Genome Analysis Lectorial

Atefeh Taherian Fard

Australian Institute for Bioengineering and Nanotechnology

The University of Queensland

BINF6000 I SCIE2100 I Bioinformatics I – Introduction

Summary from the Two Genome Analysis lectures

- **Lecture 1:**
- Overview genome sequencing and sequencing technologies
- Genome re-sequencing
- De-novo genome assembly

- **Lecture 2:**
- Gene features in prokaryotes
- Gene features in eukaryotes
- Computational approaches for gene prediction
- Functional genome annotation

Outline for today's Lectorial

- Overview of De novo and genome re-sequencing
- Identifying structural variations
- Gene feature in prokaryote and eukaryotes
- Computational approaches for gene finding
  - Support Vector Machine (SVM)
  - Hidden Markov model
- Past exam questions

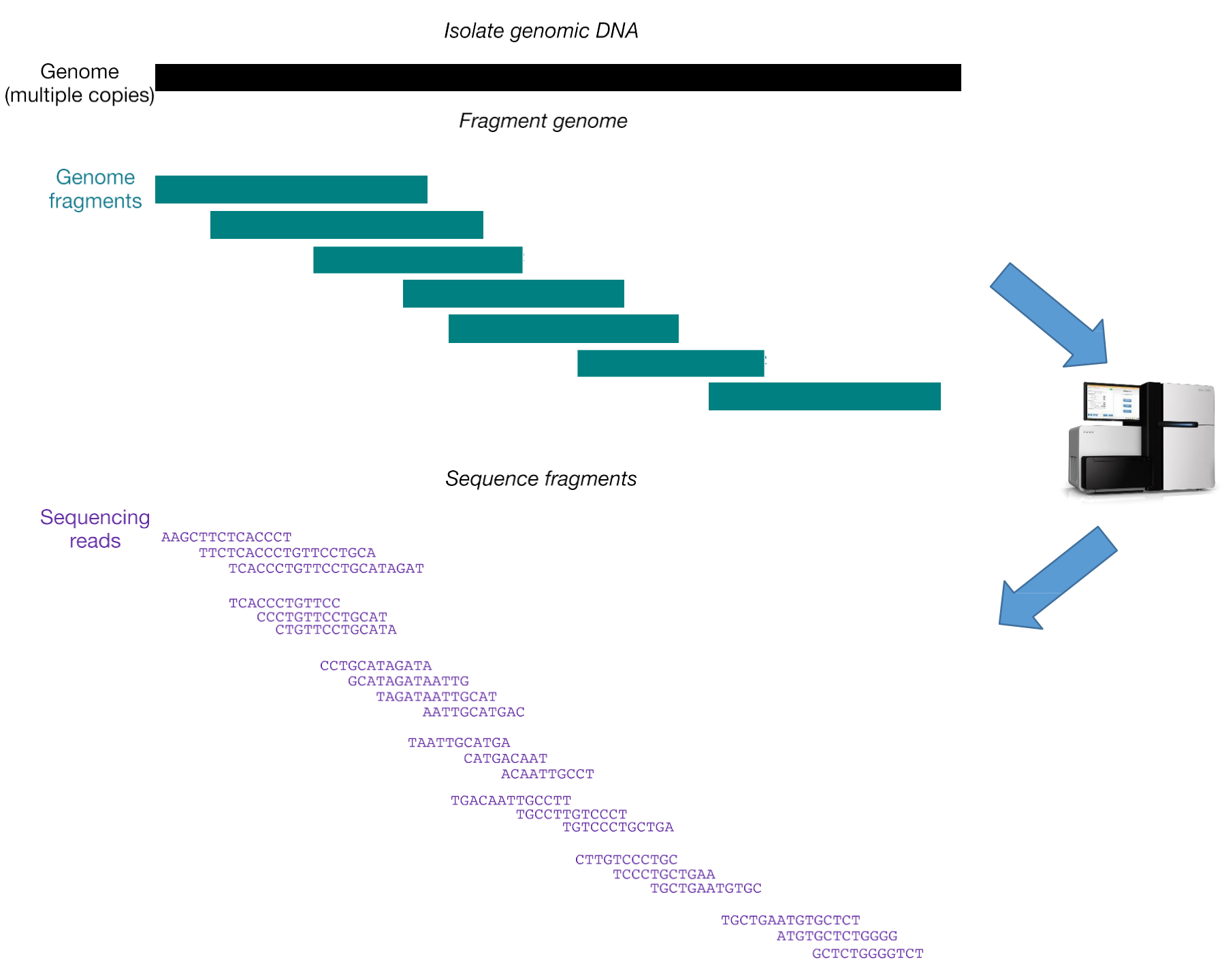
Why Do We Sequence Genomes?

Genome resequencing:

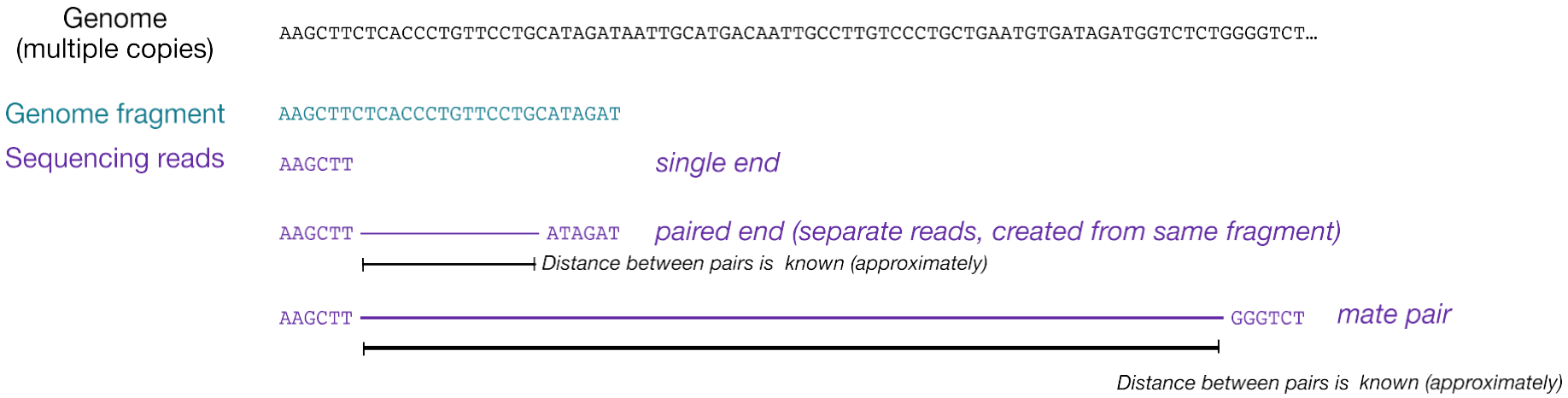
- Characterise genotype-phenotype associations
- Understand genetics of complex diseases
- Genome-based diagnosis; non-invasive prenatal testing
- Personalised medicine, genome-based prediction of optimal treatment (e.g. targetable mutations in cancer) and side-effects. Future: optimal diet, metabolic deficiencies, disease risk, …

De-novo sequencing

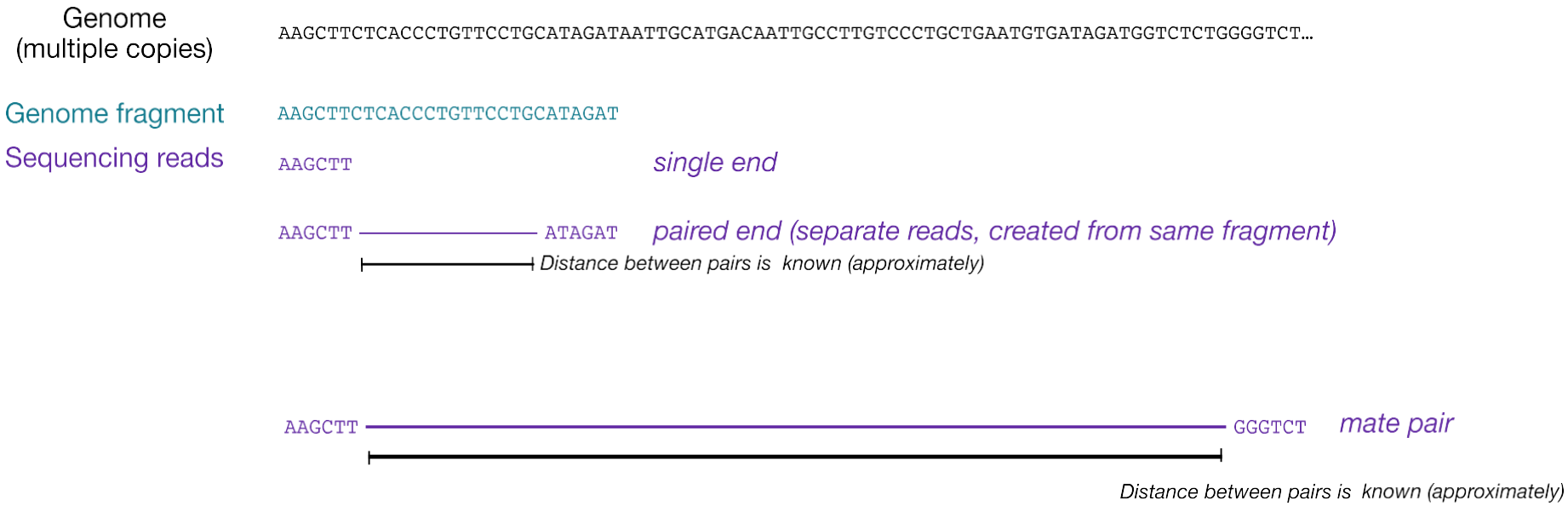
- Understand molecular biology of organisms, identify genes, gene functions, encoded pathways, metabolic capabilities, gene regulation and genome evolution



