Stochastic Models for DNA Tandem Duplication

Farzad Farnoud, with M. Schwartz, J. Bruck

Jan 18, 2016, University of Washington

Caltech

Mutations

Mutations

8.7 million species

Mutations

Data storage in *live DNA*

8.7 million species

Types of Mutations

TGATGCA

↓ Point Mutation

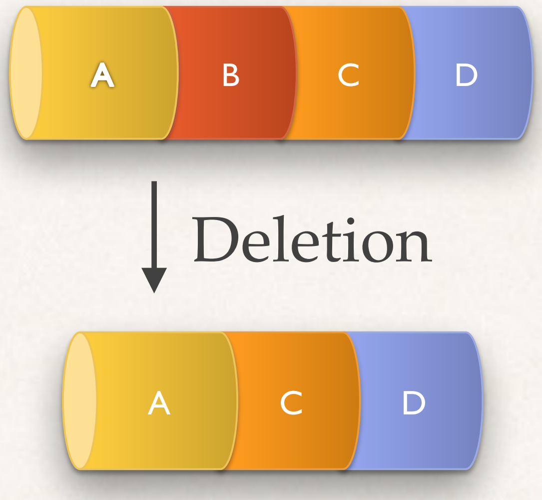
TCATGCA

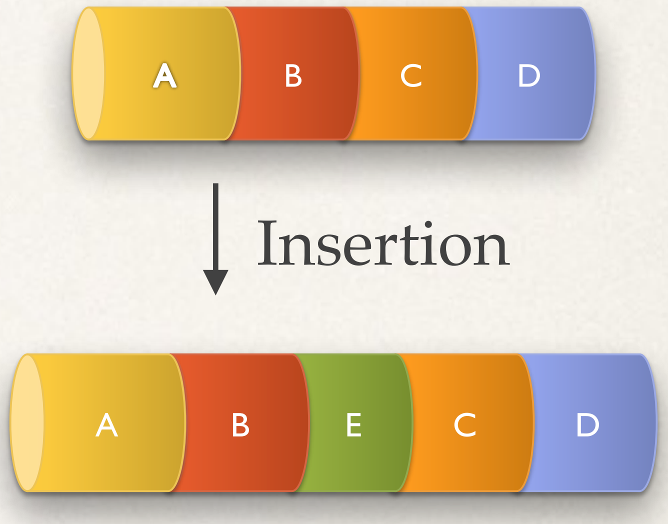
Types of Mutations

TGATGCA

↓ Point Mutation

TCATGCA


Deletion

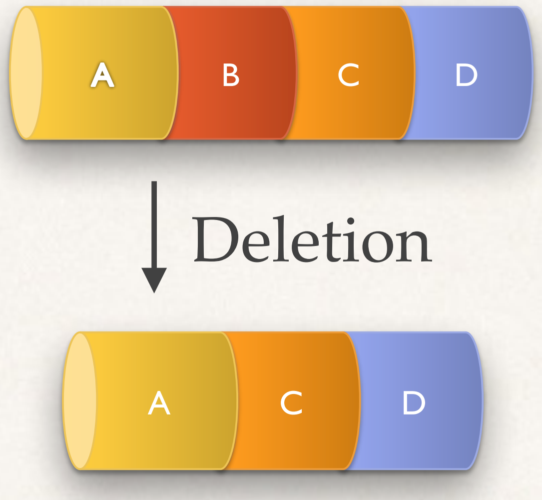

Insertion

Types of Mutations

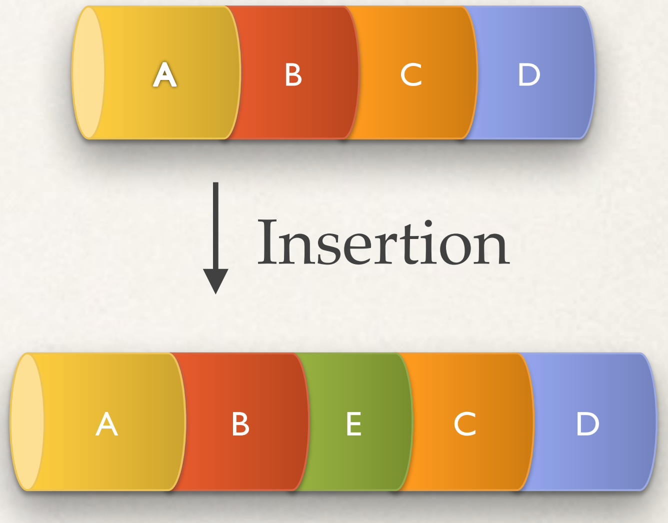
TGATGCA

↓ Point Mutation

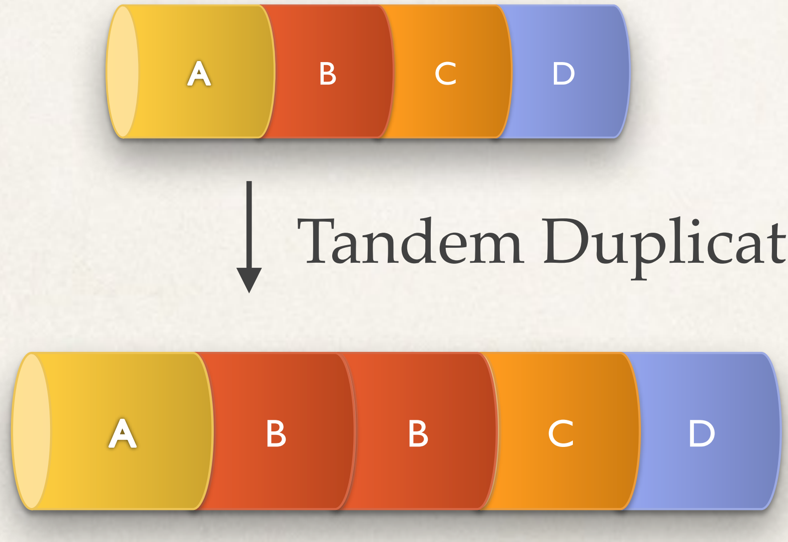
TCATGCA



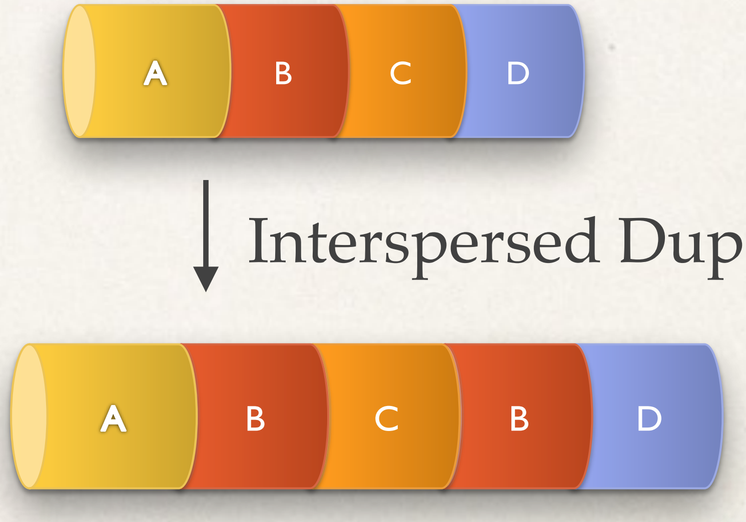
Deletion



Insertion



Tandem Duplication



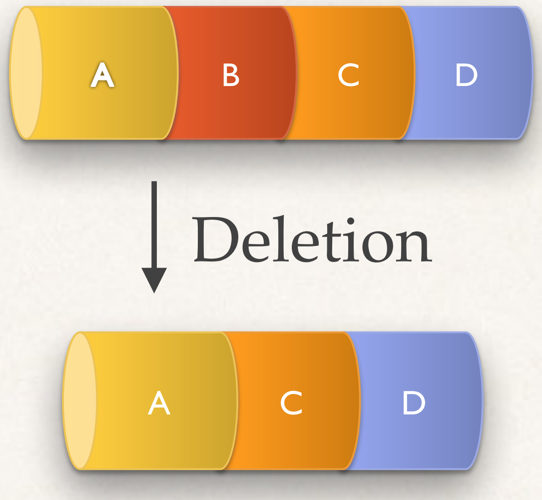
