Chapter 9

Promoter and Regulatory Element Prediction

Overview

1. Introduction
2. Introduction to Biological Databases
3. Pairwise Sequence Alignment
4. Database Similarity Searching
5. Multiple Sequence Alignment
6. Profiles and Hidden Markov Models
7. Protein Motifs and Domain Prediction
8. Gene Prediction
9. Promoter and Regulatory Element Prediction
10. Phylogenetics Basics
11. Phylogenetic Tree Construction Methods and Programs
12. Protein Structure Basics
13. Protein Structure Visualization, Comparison and Classification
14. Protein Secondary Structure Prediction
15. Protein Tertiary Structure Prediction
16. RNA Structure Prediction
17. Genome Mapping, Assembly and Comparison
18. Functional Genomics
19. Proteomics

Gene promoters

- Promoters are regulatory DNA regions located in the vicinity (mostly upstream) of transcription start sites
- Promoters determine the temporal and spatial expression pattern of the gene (*i.e.* where and when the gene is expressed, under which conditions)
- Promoters contain recognition sites for the transcription machinery (RNA polymerase, general transcription factors) and gene-specific transcription regulators (activators, repressors)
- Experimental determination of promoters and regulatory elements is time consuming and laborious

Bacterial promoters

RNA polymerase needs to bind to the promoter for transcription initiation to take place:

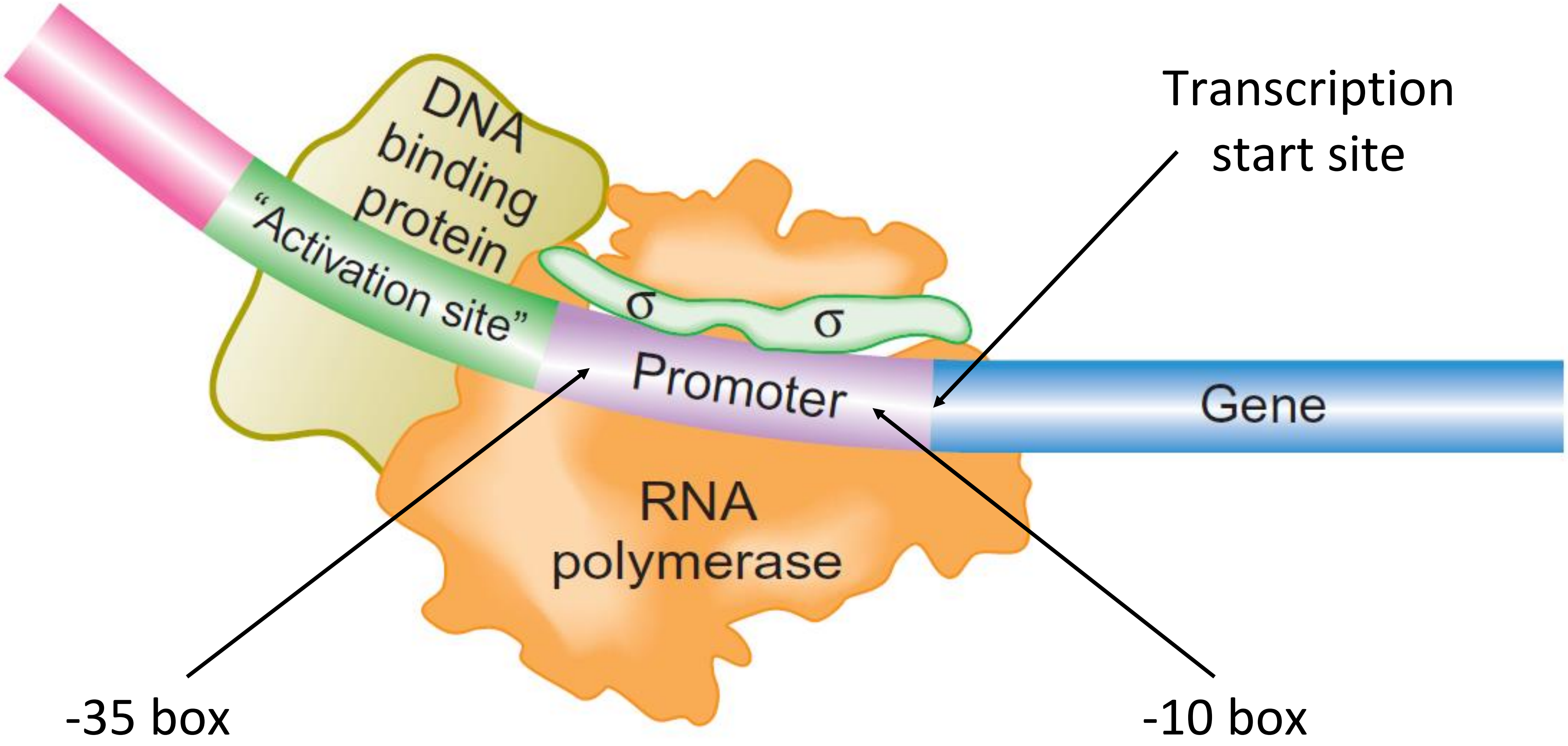
- **σ subunit** of RNA polymerase recognises specific sequence elements upstream of a gene
- **-35 and -10 boxes**: promotor sequence elements located 35 and 10 base pairs upstream from the start site
- *E.g.* consensus sequences of σ^{70} subunit of E. coli:

  -35 box: **TTGACA**

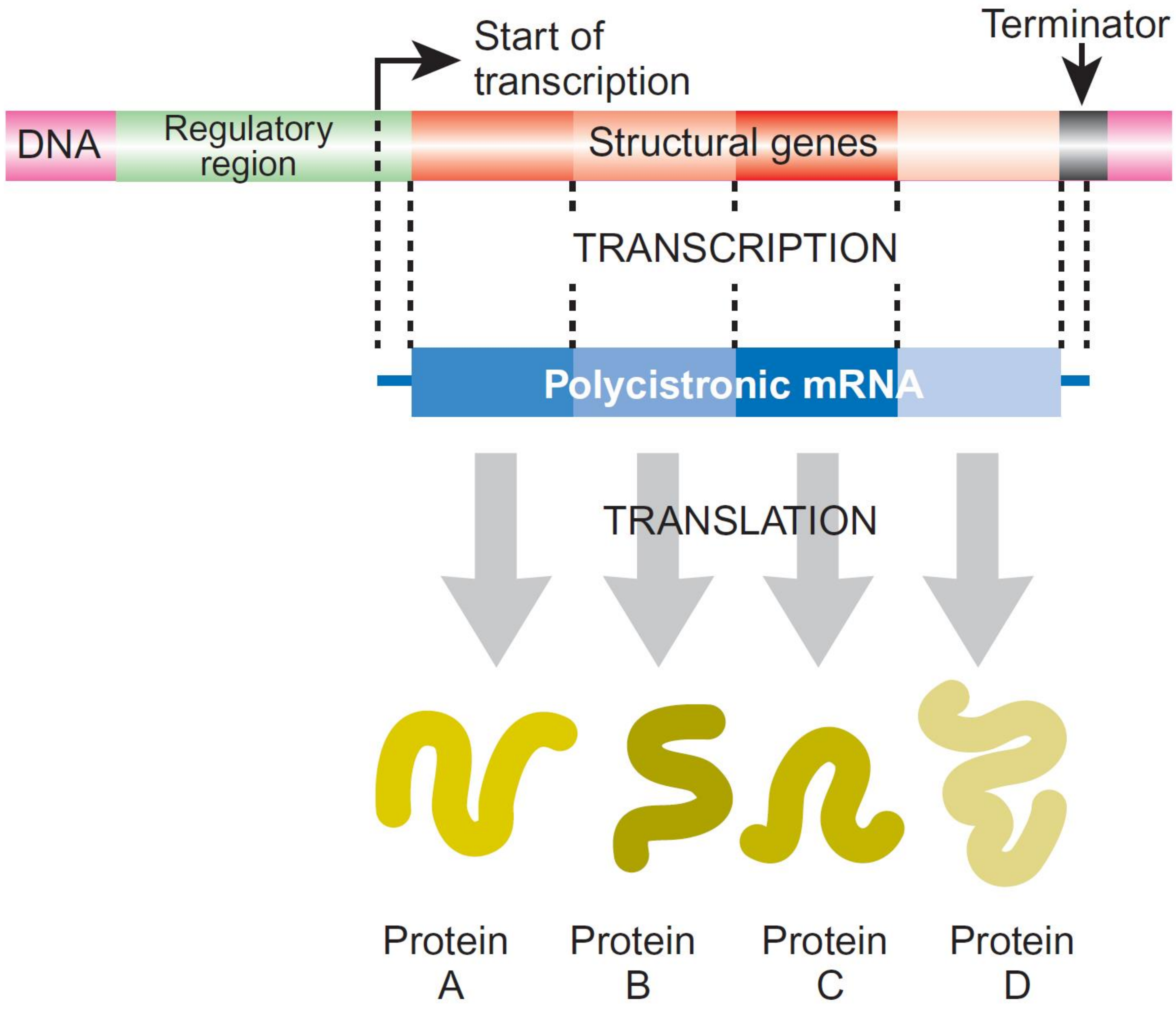
  -10 box: **TATAAT**

- Gene-specific regulatory factors directly stimulate or prevent binding of the RNA polymerase to the promoter

