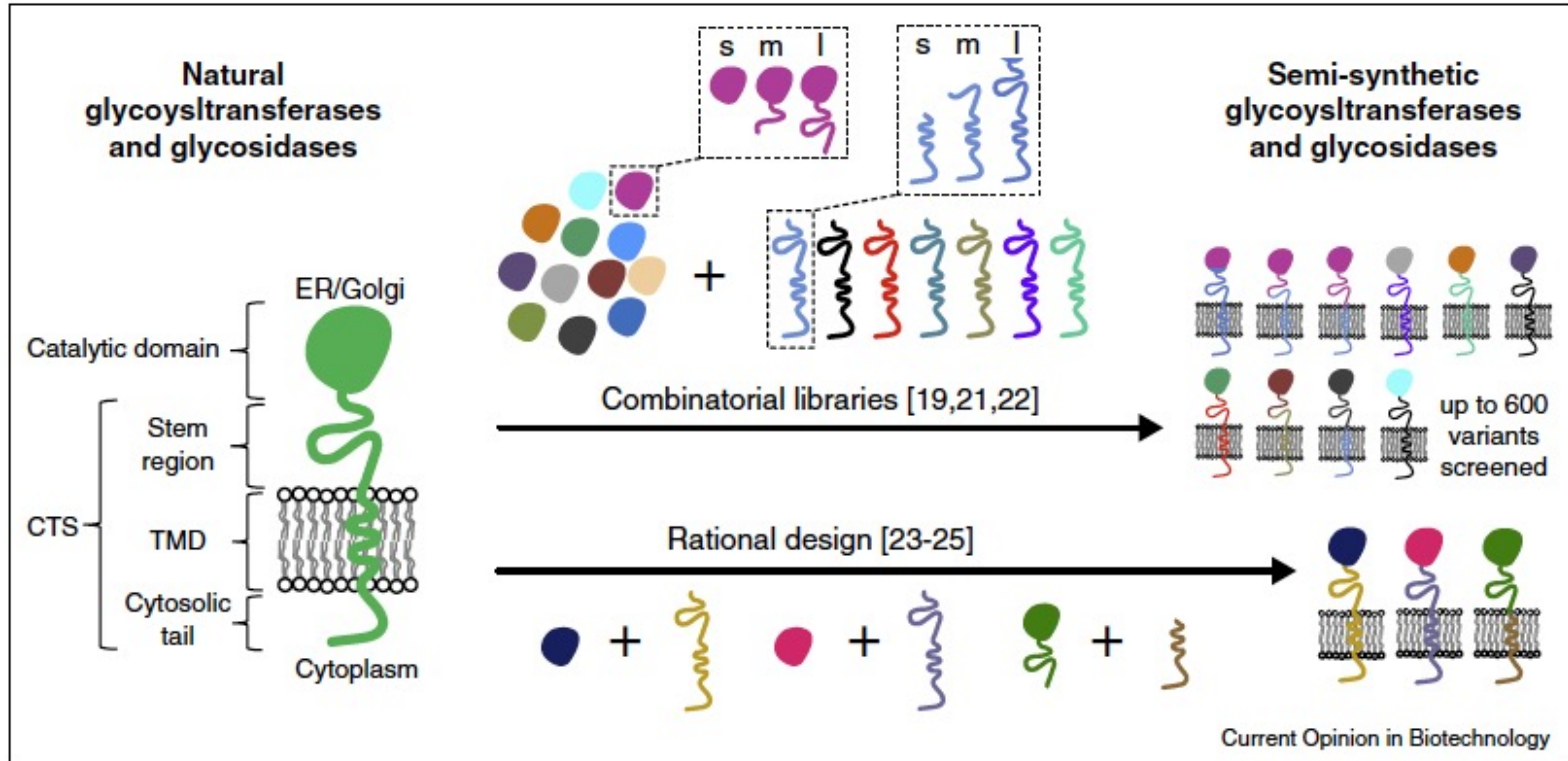
Figure 1 Overview of the biosynthesis of protein N-linked glycan side chains. As the nascent glycoprotein enters the ER, a preformed oligosaccharide, dolicol phosphate precursor (DPP) is attached cotranslationally to some asparagine residues that are part of the consensus sequence Asn-Xaa-Ser/Thr (where Xaa is any amino acid). The biosynthesis of this precursor, its attachment to asparagine residues and the subsequent steps of its processing in the ER and in the Golgi are performed by a series of glycosidases and glycosyltransferases. The exact profile of such enzymes present in the cell will largely dictate the composition and structure of the final oligosaccharide side chain, which are often described as 'complex', 'high mannose' or 'hybrid' type. Secretion at various stages of the carbohydrate elaboration yields heterogeneous glycoprotein products. (Reprinted by permission from *Nature Reviews Drug Discovery*, Macmillan Magazines Ltd., ref. 136.)