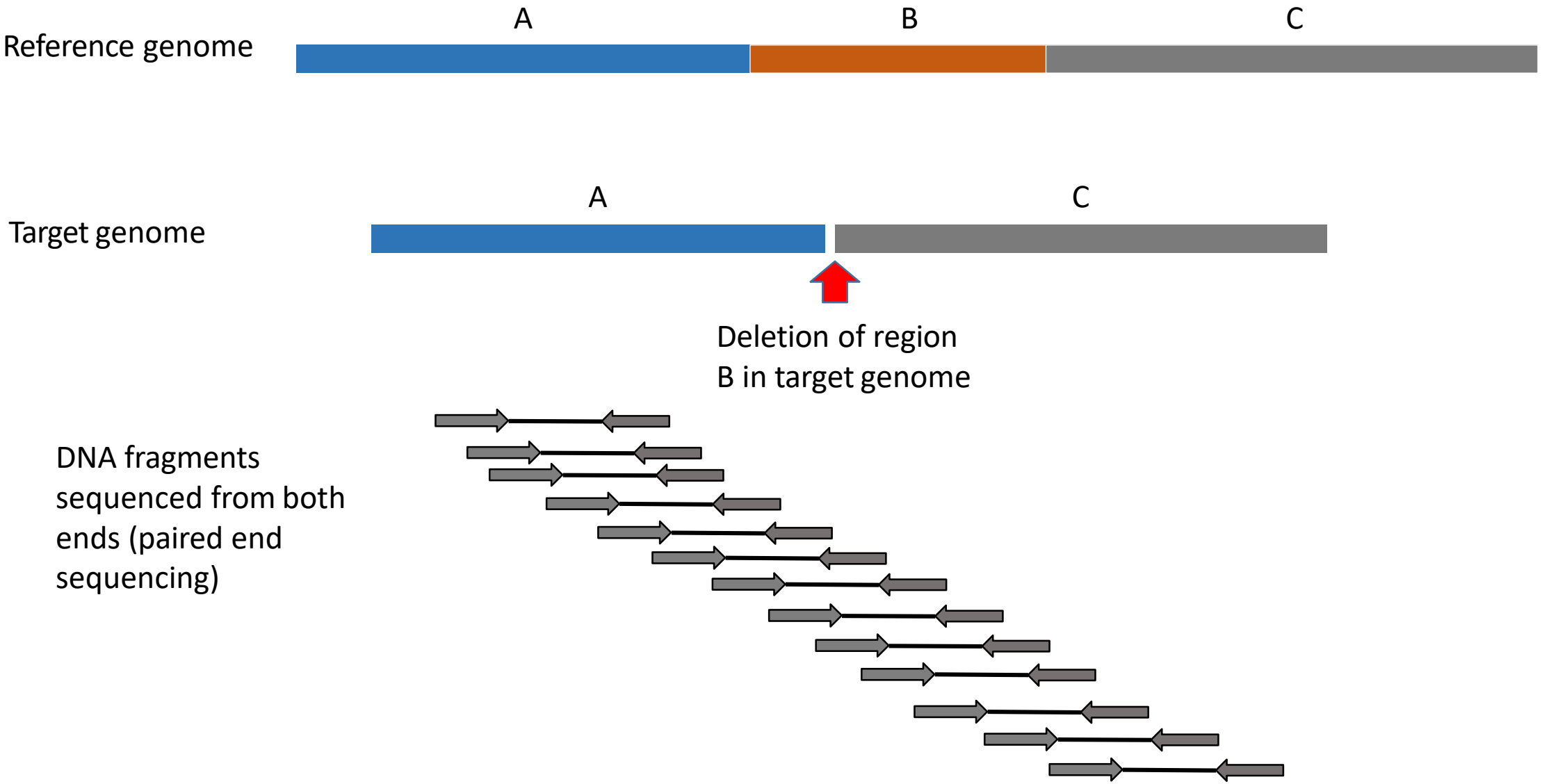
This approach is called ‘Shot Gun’ sequencing


Genome
(multiple copies)
AAGCTTCTCACCCTGTTCCTGCATAGATAATTGCATGACAATTGCCTTGTCCCTGCTGAATGTGATAGATGGTCTCTGGGGTCT…
Genome fragment
AAGCTTCTCACCCTGTTCCTGCATAGAT
Sequencing reads
AAGCTT
single end
AAGCTT
ATAGAT
paired end (separate reads, created from same fragment)
Distance between pairs is known (approximately)
AAGCTT
GGGTCT
mate pair
Distance between pairs is known (approximately)

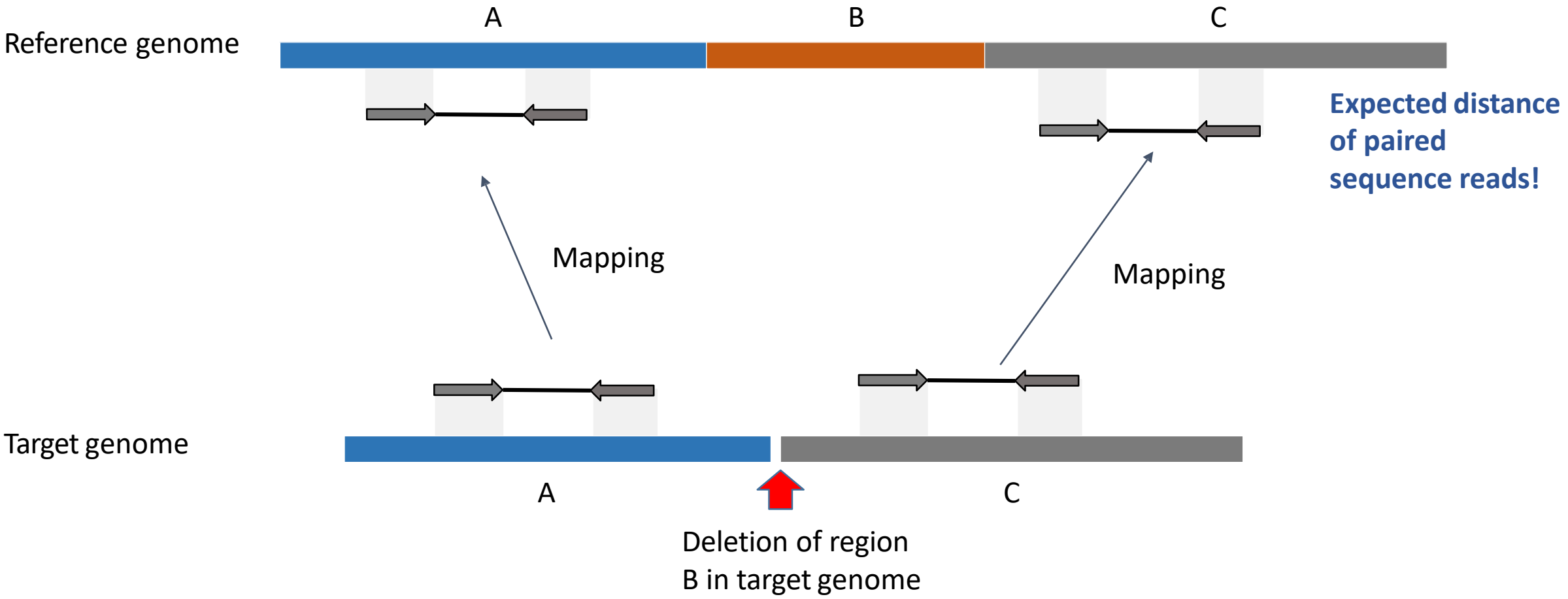

Genome
(multiple copies)
AAGCTTCTCACCCTGTTCCTGCATAGATAATTGCATGACAATTGCCTTGTCCCTGCTGAATGTGATAGATGGTCTCTGGGGTCT…
Genome fragment
AAGCTTCTCACCCTGTTCCTGCATAGAT
Sequencing reads
AAGCTT
ATAGAT
ATAGAT


Genome
(multiple copies)
AAGCTTCTCACCCTGTTCCTGCATAGATAATTGCATGACAATTGCCTTGTCCCTGCTGAATGTGATAGATGGTCTCTGGGGTCT…
Genome fragment
AAGCTTCTCACCCTGTTCCTGCATAGAT
Sequencing reads
AAGCTT
single end
AAGCTT
ATAGAT
paired end (separate reads, created from same fragment)
Distance between pairs is known (approximately)
AAGCTT
GGGTCT
mate pair
Distance between pairs is known (approximately)