Interspersed Duplication

Types of Mutations

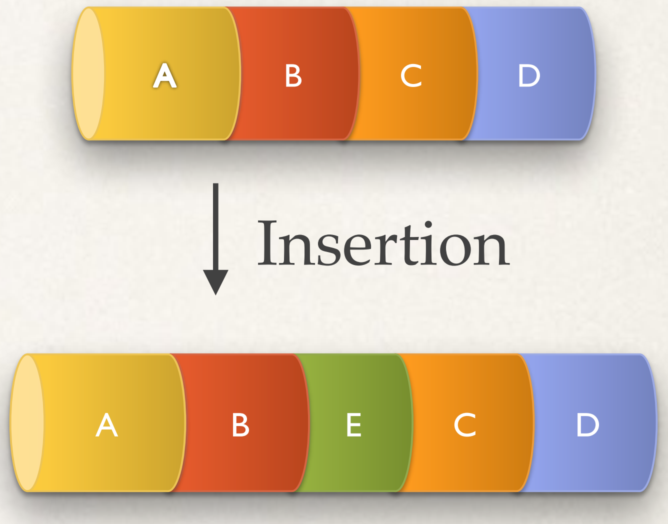
TGATGCA

↓ Point Mutation

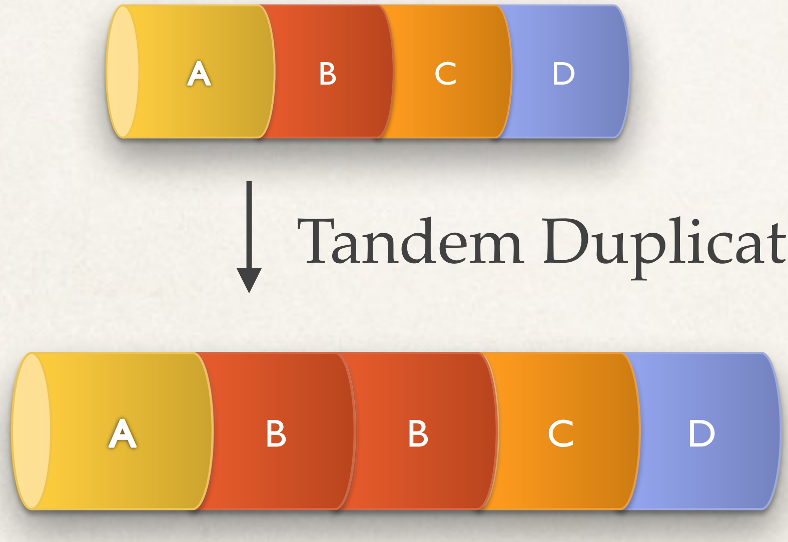
TCATGCA



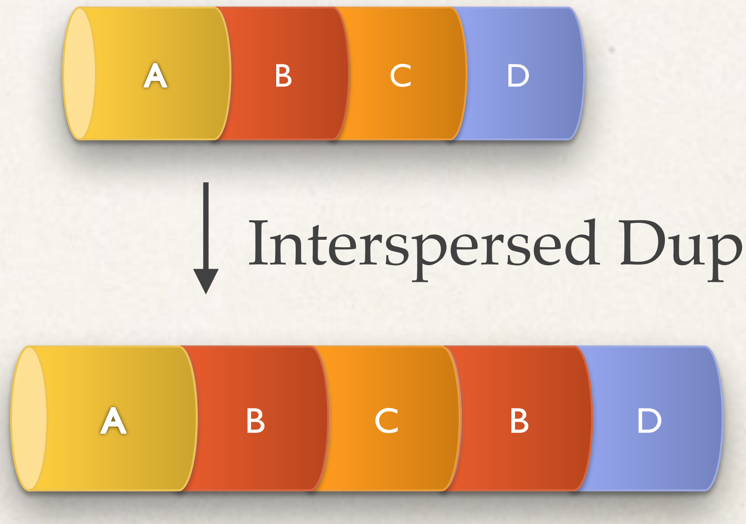
Deletion



Insertion



Tandem Duplication



Interspersed Duplication

3% of human genome

A tandem repeat region

Ch1: 933,911–935,015

GCTCCGTTACAGGTGGGCAGGGGAGGCGGCTGCGTTACAGGTGGGCAGGGGAGGCGGCTGCGTTAC
AGGTGGGCAGGGGAGGCGGCTGCGTTACAGGTGGGCAGGGGAGGCGGCTGCGTTACAGGTGGGCAG
GGGAGGCGGCTCCGTTACAGGTGGGCAGGGGAGGCGGCTGCGTTACAGGTGGGCAGGGGAGGCGGC
TGCGTTACAGGTGGGCAGGGGAGGCGGCTGCGTTACAGGTGGGCAGGGGAGGCGGCTCCGTTACAG
GTGGGCAGGGGAGGCGGCTCCGTTACAGGTGGGCAGGGGAGGCGGCTGCGTTACAGGTGGGCAGGG
GAGGCGGCTGCGTTACAGGTGGGCGGGGGAGGCGGCTGCGTTACAGGTGGGCGGGGGAGGCGGCTG
CGTTACAGGTGGGCAGGGGGGGCGGCTGCGTTACAGGTGGGCGGGGGAGGCGGCTGCGTTACAGGT