A simple bacterial promoter

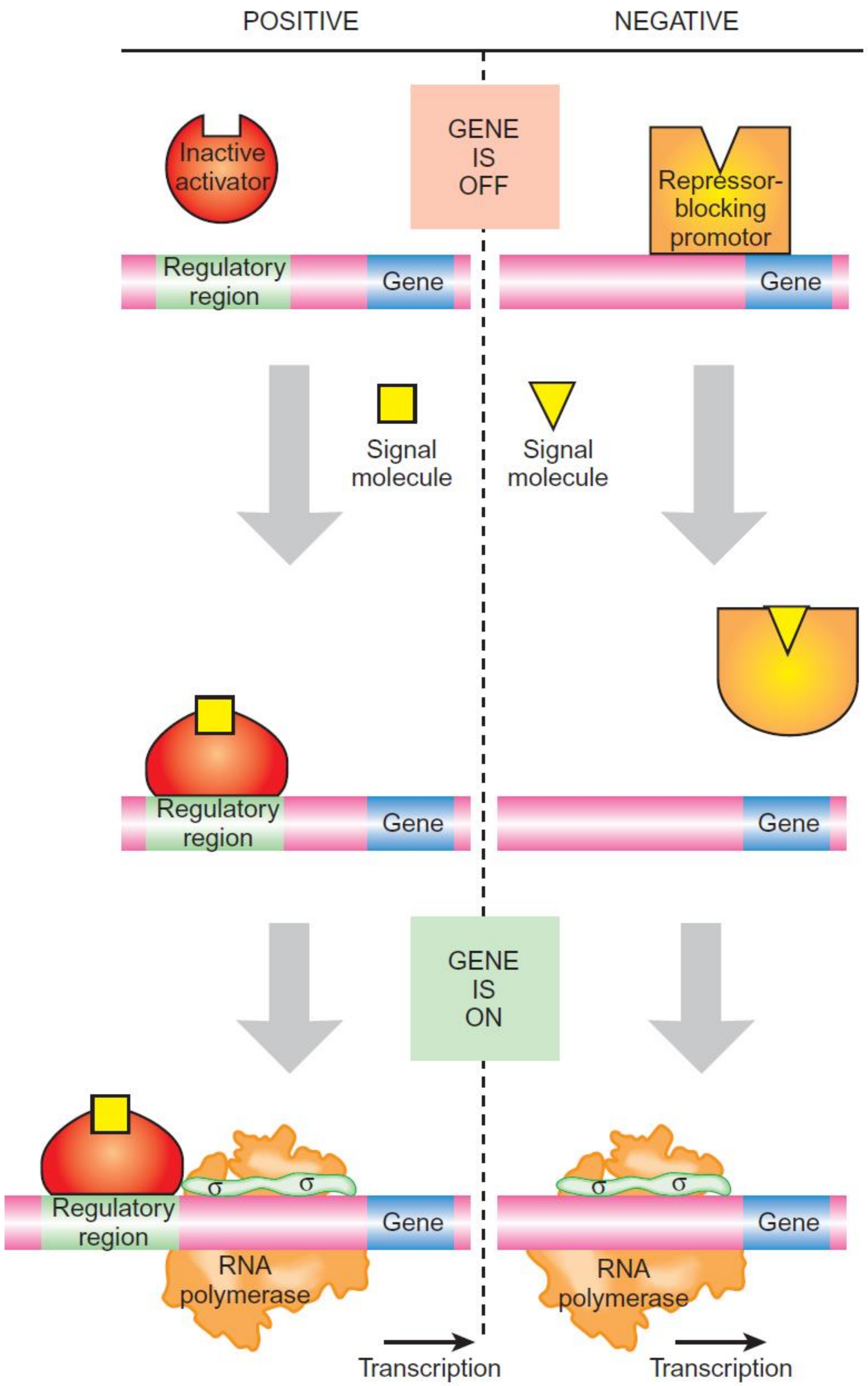


Promoters in bacteria often control operons



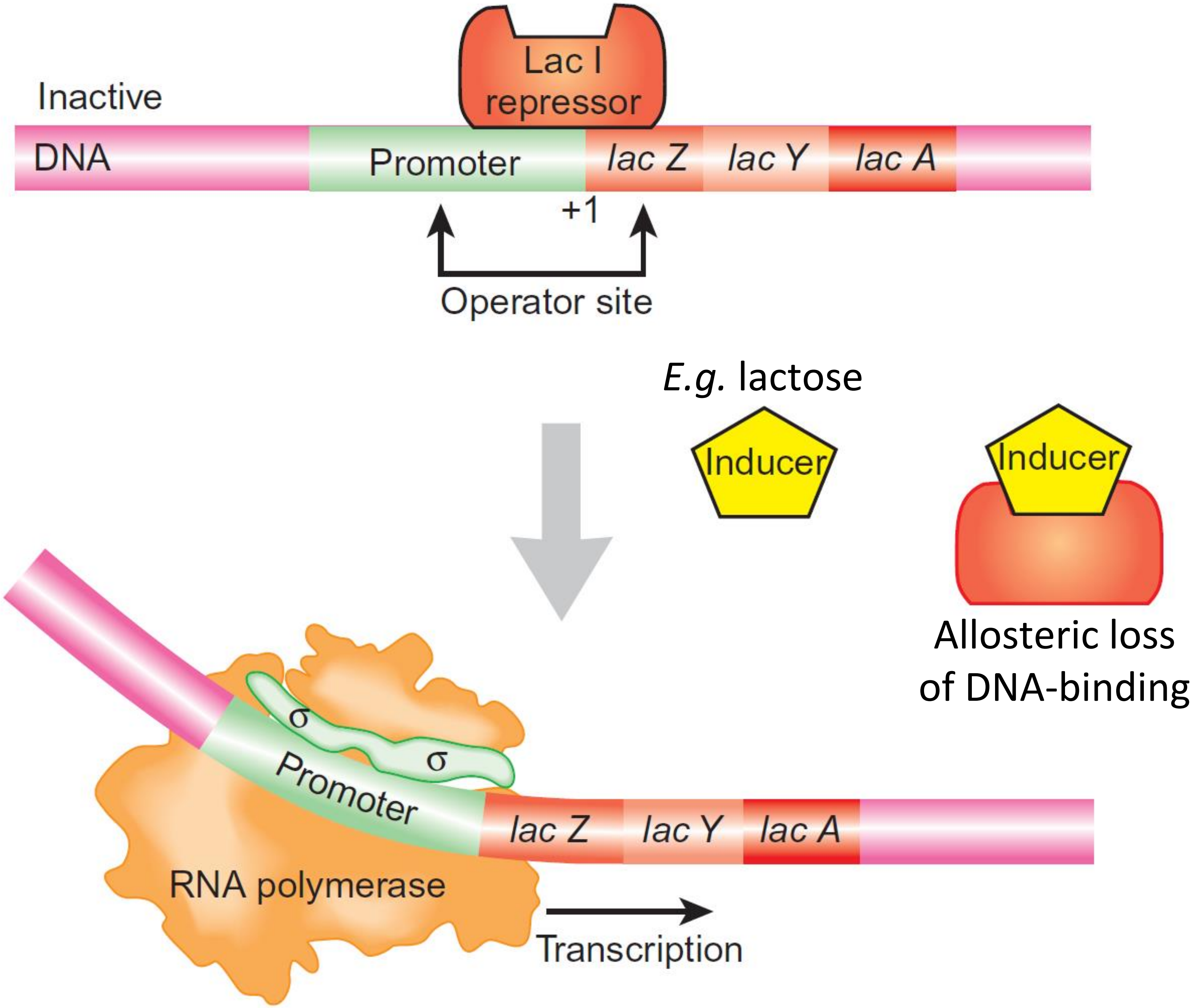
Regulatory promoter elements in bacteria