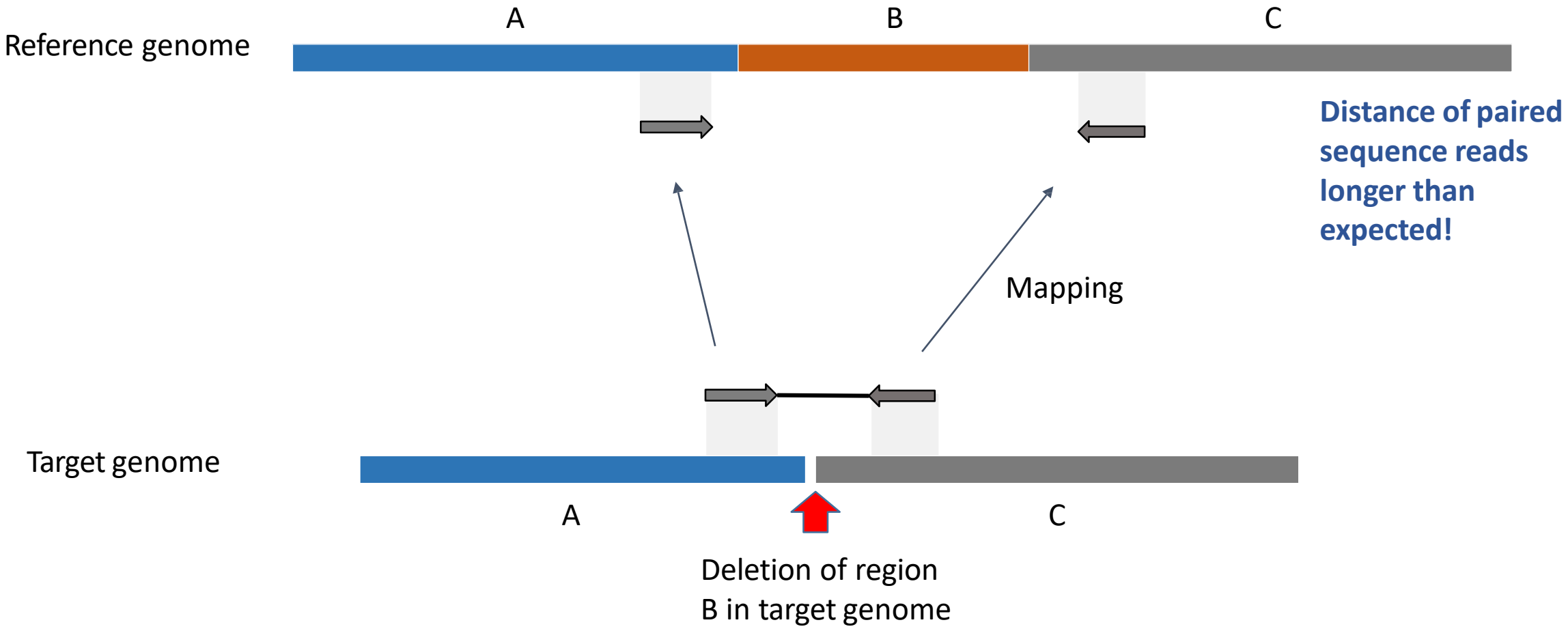
Structural Variations



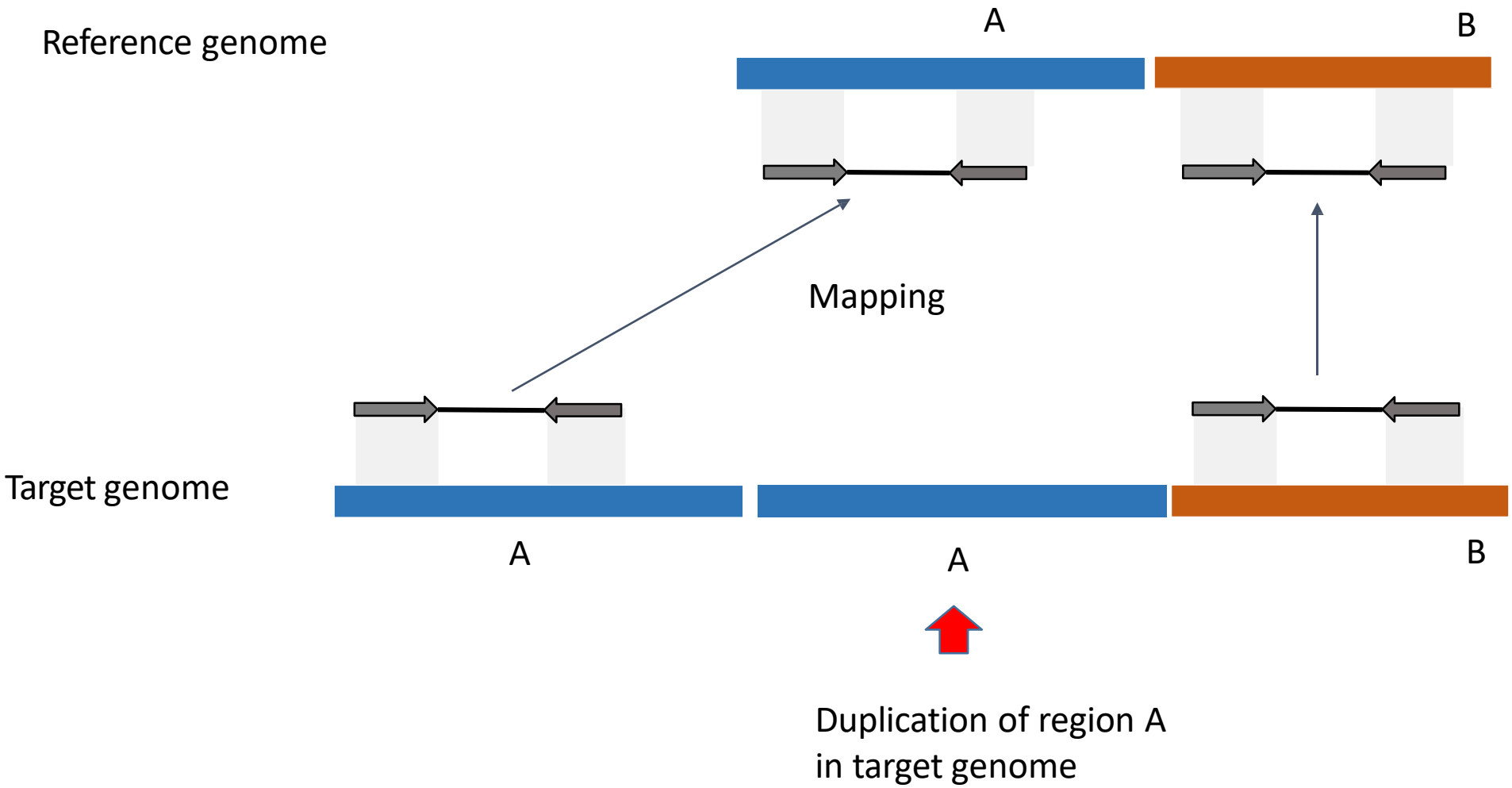
Structural Variations



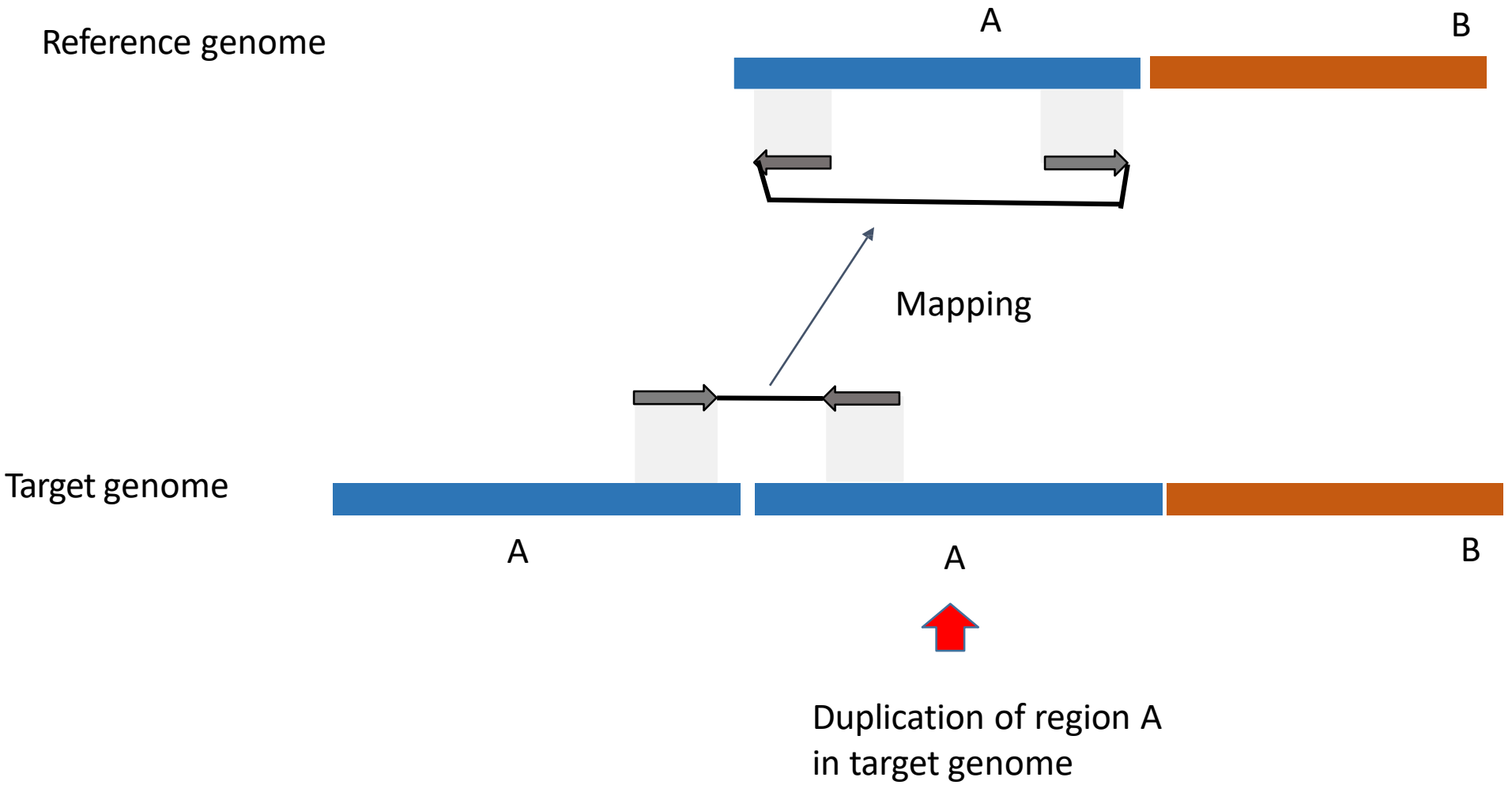
Deletion


Reference genome
A
B
C
Distance of paired sequence reads longer than expected!
Mapping
Target genome
A
C
Deletion of region
B in target genome