GGGCGGGGGAGGCGGCTCCGTTACAGGTGGGCGGGGGAGGCGGCTGCGTTACAGGTGGGCGGGGGA
GGCGGCTGCGTTACAGGTGGGCGGGGGGGGCGGCTGCGTTACAGGTGGGCGGGGGAGGCTGCTCCG
TTACAGGTGGGCGGGGGAGGCTGCTCCGTTACAGGTGGGCGGGGGGGGCGGCTGCGTTACAGGTGG
GCGGGGGGGGCGGCTGCGTTACAGGTGGGCGGGGGAGGCGGCTGCGTTACAGGTGGGCGGGGGAGG
CGGCTCCGTTACAGGTGGGCGGGGGAGGCGGCTGCGTTACAGGTGGGCGGGGGAGGCGGCTGCGTT
ACAGGTGGGCAGGGGAGGCGGCTGCGTTACAGGTGGGCAGGGGAGGCGGCTGCGTTACAGGTGGGC
GGGGGAGGCGGCTCCGTTACAGGTGGGCGGGGGAGGCGGCTGCGTTACAGGTGGGCGGGGGAGGCG
GCTGCGTTACAGGTGGGCGGGGGAGGCG

A tandem repeat region

Ch1: 933,911–935,015

```
GCTCCGTTACAGGTGGGCAGGGGAGGCG GCTGCGTTACAGGTGGGCAGGGGAGGCG
GCTGCGTTACAGGTGGGCAGGGGAGGCG GCTGCGTTACAGGTGGGCAGGGGAGGCG
GCTGCGTTACAGGTGGGCAGGGGAGGCG GCTCCGTTACAGGTGGGCAGGGGAGGCG
GCTGCGTTACAGGTGGGCAGGGGAGGCG GCTGCGTTACAGGTGGGCAGGGGAGGCG