Positive regulation *via* activators

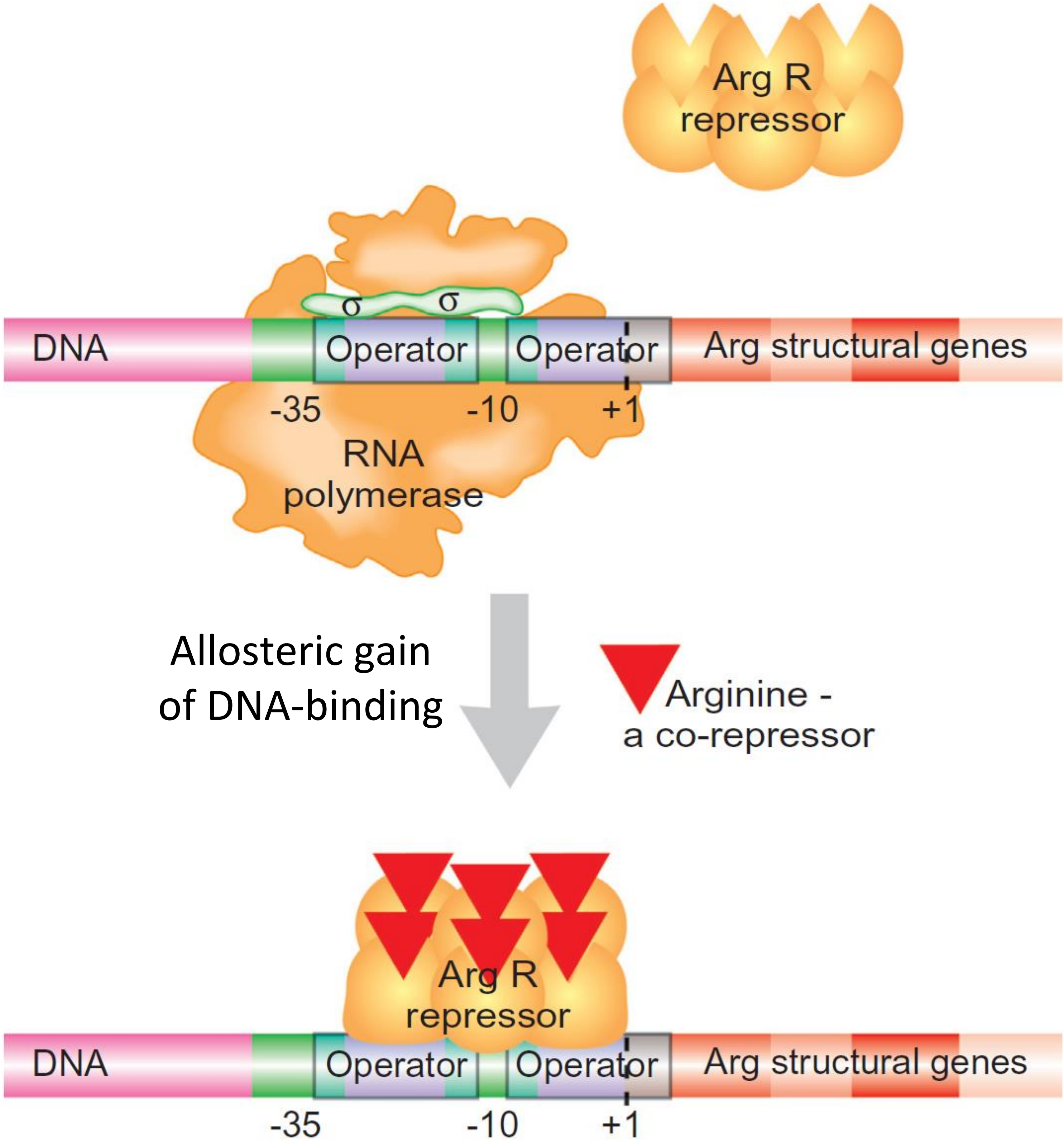


Negative regulation *via* repressors

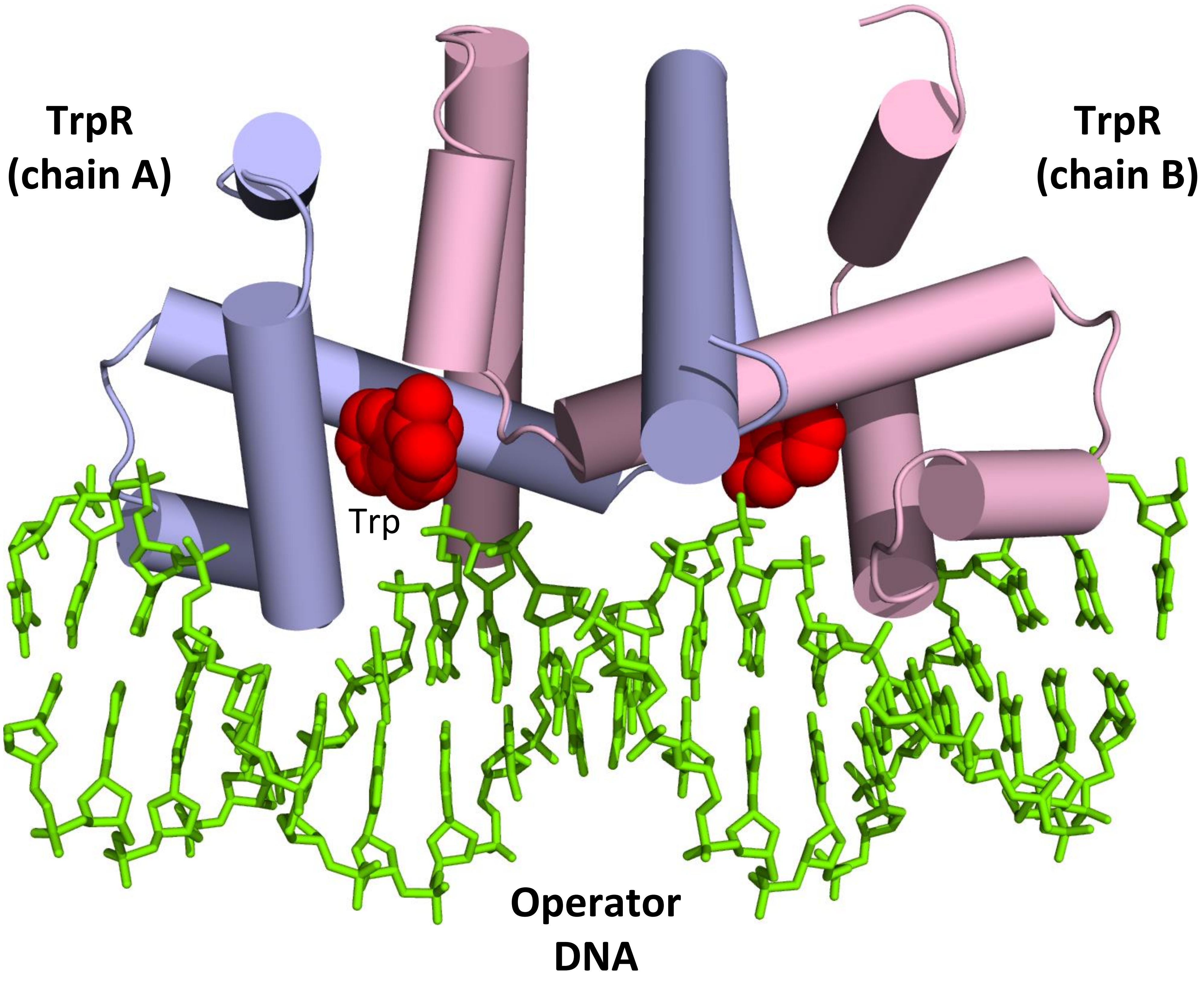
The promoter of the Lac operon



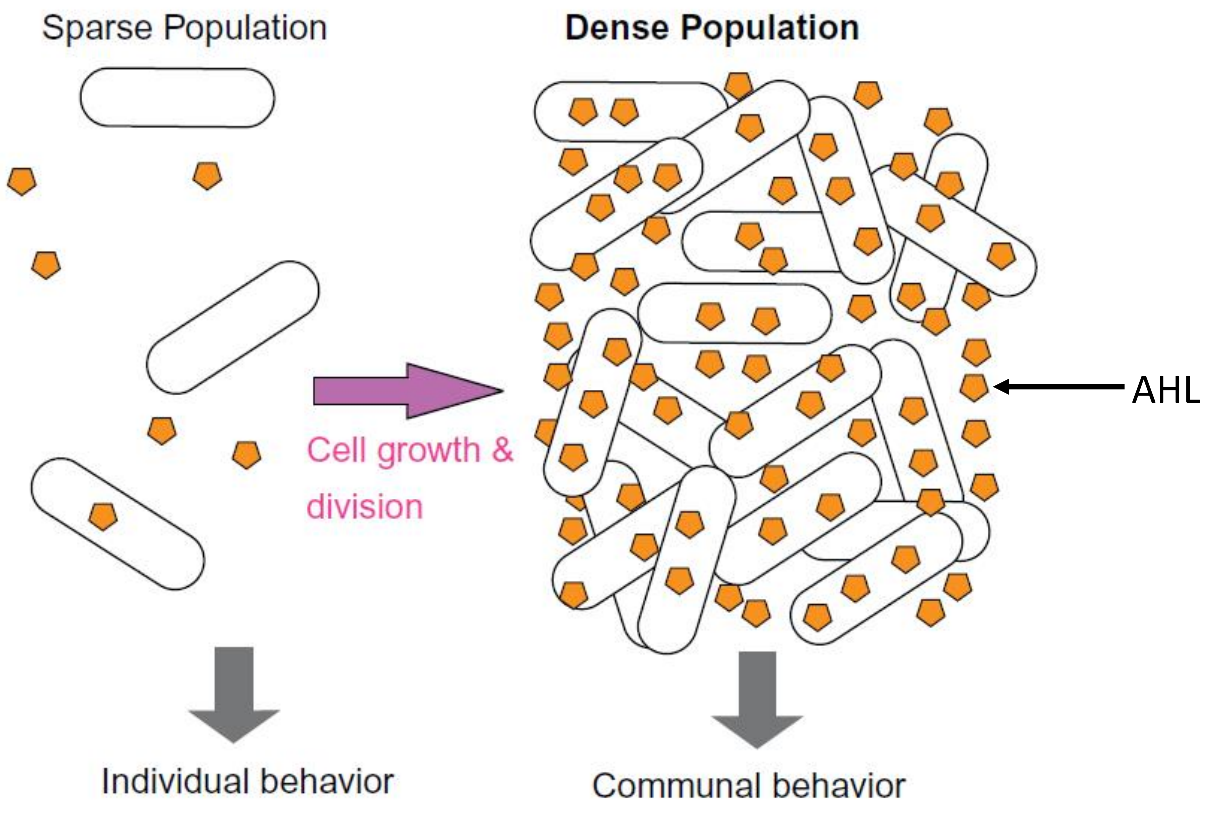
The promoter of the Arg operon

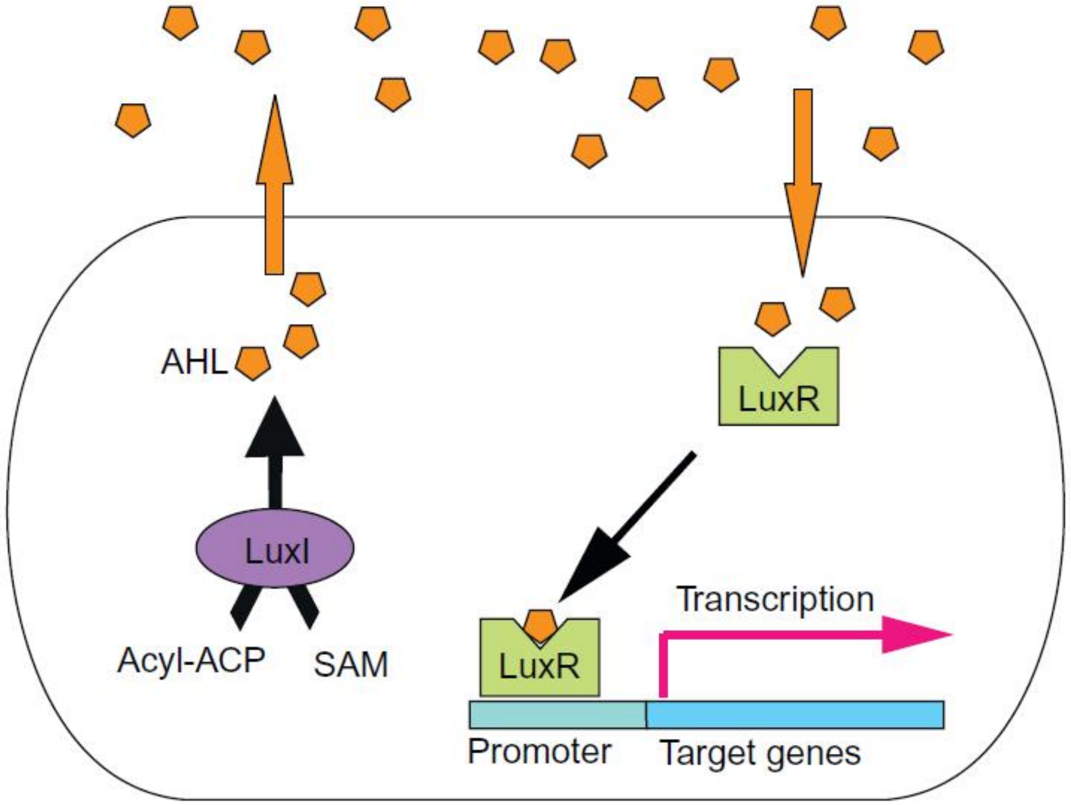


Example of an allosteric repressor: TrpR