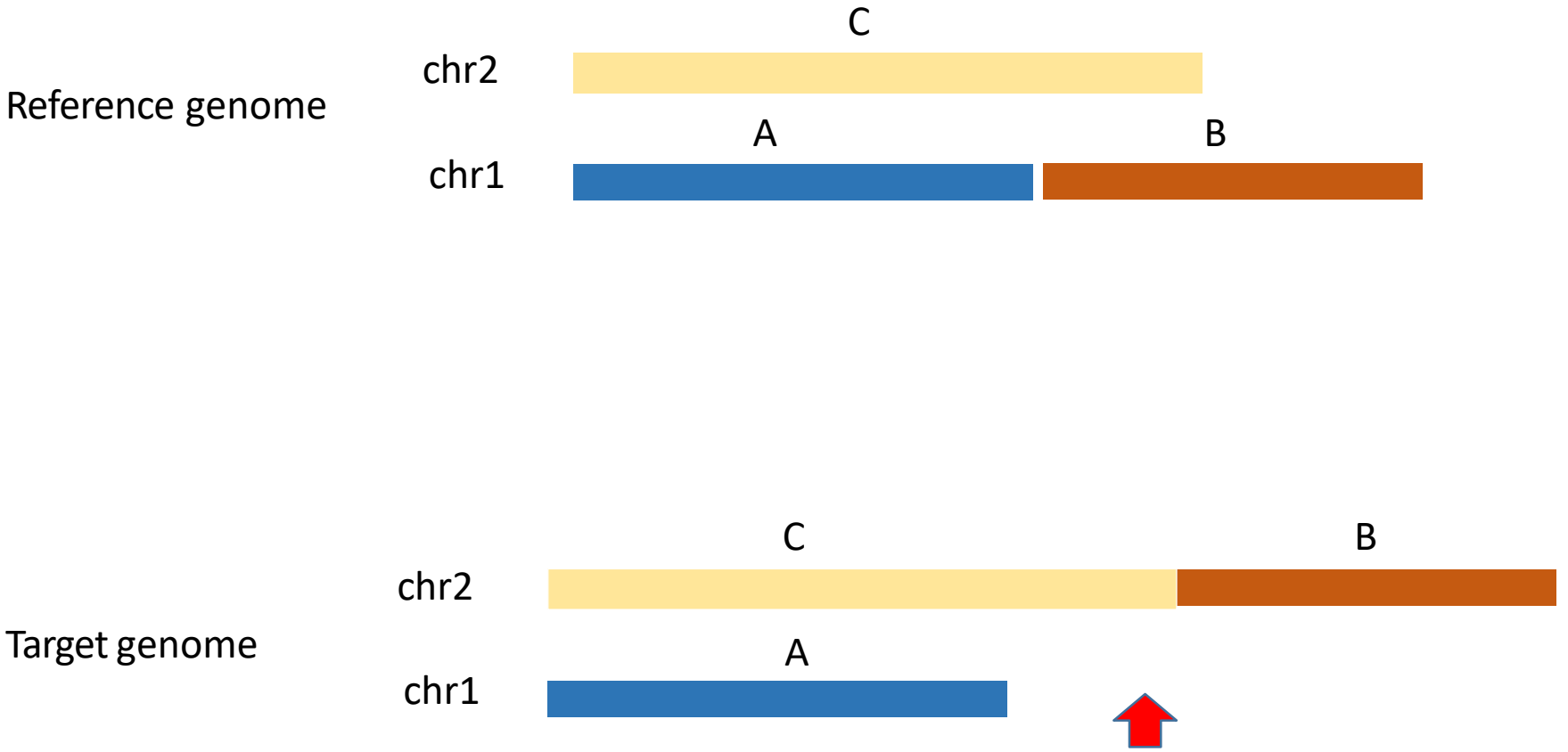
Duplications



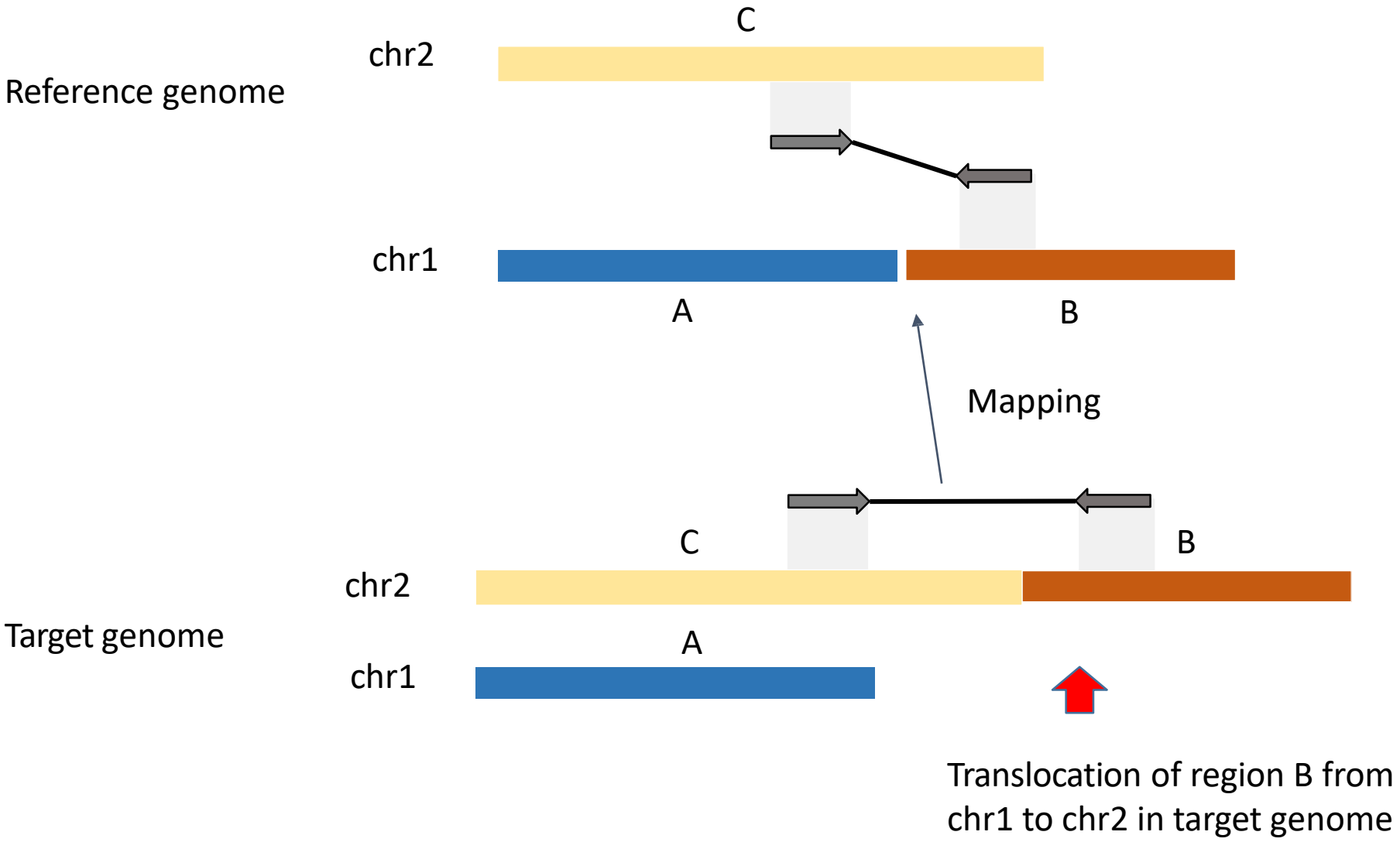
Duplications



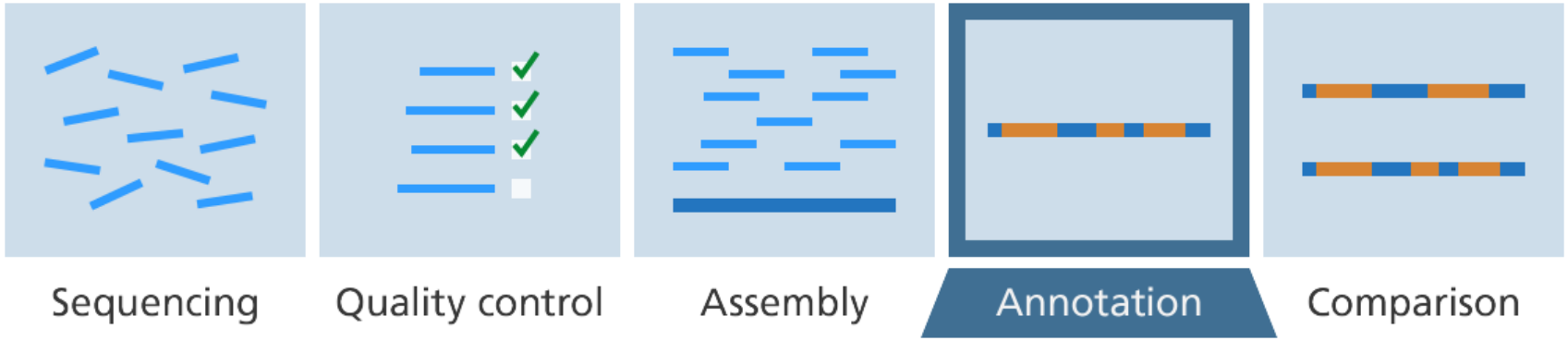
Translocations



Translocations



www.yourgenome.org



Using computational methods, find all genes (or other elements) in a long, unannotated string of nucleotides.

```
ACCGGTCAATAGCCGCAGACTACGGCATTTCAGAGGGACAGGCACTATAGCAACTAGCAACCCCCGTATAATACAAGGAGGCT
CAAGCTCCACTCTGACTCTCAACTTATTACGCTGTCACTCGATACGGCAGGGGCATTTAGACTTACGGCATATACCCGCCCGA
TCCAGCTTACGATACTACTGCTACTGGATACCCTGATAGCCAATCATTACGACTACTACTACGGCATTTCAGACCCGACAGGC
ACTAGAGCAACTAGCAACACCCGTATAATACAAGGAGGCTCAAGCTCCAGCTCTCACTGCAGCTATGTGGTGCACACATGTGC
ATCGTATGACTCAGTCGATGCTATCACGTACATCGTGTGGGTGCACACCACCCATGCCCTGATAGCCCCTGATTTTAGCCCCA
GCATTATTTTTCCGACGAGATCACGTACCCCCTACGGCATTTCAGAGGGACAGGGGACGCGCCCAATTACGACTACTACTACG
GCATTTCAGACCCGACAGGCACTAGAGCAACTAGCAACACCCGTATAATACAAGGAGGCTCAAGCTCCAGCCTTTTCAACAGA