GCTGCGTTACAGGTGGGCAGGGGAGGCG GCTCCGTTACAGGTGGGCAGGGGAGGCG
GCTCCGTTACAGGTGGGCAGGGGAGGCG GCTGCGTTACAGGTGGGCAGGGGAGGCG
GCTGCGTTACAGGTGGGCGGGGGAGGCG GCTGCGTTACAGGTGGGCGGGGGAGGCG
GCTGCGTTACAGGTGGGCAGGGGGGGCG GCTGCGTTACAGGTGGGCGGGGGAGGCG
GCTGCGTTACAGGTGGGCGGGGGAGGCG GCTCCGTTACAGGTGGGCGGGGGAGGCG
GCTGCGTTACAGGTGGGCGGGGGAGGCG GCTGCGTTACAGGTGGGCGGGGGGGGCG
GCTGCGTTACAGGTGGGCGGGGGAGGCT GCTCCGTTACAGGTGGGCGGGGGAGGCT
GCTCCGTTACAGGTGGGCGGGGGGGGCG GCTGCGTTACAGGTGGGCGGGGGGGGCG
GCTGCGTTACAGGTGGGCGGGGGAGGCG GCTGCGTTACAGGTGGGCGGGGGAGGCG
GCTCCGTTACAGGTGGGCGGGGGAGGCG GCTGCGTTACAGGTGGGCGGGGGAGGCG
GCTGCGTTACAGGTGGGCAGGGGAGGCG GCTGCGTTACAGGTGGGCAGGGGAGGCG
GCTGCGTTACAGGTGGGCGGGGGAGGCG GCTCCGTTACAGGTGGGCGGGGGAGGCG
GCTGCGTTACAGGTGGGCGGGGGAGGCG GCTGCGTTACAGGTGGGCGGGGGAGGCG
```