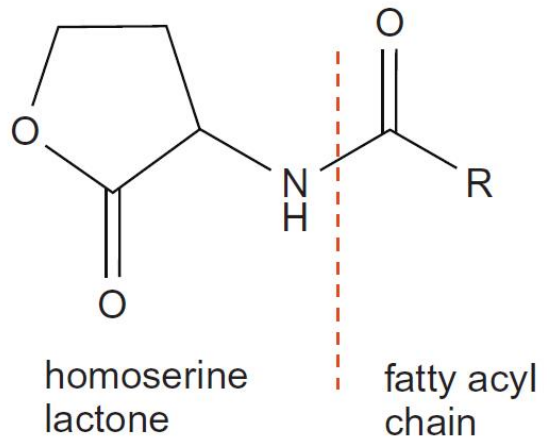


Quorum sensing in bacteria

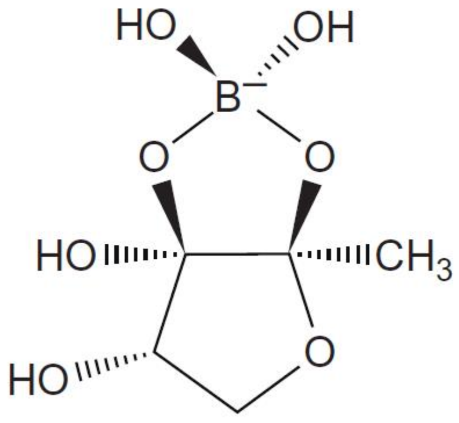




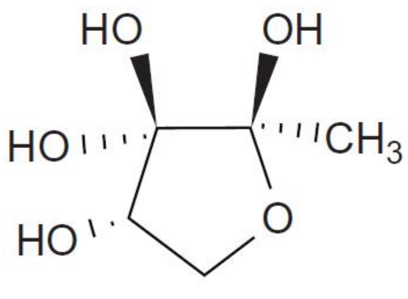
N-acyl homoserine lactone
“Autoinducer I”



Autoinducer 2



AI-2 of *Vibrio*



AI-2 of *Salmonella*

Promoters and regulatory elements in eukaryotes

Three different types of eukaryotic RNA polymerase complexes exist:

RNA polymerase I: transcription of ribosomal RNA

<u>**RNA polymerase II**</u>: transcription of protein-encoding genes

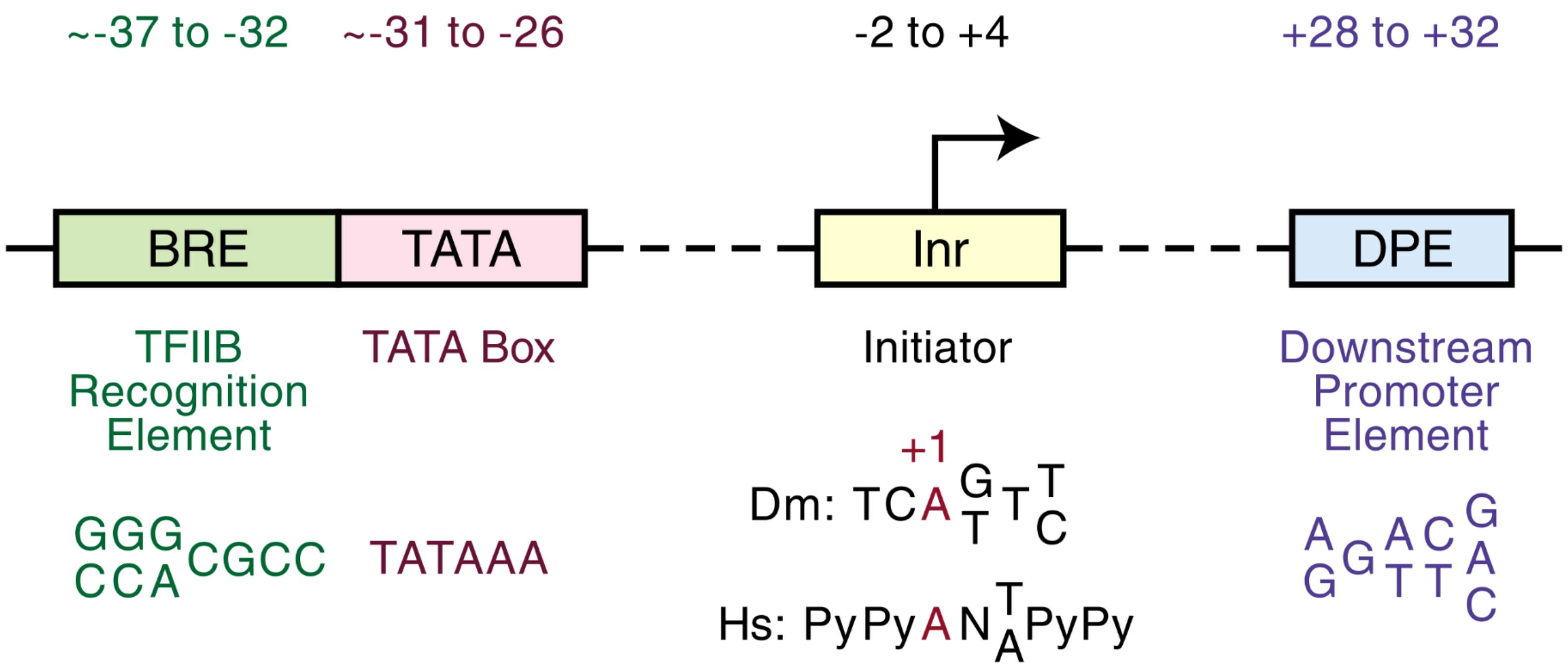
RNA polymerase III: transcription of tRNAs

Each eukaryotic gene has its own unique promoter; particularly RNA polymerase II promoters can be <u>extremely complex</u>

RNA polymerase II promoters

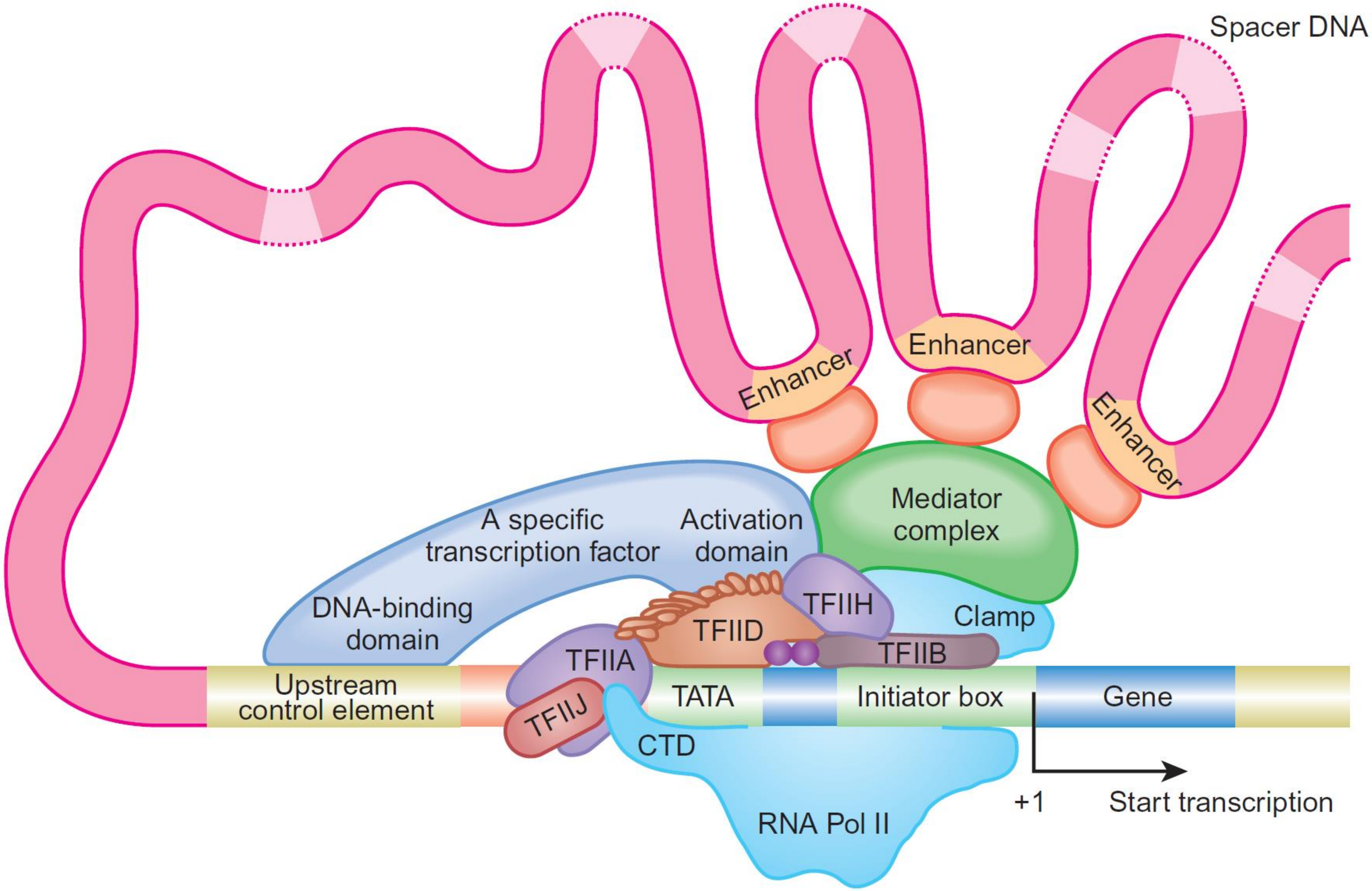
- Eukaryotic RNA polymerase II cannot directly bind to promoters, but relies on a dozen or more transcription factors to guide and position it on the DNA

- Core promoter elements (not all need be present!):



- Regulatory elements can be near the core promoter but may also be thousands of base pairs away (in so-called *enhancers*)

A “simple” eukaryotic promoter



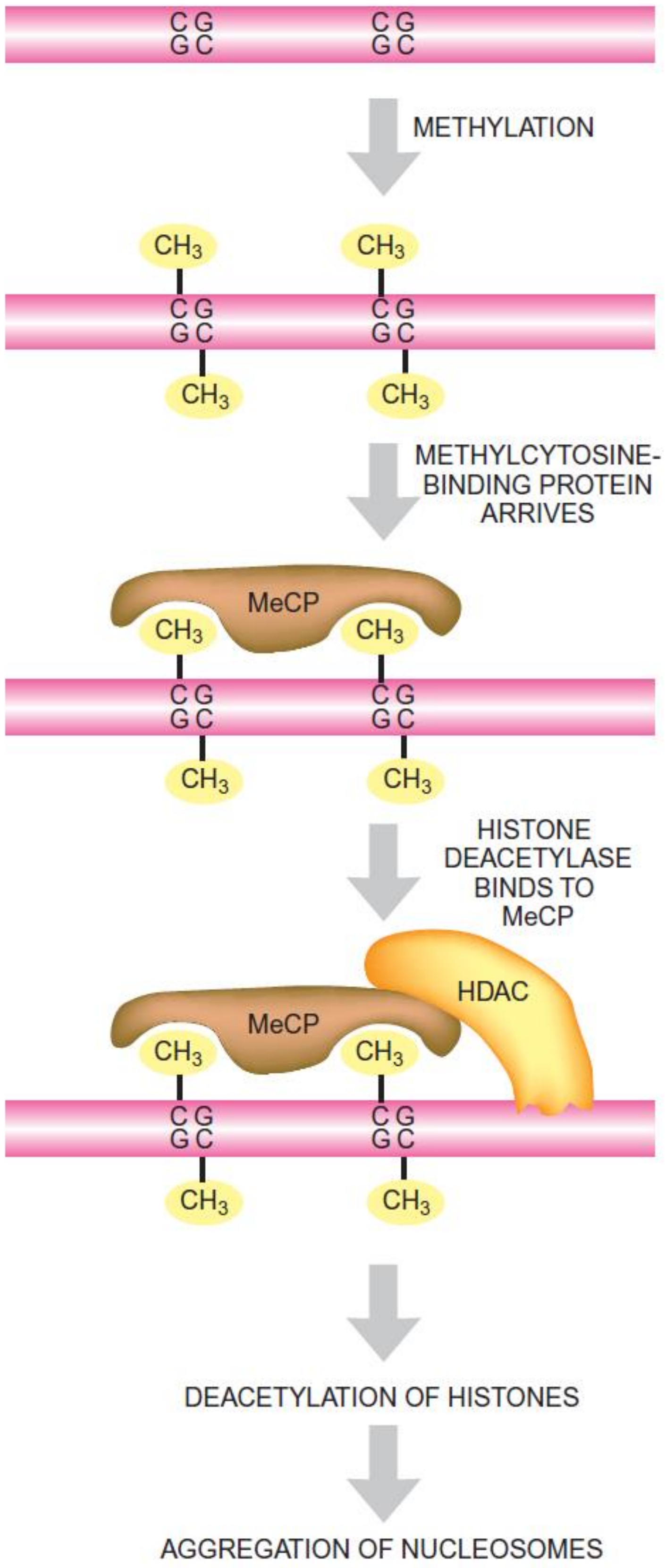
Eukaryotic promoters are often flanked by CpG islands