CCGGGCGTTACGGTAAAAAAAAATCCGGCCGTACGGACTACTGGATACCGCAGACTACGGCATTTCAGAGGGACAGGCACTAT
AGCAACTAGCAACCCCCGTATAATACAAGGAGGCTCAAGCTCCACTCTGACTCTCAACTTATGACAGGGGACGATGACTCAGT
CGATTTCGCTATCACGTAAAACATCGTGTGGGGTGCACACCACCGCATGCCCTTTCAGGATAGCCCCTGATTTTAAGCCCCAG
CATTATTTGGTTCCGACGAGATCACGTACCCCACTACGGCATTTCAGAGGGACACTCAGTCGATGCTATCACGTACATCGTGT
GGGTGCTTACACCACGCCATGCCCTGATAGCCCCTGGGGATTTTAGCCCCAGCATTAATTTCTTCCGACGAGCCCTCAGACCC
GACAGGGGCACTAGAGCAACTATATAAGCAACACCCGTATACCCATACCAAGGAGGCTCAAGCTCCAGCCTTCTTCAACAGGA
CCGGGCGGGATTCCACATCATTCATGGGCAGCATCCCAGCAAACCCACGGCATAAGGACCACCCCTCGGCTAAGCAATCGCAT
AATACGGCGCTGCGCTCTACGTCTAGAGCTCACCATCTTACGAGGCTCTACCTCTATG...[3 BILLION or so MORE]
```

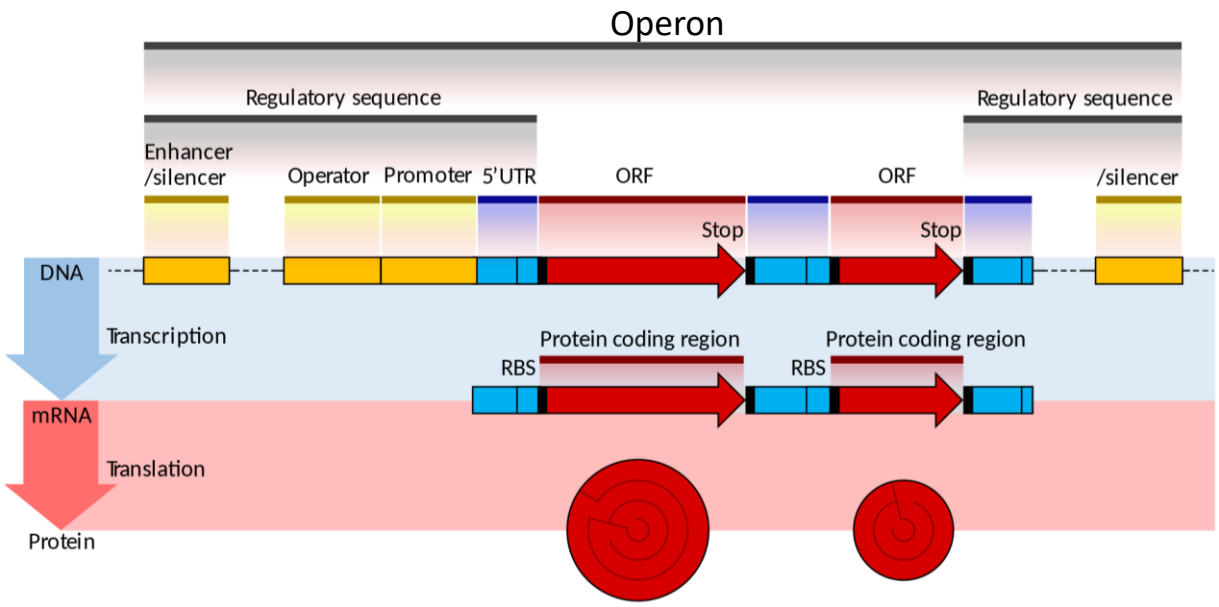
Aim: To identify transcriptional unit

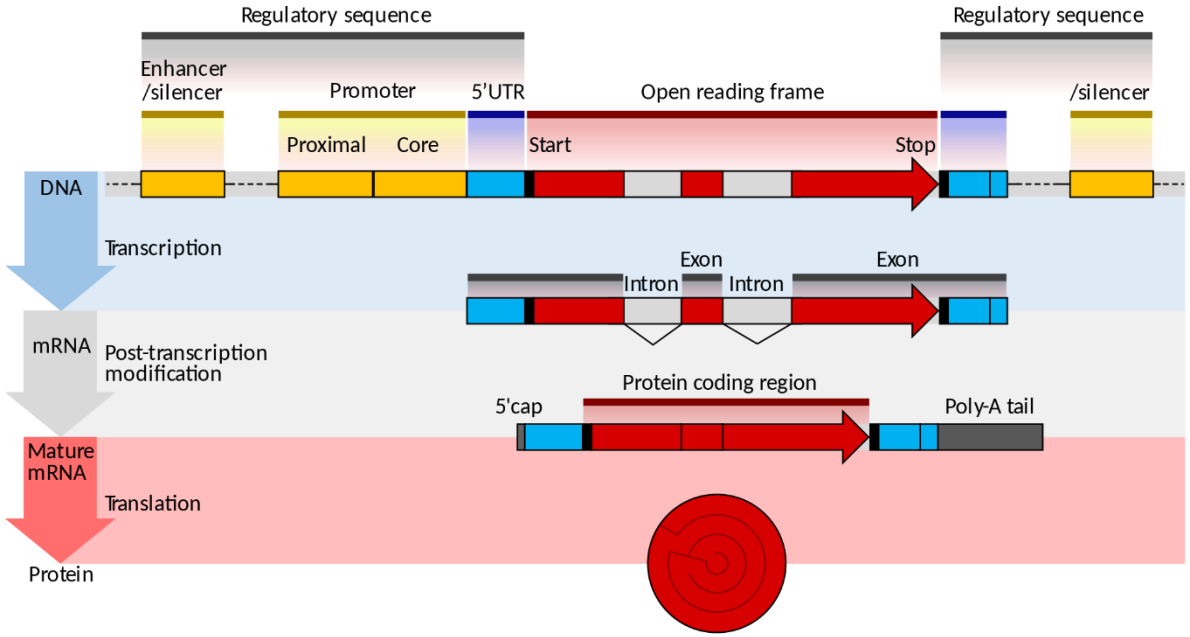
What do we know? Know only approximately what they look like

How? Find their locations and boundaries as accurately as possible, overlook as few as possible, and report as few non-genes as possible.