A tandem repeat region

Ch1: 933,911–935,015

repeat unit length

repeat unit

```
GCTCCGTTACAGGTGGGCAGGGGAGGCG GCTGCGTTACAGGTGGGCAGGGGAGGCG
GCTGCGTTACAGGTGGGCAGGGGAGGCG GCTGCGTTACAGGTGGGCAGGGGAGGCG
GCTGCGTTACAGGTGGGCAGGGGAGGCG GCTCCGTTACAGGTGGGCAGGGGAGGCG
GCTGCGTTACAGGTGGGCAGGGGAGGCG GCTGCGTTACAGGTGGGCAGGGGAGGCG
GCTGCGTTACAGGTGGGCAGGGGAGGCG GCTCCGTTACAGGTGGGCAGGGGAGGCG
GCTCCGTTACAGGTGGGCAGGGGAGGCG GCTGCGTTACAGGTGGGCAGGGGAGGCG
GCTGCGTTACAGGTGGGCGGGGGAGGCG GCTGCGTTACAGGTGGGCGGGGGAGGCG
GCTGCGTTACAGGTGGGCAGGGGGGGCG GCTGCGTTACAGGTGGGCGGGGGAGGCG
GCTGCGTTACAGGTGGGCGGGGGAGGCG GCTCCGTTACAGGTGGGCGGGGGAGGCG
GCTGCGTTACAGGTGGGCGGGGGAGGCG GCTGCGTTACAGGTGGGCGGGGGGGGCG
GCTGCGTTACAGGTGGGCGGGGGAGGCT GCTCCGTTACAGGTGGGCGGGGGAGGCT
GCTCCGTTACAGGTGGGCGGGGGGGGCG GCTGCGTTACAGGTGGGCGGGGGGGGCG
GCTGCGTTACAGGTGGGCGGGGGAGGCG GCTGCGTTACAGGTGGGCGGGGGAGGCG
GCTCCGTTACAGGTGGGCGGGGGAGGCG GCTGCGTTACAGGTGGGCGGGGGAGGCG
GCTGCGTTACAGGTGGGCAGGGGAGGCG GCTGCGTTACAGGTGGGCAGGGGAGGCG
GCTGCGTTACAGGTGGGCGGGGGAGGCG GCTCCGTTACAGGTGGGCGGGGGAGGCG
GCTGCGTTACAGGTGGGCGGGGGAGGCG GCTGCGTTACAGGTGGGCGGGGGAGGCG
```

A tandem repeat region

Ch1: 933,911–935,015

GCT**C**CGTTACAGGTGGGCAGGGGAGGCG GCT**G**CGTTACAGGTGGGCAGGGGAGGCG
GCTGCGTTACAGGTGGGCAGGGGAGGCG GCTCCGTTACAGGTGGGCAGGGGAGGCG
GCTGCGTTACAGGTGGGCAGGGGAGGCG GCTCCGTTACAGGTGGGCAGGGGAGGCG
GCTGCGTTACAGGTGGGCAGGGGAGGCG GCTGCGTTACAGGTGGGCAGGGGAGGCG
GCTGCGTTACAGGTGGGCAGGGGAGGCG GCTCCGTTACAGGTGGGCAGGGGAGGCG
GCTCCGTTACAGGTGGGCAGGGGAGGCG GCTGCGTTACAGGTGGGCAGGGGAGGCG
GCTGCGTTACAGGTGGGCGGGGGAGGCG GCTGCGTTACAGGTGGGCGGGGGAGGCG
GCTGCGTTACAGGTGGGCAGGGGGGGCG GCTGCGTTACAGGTGGGCGGGGGAGGCG
GCTGCGTTACAGGTGGGCGGGGGAGGCG GCTCCGTTACAGGTGGGCGGGGGAGGCG
GCTGCGTTACAGGTGGGCGGGGGAGGCG GCTGCGTTACAGGTGGGCGGGGGGGGCG
GCTGCGTTACAGGTGGGCGGGGGAGGCT GCTCCGTTACAGGTGGGCGGGGGAGGCT
GCTCCGTTACAGGTGGGCGGGGGGGGCG GCTGCGTTACAGGTGGGCGGGGGGGGCG
GCTGCGTTACAGGTGGGCGGGGGAGGCG GCTGCGTTACAGGTGGGCGGGGGAGGCG
GCTCCGTTACAGGTGGGCGGGGGAGGCG GCTGCGTTACAGGTGGGCGGGGGAGGCG
GCTGCGTTACAGGTGGGCAGGGGAGGCG GCTGCGTTACAGGTGGGCAGGGGAGGCG
GCTGCGTTACAGGTGGGCGGGGGAGGCG GCTCCGTTACAGGTGGGCGGGGGAGGCG
GCTGCGTTACAGGTGGGCGGGGGAGGCG GCTGCGTTACAGGTGGGCGGGGGAGGCG