CpG islands can be methylated by regulatory methyltransferases

Factors that recognise methylated DNA recruit histone deacetylases

Deacetylases deacetylate histones

Deacetylation of histones leads to formation of *heterochromatin*



Difficulties in computational promoter prediction

- Regulatory sequences are not always well-defined and can be quite divergent
- Each gene has a unique combination of regulatory motifs
- Individual regulatory elements tend to be short (6-8 nucleotides): random chance of sequence similarity results in high rate of false positives
- Promoters cannot be translated into protein sequences to increase sensitivity of detection

Categories of prediction algorithms

- *Ab initio*: *de novo* predictions by scanning a genome sequence for a known pattern
- Similarity-based: predictions based on alignment of homologous sequences ("phylogenetic footprinting")
- Expression profile based: using profiles constructed from a number of co-expressed gene sequences from the same organism

Ab initio algorithms

- Prediction of prokaryotic/eukaryotic promoters and regulatory elements based on characteristic sequence patterns corresponding to known transcription factor recognition sites
- Examples: the -35/-10 boxes in bacteria and the TATA box in eukaryotes
- *A priori* knowledge about recognition sites is needed
- Impossible to discover new, unknown motifs
- Prediction programs are highly species-specific

Ab initio algorithms

The actual methods are very similar to those used in protein motif and domain searches (Chapter 7):

- Regular expressions, position-specific scoring matrices (PSSMs), Hidden Markov Models (HMMs), …
- Regular expression / PSSM / HMM constructed from well-characterized binding sites usually covering 6 to 10 bases
- Log-odds score evaluated for statistical significance in the case of a PSSM

Main problem and difference w.r.t. protein domain search: high rates of false positives due to much shorter sequence and (consequently) high chance of random sequence matches

Ab initio prediction of eukaryotic promoters

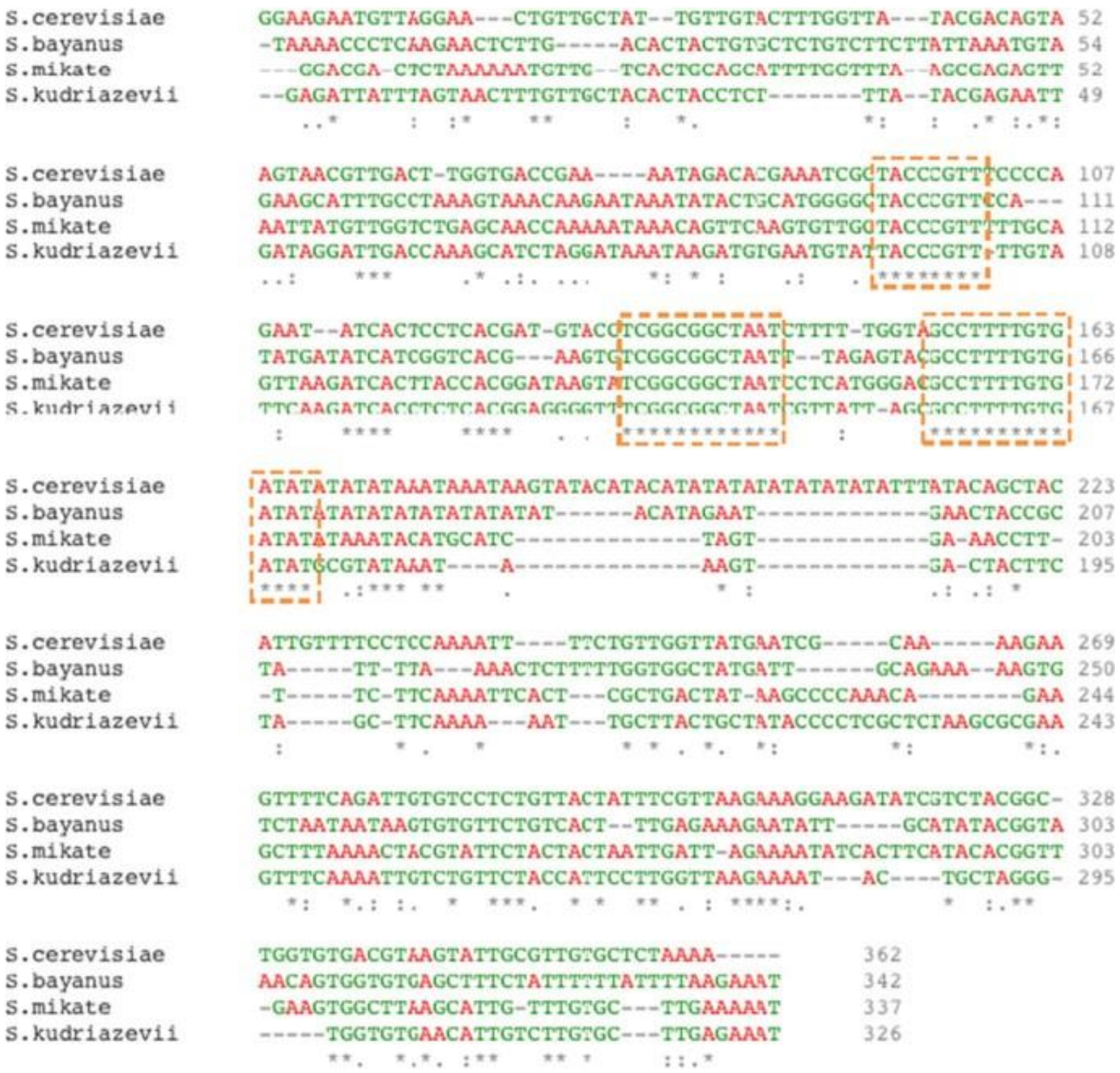
Even more complicated than prokaryotic prediction, but:

- Improved accuracy of prediction by taking into account the presence of CpG islands
  - Promoters can be found in the immediate vicinity of the islands
- Eukaryotic transcription initiation requires cooperation of a large number of transcription factors
  - Finding a cluster of transcription factor binding sites increases the probability that individual binding site prediction are correct

Phylogenetic footprinting

- Promoter and regulatory elements from <u>closely</u> related organisms such as human and mouse are highly conserved
- Promoter sequences for a particular gene are identified by aligning upstream regions between species
- Conserved non-coding DNA elements, called *phylogenetic footprints*, are likely to be transcription factor recognition sites