Humanization of glycosylation in the yeast *Pichia pastoris*

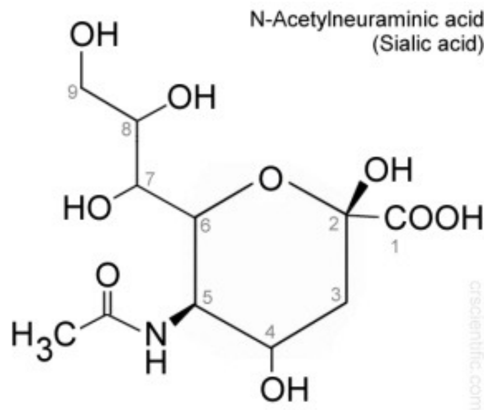
N-linked glycosylation pathways in humans and yeast. Representative N-linked glycosylation pathways in humans and P. pastoris (a). An alternative humanized N-linked glycosylation pathway in P. pastoris (b). Mns; α 1,2-mannosidase, MnsII; mannosidase II, GnTI; β 1,2-Nacetylglucosaminyltransferase I, GnTII; β 1,2-N-acetylglucosaminyltransferase II, GalT; β 1,4-galactosyltransferase, SiaT; α 2,6-sialyltransferase, MnT; mannosyltransferase. For simplicity the two GlcNAc residues present at the reducing end of all glycans have been omitted.



Strategies to create hybrid glycosyltransferases



Design strategies to create semi-synthetic glycosyltransferases and glycosidases for glycoengineering. On the left side, the general domain structure of glycosyltransferases and glycosidases is shown. These type II membrane proteins consist of an N-terminal cytosolic tail, a transmembrane domain (TMD), a stem region (these elements are referred to as CTS), and a C-terminal catalytic domain. In the middle and on the right side, design strategies for creating tailor-made enzymes with the desired catalytic activity and the proper localization in the sec pathway are shown. The combinatorial library approach involved the combination of large sets of catalytic domains with CTS fragments to fusion proteins, which were then screened for the desired activity [19,21,22]. Different lengths of the catalytic domains and the CTS fragments were tested (referred to as 's' for short, 'm' for medium, 'l' for long and shown exemplarily for one catalytic domain and one CTS). Rational approaches were also used to design these chimeric enzymes [23–25]. The schematic for the domain architecture and the combinatorial libraries is based on Czlapinski *et al*. [20] and Nett *et al*. [26••].



Synthesis and incorporation of sialic acid in glycoproteins in yeast

Glycoengineering steps required for sialic acid transfer in the yeast Golgi. Endogenous UDP-GlcNAc, present in the yeast cytoplasm, is converted to CMP-sialic acid by UDP-N-acetylglucosamine-2-epimerase/N-acetylmannosamine kinase (GNE), N-acetylneuraminate-9-phosphate synthase (SPS), sialylate-9-P phosphatase (SPP) and CMP-sialic acid synthase (CSS). Subsequently, the product is translocated into the Golgi by the CMP-sialic acid transporter (CST) and sialic acid is transferred onto the acceptor glycan by sialyltransferase (ST). Enzymes are indicated by blue text and metabolic intermediates by black text. For simplicity the two GlcNAc residues present at the reducing end of all glycans have been omitted.

ST
Gal2GlcNAc2Man3 (GS5.0)
CMP-Sialic Acid
Sia2Gal2GlcNAc2Man3 (GS6.0)
Golgi
Cytoplasm
CST
CMP-Sialic Acid
CSS
UDP-GlcNAc → ManNAc-6-P → Sialylate-9-P → Sialic acid
GNE SPS SPP
Current Opinion in Biotechnology