Point mutation (PM)

Point mutations are in the same positions

```
GCTCCGTTACAGGTGGGCAGGGGAGGCG
GCTGCGTTACAGGTGGGCAGGGGAGGCG
GCTGCGTTACAGGTGGGCAGGGGAGGCG
GCTGCGTTACAGGTGGGCAGGGGAGGCG
GCTGCGTTACAGGTGGGCAGGGGAGGCG
GCTCCGTTACAGGTGGGCAGGGGAGGCG
GCTGCGTTACAGGTGGGCGGGGGAGGCG
GCTGCGTTACAGGTGGGCAGGGGGGGCG
GCTGCGTTACAGGTGGGCGGGGGAGGCG
GCTGCGTTACAGGTGGGCGGGGGAGGCG
GCTGCGTTACAGGTGGGCGGGGGAGGCT
GCTCCGTTACAGGTGGGCGGGGGGGGCG
GCTGCGTTACAGGTGGGCGGGGGAGGCG
GCTCCGTTACAGGTGGGCGGGGGAGGCG
GCTGCGTTACAGGTGGGCAGGGGAGGCG
GCTGCGTTACAGGTGGGCGGGGGAGGCG
GCTGCGTTACAGGTGGGCGGGGGAGGCG
GCTGCGTTACAGGTGGGCAGGGGAGGCG
GCTGCGTTACAGGTGGGCAGGGGAGGCG
GCTCCGTTACAGGTGGGCAGGGGAGGCG
GCTGCGTTACAGGTGGGCAGGGGAGGCG
GCTCCGTTACAGGTGGGCAGGGGAGGCG
GCTGCGTTACAGGTGGGCAGGGGAGGCG
GCTGCGTTACAGGTGGGCGGGGGAGGCG
GCTGCGTTACAGGTGGGCGGGGGAGGCG
GCTCCGTTACAGGTGGGCGGGGGAGGCG
GCTGCGTTACAGGTGGGCGGGGGGGGCG
GCTCCGTTACAGGTGGGCGGGGGAGGCT
GCTGCGTTACAGGTGGGCGGGGGGGGCG
GCTGCGTTACAGGTGGGCGGGGGAGGCG
GCTGCGTTACAGGTGGGCGGGGGAGGCG
GCTGCGTTACAGGTGGGCAGGGGAGGCG
GCTCCGTTACAGGTGGGCGGGGGAGGCG
GCTGCGTTACAGGTGGGCGGGGGAGGCG
```

Stochastic Model

Start from one repeat unit (***seed***).
Random mutations:
- Tandem duplications (TD)
  of one or more repeat units
- Point mutations (PM)

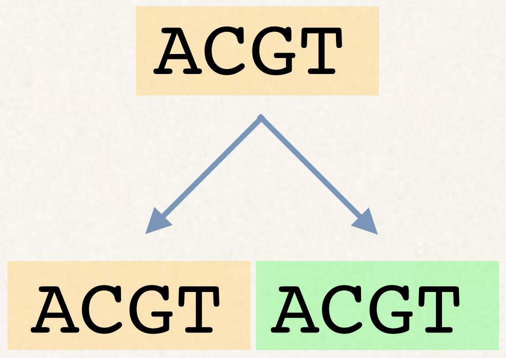
Stochastic Model

`ACGT`

Start from one repeat unit (***seed***).
Random mutations:
- Tandem duplications (TD) of one or more repeat units
- Point mutations (PM)

Stochastic Model

Start from one repeat unit (***seed***).
Random mutations:
- Tandem duplications (TD)
  of one or more repeat units
- Point mutations (PM)