Phylogenetic footprinting: example



```
S.cerevisiae     GGAAGAATGTTAGGAA---CTGTTGCTAT--TGTTGTACTTTGGTTA---TACGACAGTA 52
S.bayanus        -TAAAACCCTCAAGAACTCTTG-----ACACTACTGTGCTCTGTCTTCTTATTAAATGTA 54
S.mikate         ---GGACGA-CTCTAAAAAATGTTG--TCACTGCAGCATTTTGGTTTA--AGCGAGAGTT 52
S.kudriazevii    --GAGATTATTTAGTAACTTTGTTGCTACACTACCTCT-------TTA--TACGAGAATT 49
                    ..*     :  :*    **     :   *.            *:   :  .* :.*:

S.cerevisiae     AGTAACGTTGACT-TGGTGACCGAA----AATAGACACGAAATCGCTACCCGTTTCCCCA 107
S.bayanus        GAAGCATTTGCCTAAAGTAAACAAGAATAAATATACTGCATGGGGCTACCCGTTCCA--- 111
S.mikate         AATTATGTTGGTCTGAGCAACCAAAAATAAACAGTTCAAGTGTTGGTACCCGTTTTTGCA 112
S.kudriazevii    GATAGGATTGACCAAAGCATCTAGGATAAATAAGATGTGAATGTATTACCCGTT-TTGTA 108
                  ..:   ***      .* .:. ...    *: * :    .:   . ********

S.cerevisiae     GAAT--ATCACTCCTCACGAT-GTACCTCGGCGGCTAATCTTTT-TGGTAGCCTTTTGTG 163
S.bayanus        TATGATATCATCGGTCACG---AAGTGTCGGCGGCTAATT--TAGAGTACGCCTTTTGTG 166
S.mikate         GTTAAGATCACTTACCACGGATAAGTATCGGCGGCTAATCCTCATGGGACGCCTTTTGTG 172
S.kudriazevii    TTCAAGATCACCTCTCACGGAGGGGTTTCGGCGGCTAATCGTTATT-AGCGCCTTTTGTG 167
                  :    ****     ****   . .  ************    :      **********

S.cerevisiae     ATATATATATAAATAAATAAGTATACATACATATATATATATATATATTTATACAGCTAC 223
S.bayanus        ATATATATATATATATATATAT------ACATAGAAT-------------GAACTACCGC 207
S.mikate         ATATATAAATACATGCATC--------------TAGT-------------GA-AACCTT- 203
S.kudriazevii    ATATGCGTATAAAT----A--------------AAGT-------------GA-CTACTTC 195
                 ****  .:*** **    .               * :             .: .: *

S.cerevisiae     ATTGTTTTCCTCCAAAATT----TTCTGTTGGTTATGAATCG-----CAA-----AAGAA 269
S.bayanus        TA-----TT-TTA---AAACTCTTTTTGGTGGCTATGATT------GCAGAAA--AAGTG 250
S.mikate         -T-----TC-TTCAAAATTCACT---CGCTGACTAT-AAGCCCCAAACA--------GAA 244
S.kudriazevii    TA-----GC-TTCAAAA---AAT---TGCTTACTGCTATACCCCTCGCTCTAAGCGCGAA 243
                  :        * .   *          * * . *.   *:       *:        *:.

S.cerevisiae     GTTTTCAGATTGTGTCCTCTGTTACTATTTCGTTAAGAAAGGAAGATATCGTCTACGGC- 328
S.bayanus        TCTAATAATAAGTGTGTTCTGTCACT--TTGAGAAAGAATATT-----GCATATACGGTA 303
S.mikate         GCTTTAAAACTACGTATTCTACTACTAATTGATT-AGAAAATATCACTTCATACACGGTT 303
S.kudriazevii    GTTTCAAAATTGTCTGTTCTACCATTCCTTGGTTAAGAAAAT---AC----TGCTAGGG- 295
                   *:  *.: :.  *  ***.  * *  **  . : ****:.          *  :.**

S.cerevisiae     TGGTGTGACGTAAGTATTGCGTTGTGCTCTAAAA-----       362
S.bayanus        AACAGTGGTGTGAGCTTTCTATTTTTTATTTTAAGAAAT       342
S.mikate         -GAAGTGGCTTAAGCATTG-TTTGTGC---TTGAAAAAT       337
S.kudriazevii    -----TGGTGTGAACATTGTCTTGTGC---TTGAGAAAT       326
                      **.  *.*. :**   ** *     ::.*
```

Phylogenetic footprinting

Phylogenetic footprinting requires sequences from moderately divergent species:

- If the organisms selected are too closely related (*e.g.* human and chimpanzee), the sequence differences may not be sufficient to reveal functional elements
- If evolutionary distances are too large (*e.g.* human and yeast), promoter and other elements are no longer conserved
- *E.g.* human and mouse (vertebrate) sequences often yield informative results

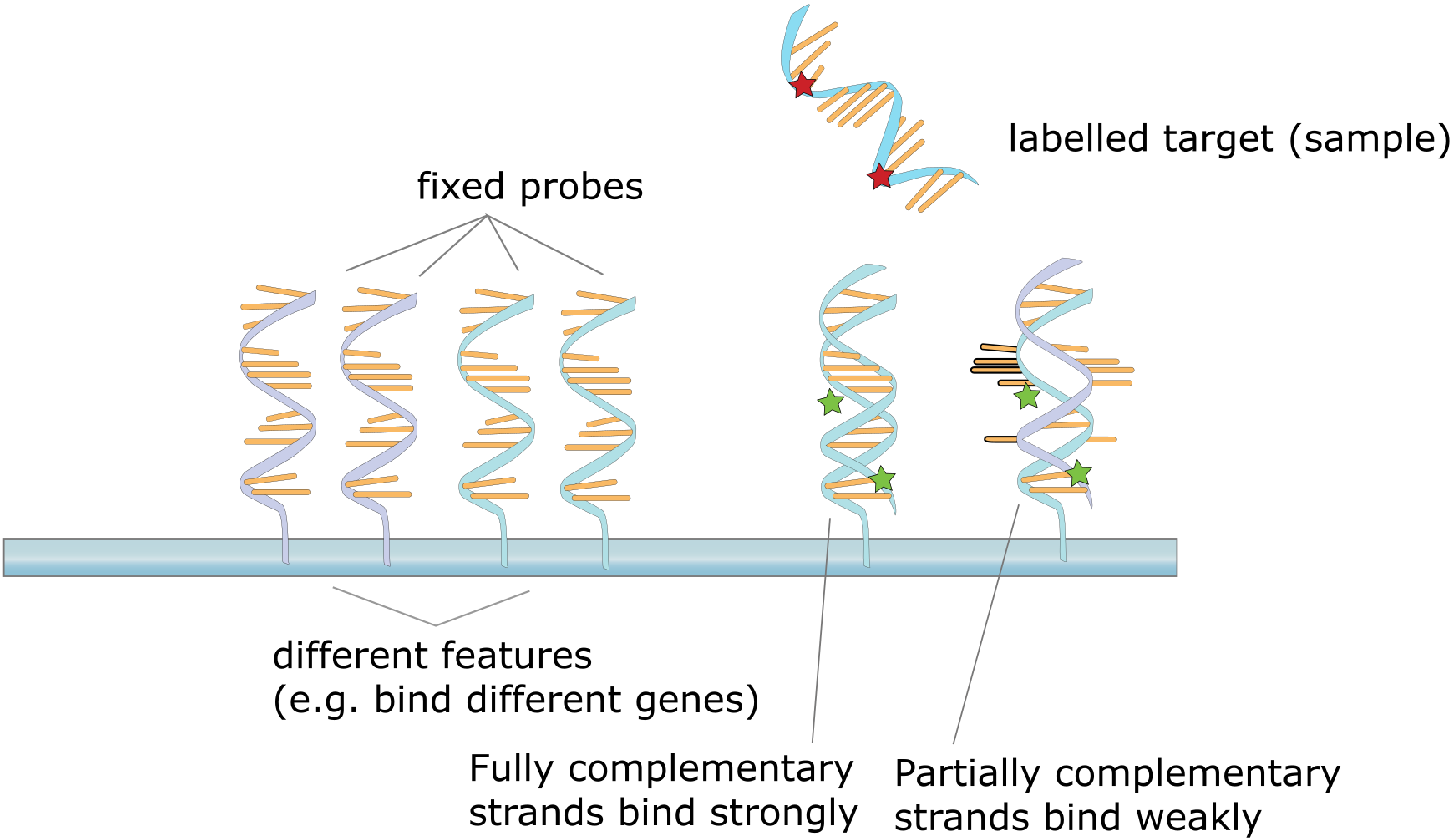
Phylogenetic footprinting

- Predictive value depends on the quality of the sequence alignments
- No training of a model is required, hence broadly applicable
- Potential to discover new regulatory motifs shared among organisms

Expression profiling based method

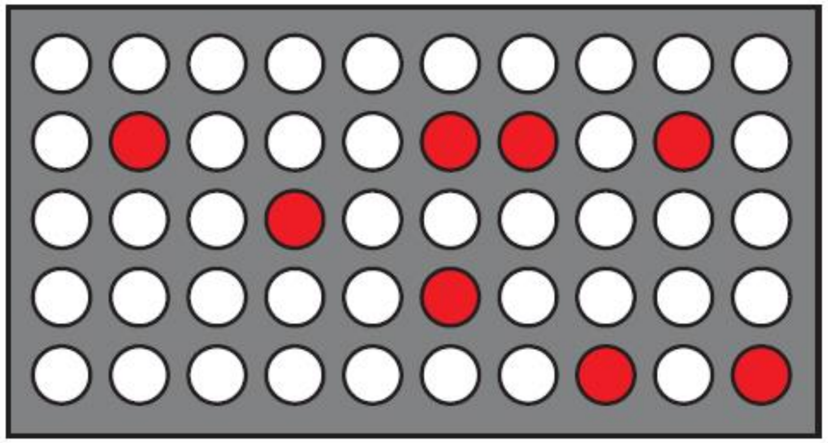
- *DNA microarray methods* and *RNA-Seq* allow simultaneous monitoring of expression levels of thousands of genes