Adapted from Mark Ragan

Summary: Gene features in Prokaryotics vs Eukaryotes





- Codon bias and GC rich regions
- Transcriptional start and stop sites
- ORFs: Start and stop codons
- 5’ UTR: Ribosomal Binding Sequence site
- 3’UTR

- Codon bias and GC rich regions
- Promoter regions
- Intro and Exon splice site
- ORFs: Start and Stop codons
- 5’ UTR: 5’ Cap (G cap site)
- 3’ UTR: PolyAs

Shafee & Lowe (2017)

Gene finding Approaches

- Physical, genetic or other ***experimental approaches***
  - *e.g.* Genetic knockouts

- *Computational approaches*

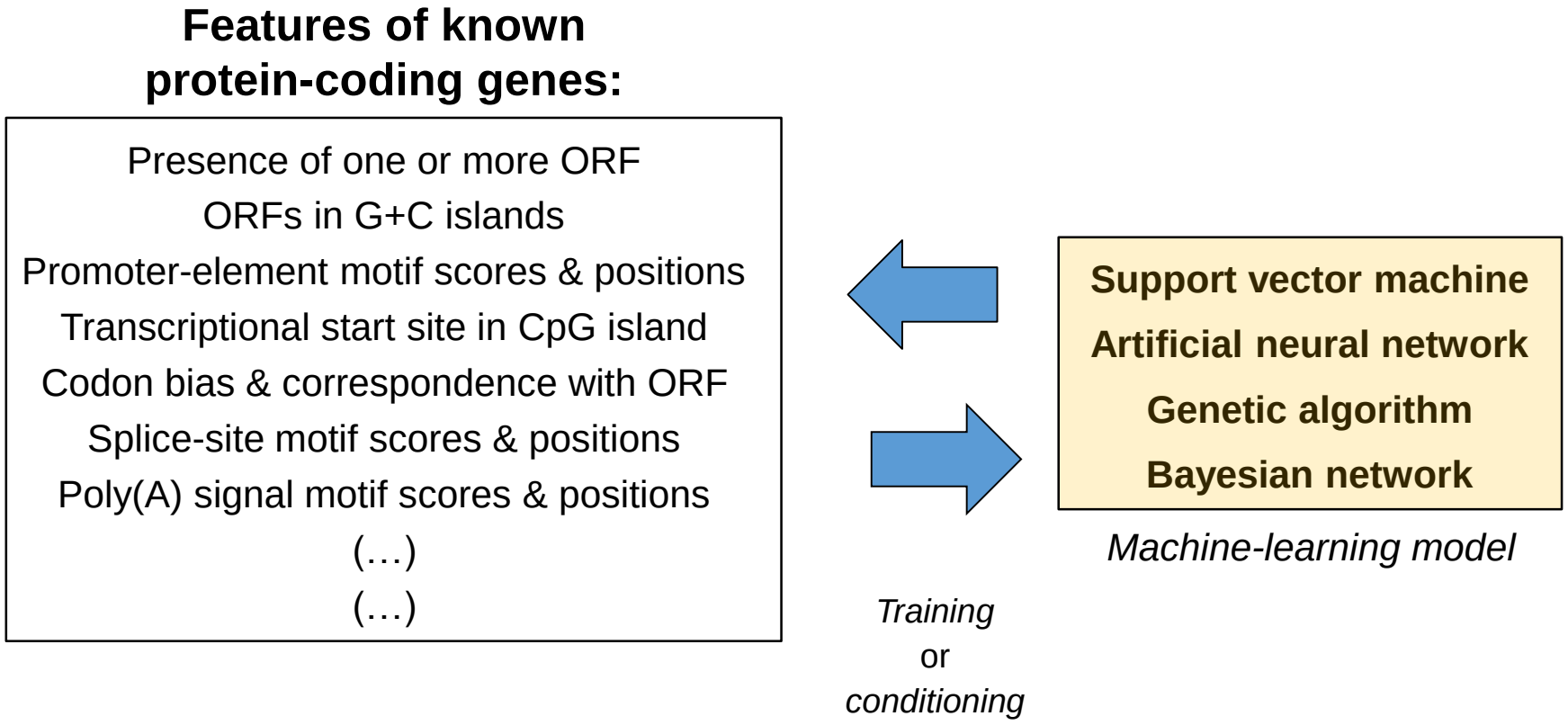
1) Identity search

2) Similarity search Homology based

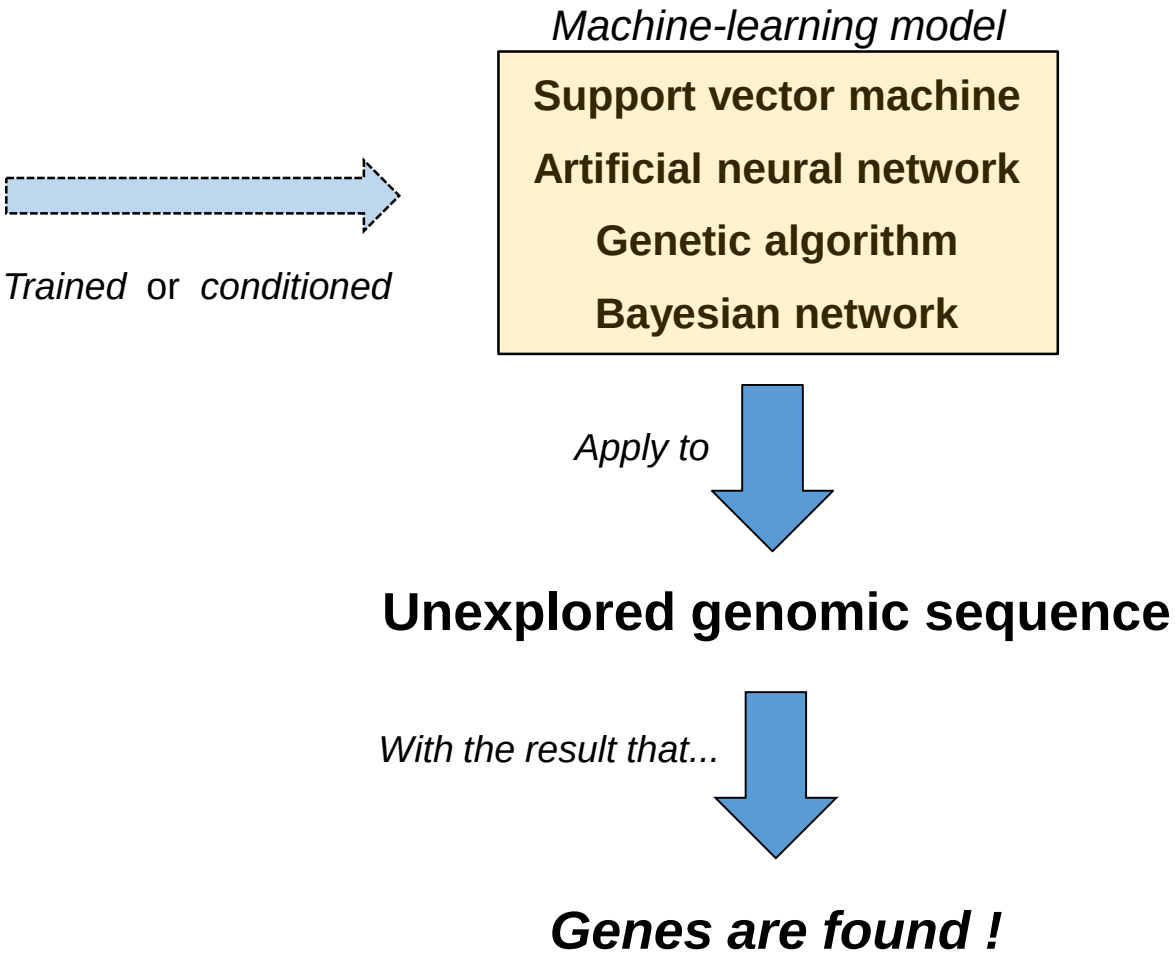
3) *Ab initio* approaches

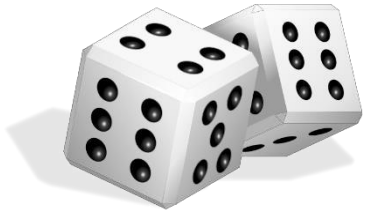
Slides adapted from Lutz Krause

Machine learning approach to ab initio gene finding



Machine learning approach to ab initio gene finding





Hidden Markov Models

The dishonest casino:

Known information:

- Casino has 2 die, fair dice, loaded dice
- Casino player switches back & forth between dies
- Once either of the dice is used, it will continue to be used for a while

Number	Probability	
	Fair	Loaded
1	1/6	1/10
2	1/6	1/10
3	1/6	1/10
4	1/6	1/10
5	1/6	1/10
6	1/6	**1/2**

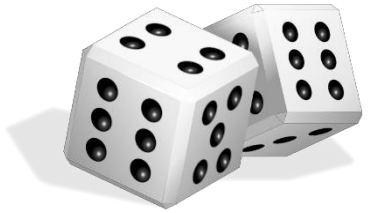
Observations:

- Sequence of roles:

  3 5 3 1 3 6 3 6 4 4 1 6 2 …

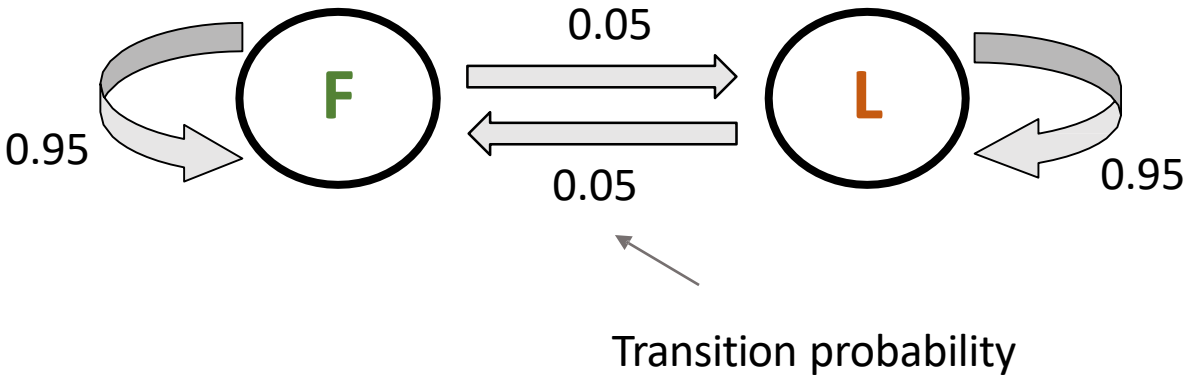
Question:

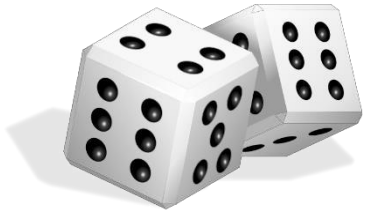
- Which dice used for each role?



Dishonest Casino Example

Number	Probability	
	Fair	Loaded
1	1/6	1/10
2	1/6	1/10
3	1/6	1/10
4	1/6	1/10
5	1/6	1/10
6	1/6	1/2



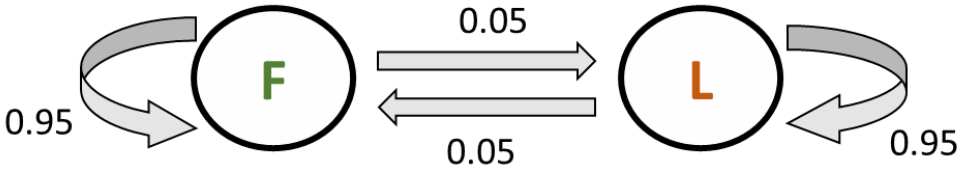


Dishonest Casino Example

Observation:

Sequence of roles:

Obs: 3 1 **6** 2 5 2 3 1 3 **6** 3 **6** **6** 4 **6** **6** 2 **6** …



Hidden information:

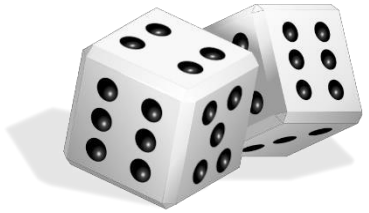
Sequence of states, e.g.

S1: F F F F F F F F F L L L L L L L L L….

S2: F F F F F F F F F F F F F F F F F F….

S3: L L L F F F F F F L L L L L L L L L…..

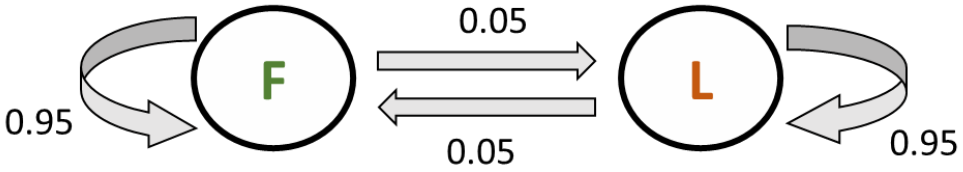
Number	Probability	
	Fair	Loaded
1	1/6	1/10
2	1/6	1/10
3	1/6	1/10
4	1/6	1/10
5	1/6	1/10
6	1/6	**1/2**



Dishonest Casino Example

Obs: 3 1 **6** 2 5 2 3 1 3 **6** 3 **6** **6** 4 **6** **6** 2 **6**

S1: F F F F F F F F F L L L L L L L L L



Transition to L state

P(Obs|S1) = 1/6 * 0.95 * 1/6 * 0.95 … * 0.05 * 1/2 * 0.95 * 1/10 * 0.95 * 1/2 … = 3.4e-14

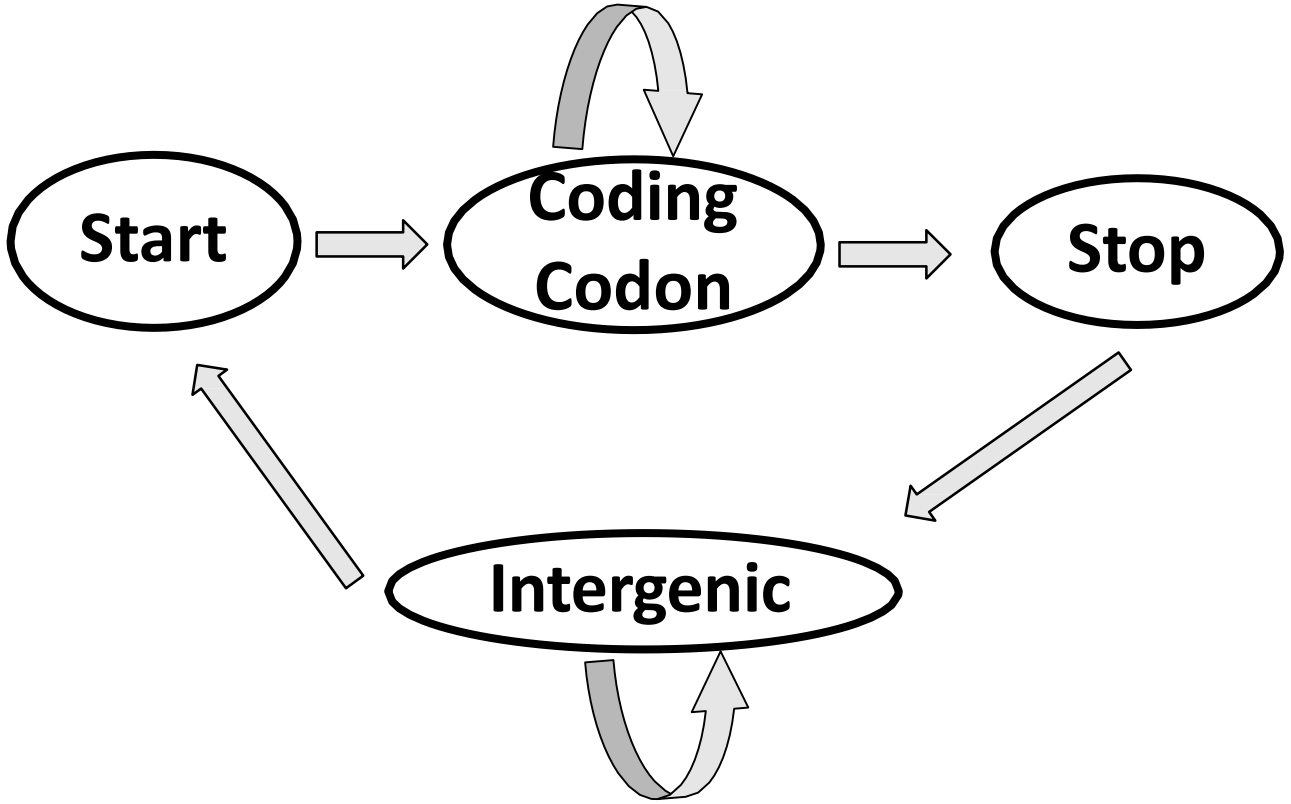
P(Obs|S2) = 4.1e-15

…..