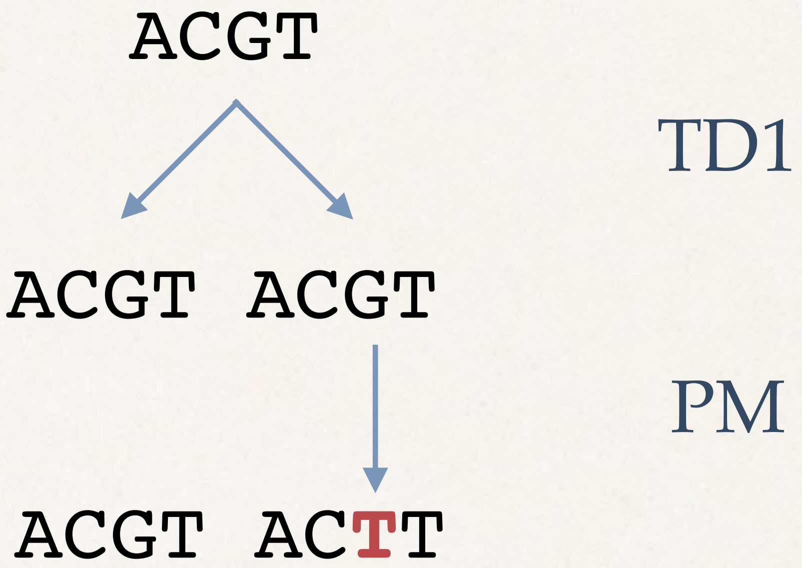
ACGT

ACGT ACGT

TD1

Stochastic Model

Start from one repeat unit (***seed***).
Random mutations:
- Tandem duplications (TD) of one or more repeat units
- Point mutations (PM)