Microarray analysis of gene expression levels

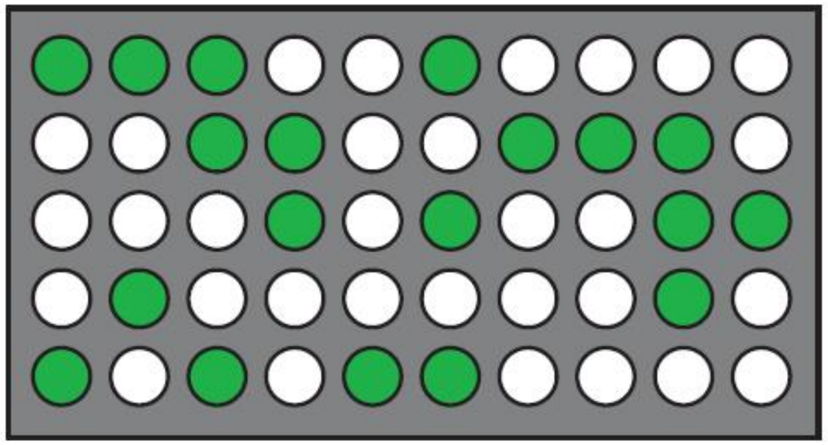


Microarray analysis of gene expression levels

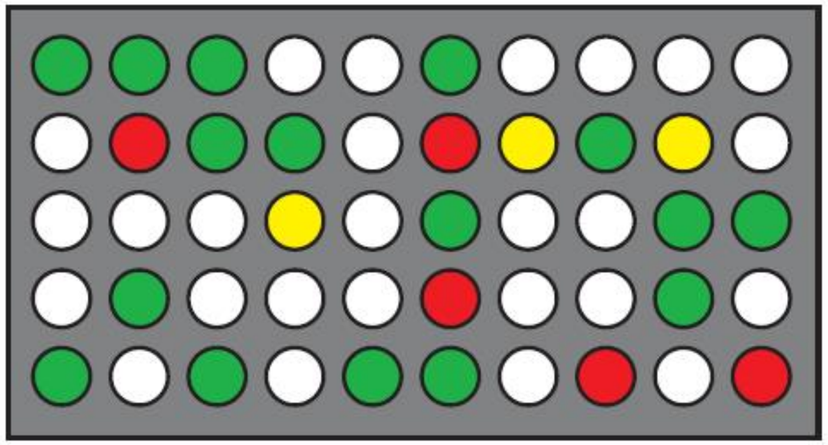
RNA isolated from cells grown in **condition 1** and labelled with red fluorescent dye



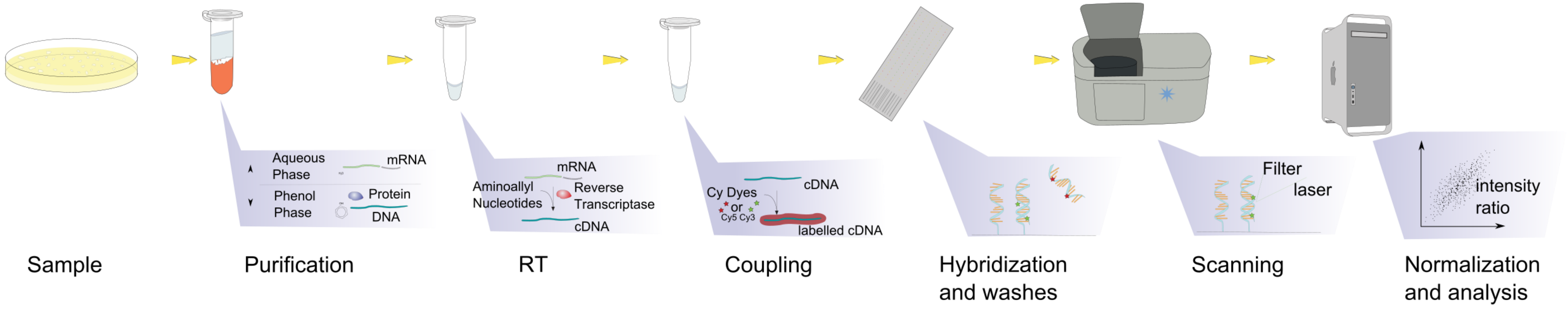
RNA isolated from cells grown in **condition 2** and labelled with green fluorescent dye

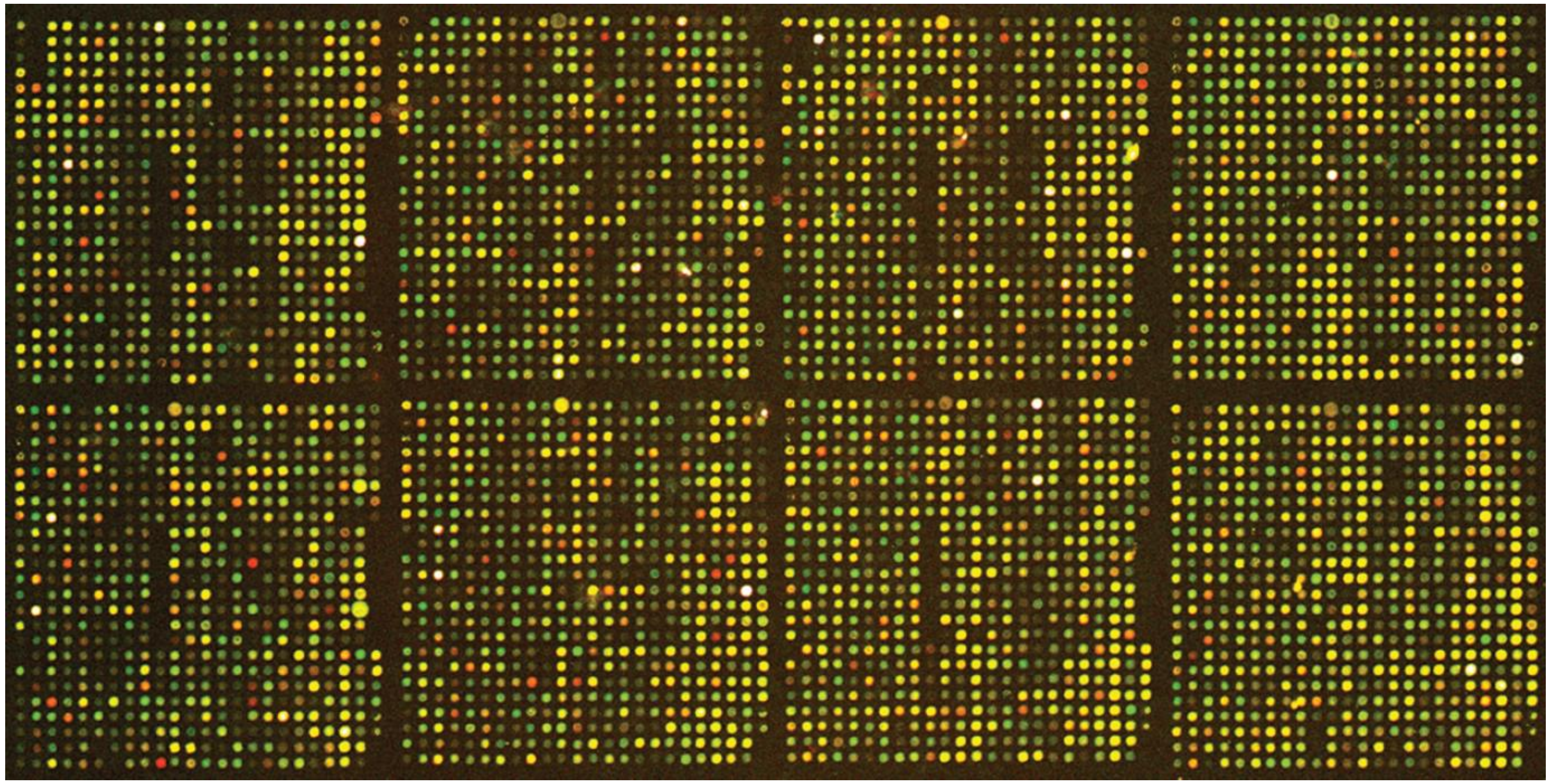


RNA from both samples



Microarray analysis of gene expression levels





Expression profiling based method

- *DNA microarray methods* (as well as *RNA-Seq*) allow monitoring of expression levels of thousands of genes simultaneously
- Genes with similar expression profiles are considered "co-expressed"
- It is assumed that co-expression is due to common promoters and regulatory elements
- Upstream sequence of co-expressed genes is aligned to reveal common regulatory elements

Expression profiling based method: problems

- Identification of co-expressed genes is error-prone (depends on clustering approaches)
- Co-expression can also be caused by parallel signalling pathways and distinct transcription regulatory mechanisms

In conclusion…

- Identification of promoters and regulatory elements, especially in eukaryotes, essentially remains an unsolved problem
- Prediction results may nonetheless be helpful, but should really be treated as hypotheses
- Experimental verification remains essential
- Focus on specific regions (non-coding sequences upstream of genes) to prevent false positives