Aim: Identify most likely path through model, which is S1 in this case, 9 roles fair dice, 9 roles loaded dice

Number	Probability Fair	Loaded
1	1/6	1/10
2	1/6	1/10
3	1/6	1/10
4	1/6	1/10
5	1/6	1/10
6	1/6	**1/2**

Simple HMM for Gene Identification in Prokaryotes



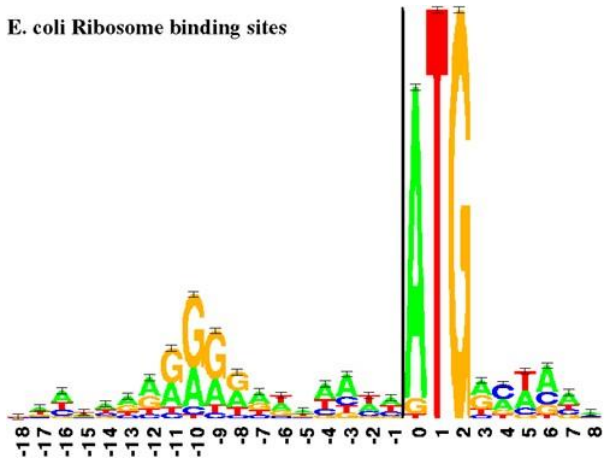
Training: Train model to learn codon frequencies of coding and non-coding sequences

Classification: Given observed DNA sequence, find most likely path through model to divide sequence into coding and non-coding regions

…CCTATC **ATG** GCT ATC GAC GAA AAC AAA … **TAA** CCTTATACTAG…

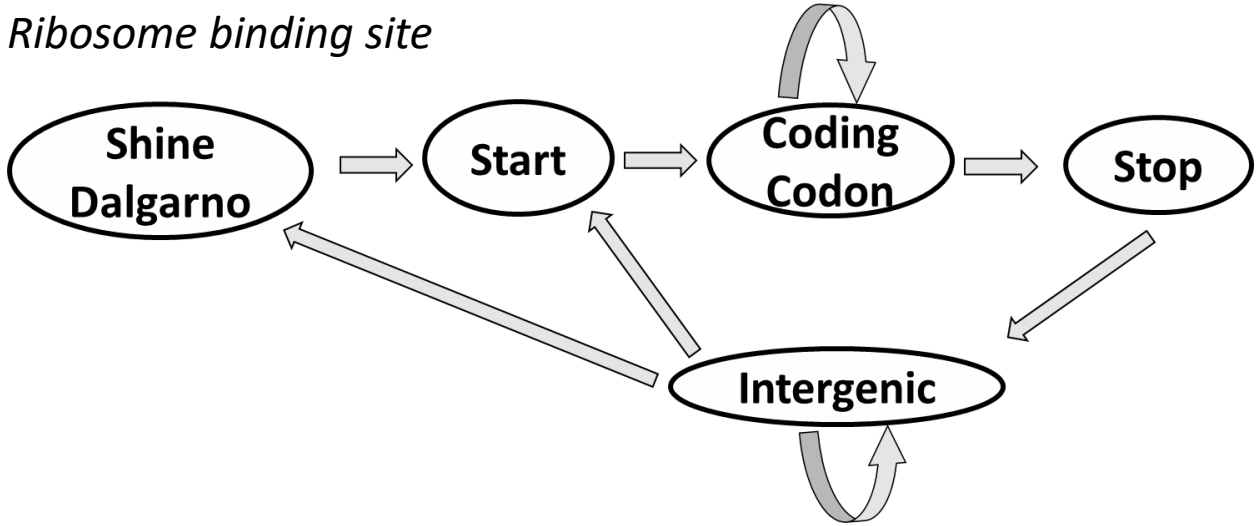
More Complex HMM for Gene Identification in Prokaryotes

Include signal for ribosomal binding site



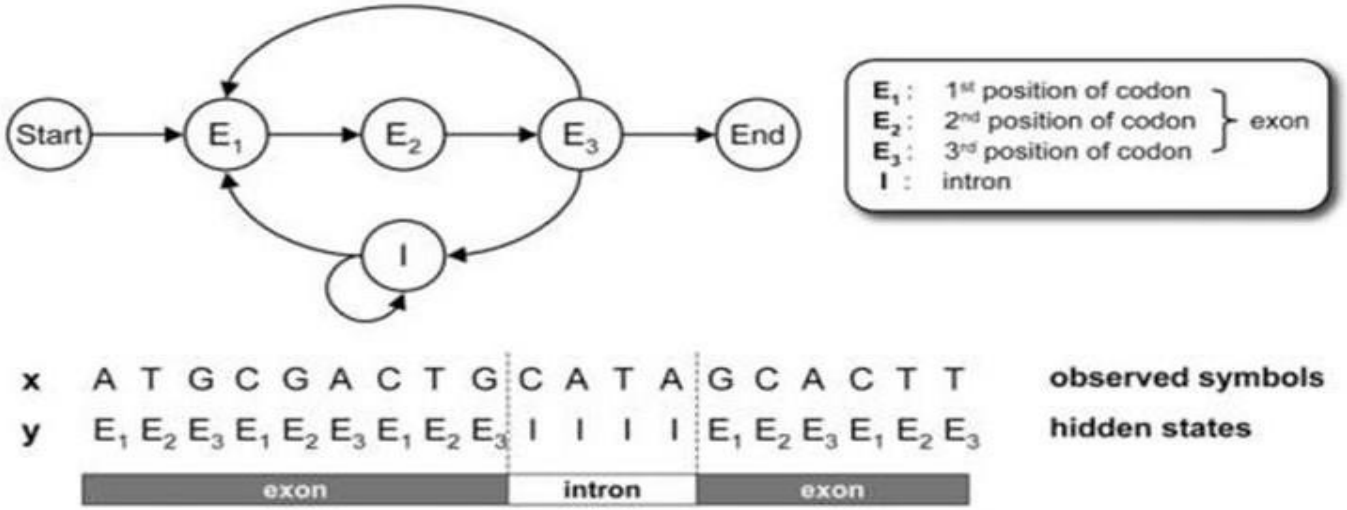
A/G-rich region about 10 bases upstream of the start codon
Helps recruit ribosome to mRNA

Ribosome binding site

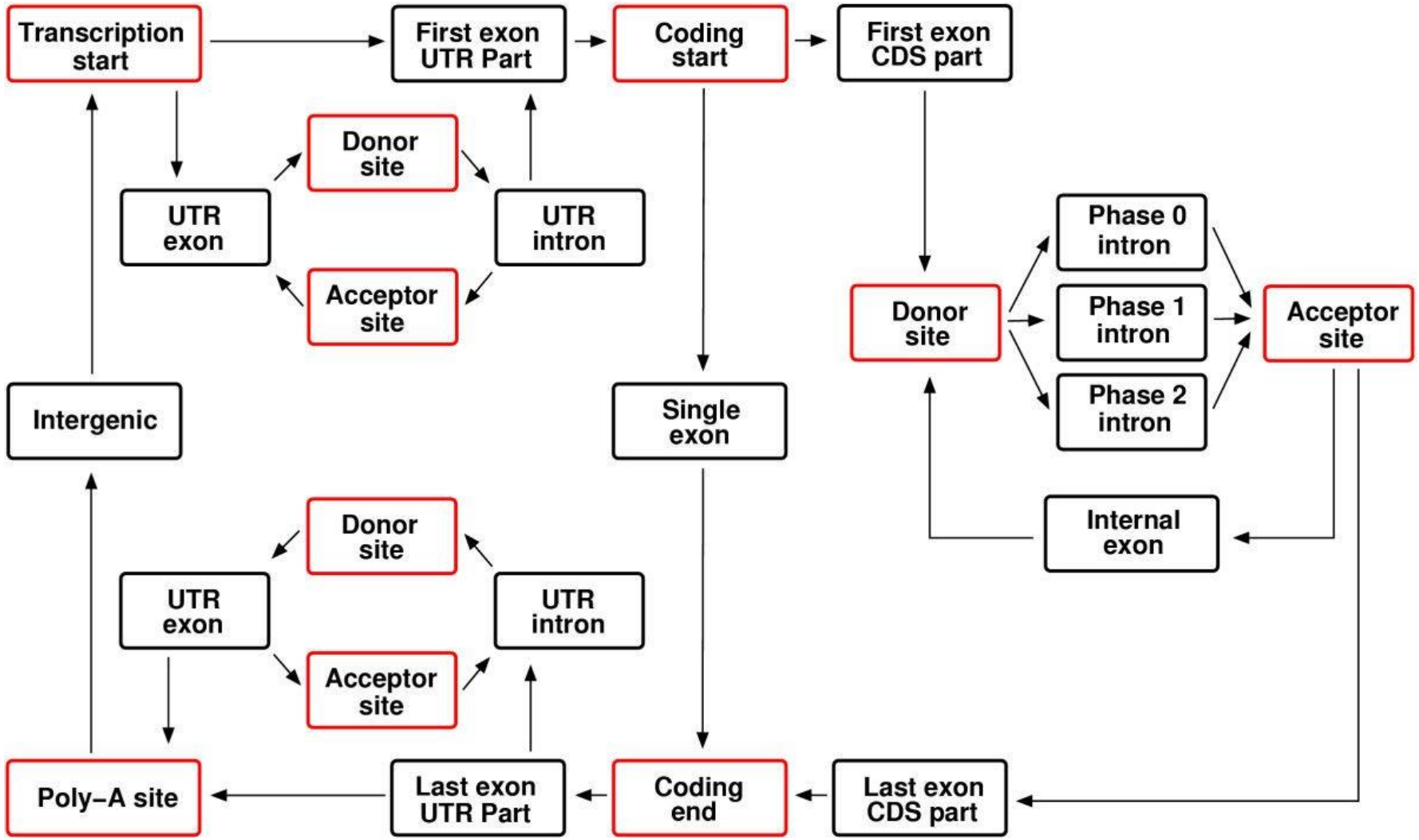


Gene Prediction in Eukaryotic Genomes

- **A Simple HMM for Modeling Eukaryotic Genes**



- hmmlearn() Python package. https://github.com/hmmlearn/hmmlearn


Transcription start
First exon UTR Part
Coding start
First exon CDS part
Donor site
UTR exon
UTR intron
Acceptor site
Phase 0 intron
Donor site
Phase 1 intron
Acceptor site
Phase 2 intron
Intergenic
Single exon
Internal exon
Donor site
UTR exon
UTR intron
Acceptor site
Poly–A site
Last exon UTR Part
Coding end
Last exon CDS part