ACGT

TD1

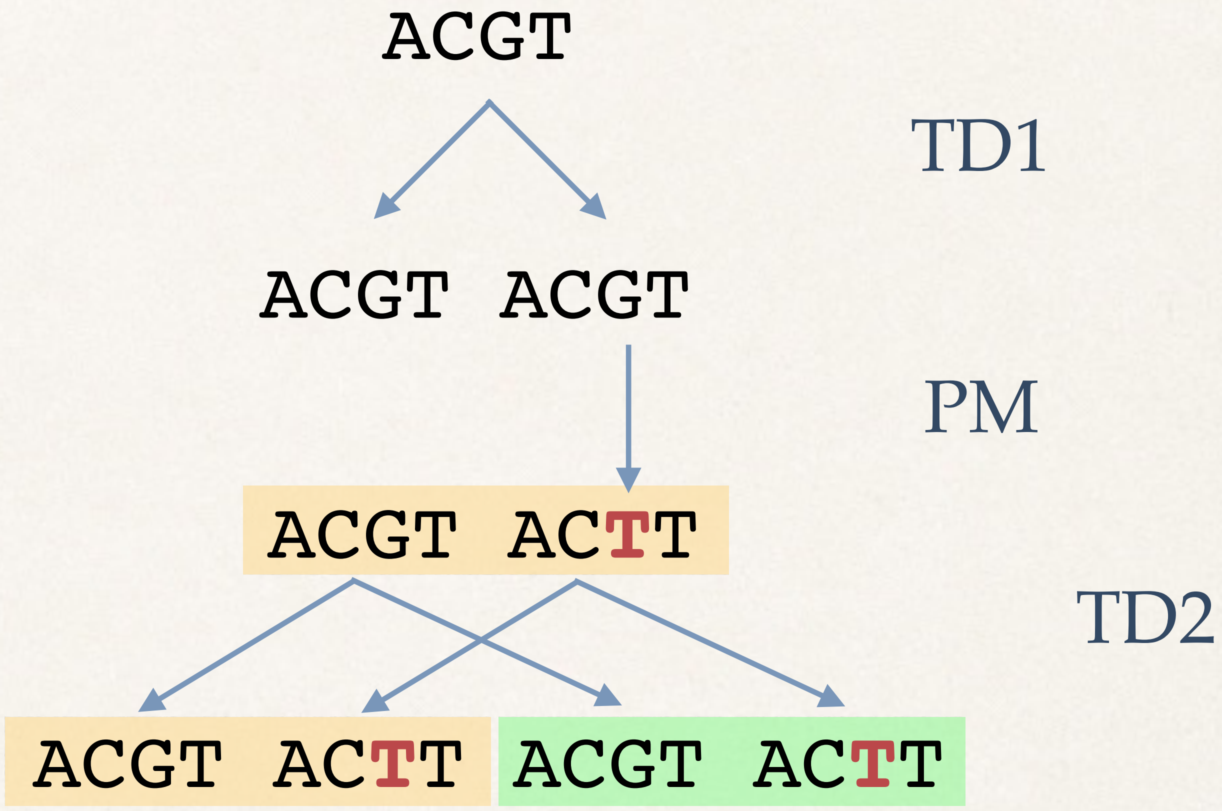
ACGT ACGT

PM

ACGT ACTT

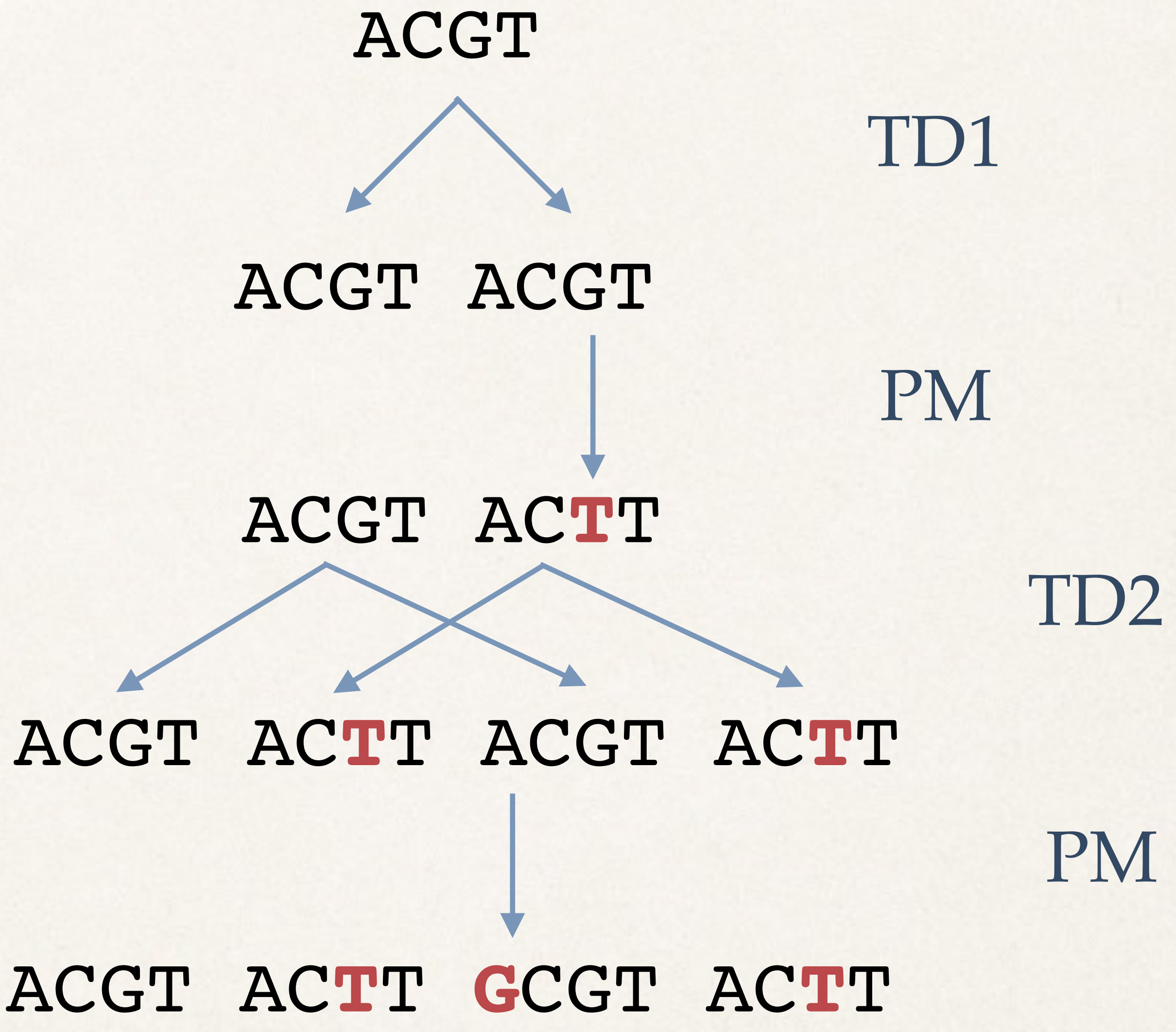
Stochastic Model

Start from one repeat unit (***seed***).
Random mutations:
- Tandem duplications (TD)
  of one or more repeat units
- Point mutations (PM)


ACGT
TD1
ACGT ACGT
PM
ACGT ACTT
TD2
ACGT ACTT ACGT ACTT

Stochastic Model

Start from one repeat unit (***seed***).
Random mutations:
- Tandem duplications (TD)
  of one or more repeat units
- Point mutations (PM)



ACGT

TD1

ACGT ACGT

PM

ACGT ACTT

TD2

ACGT ACTT ACGT ACTT

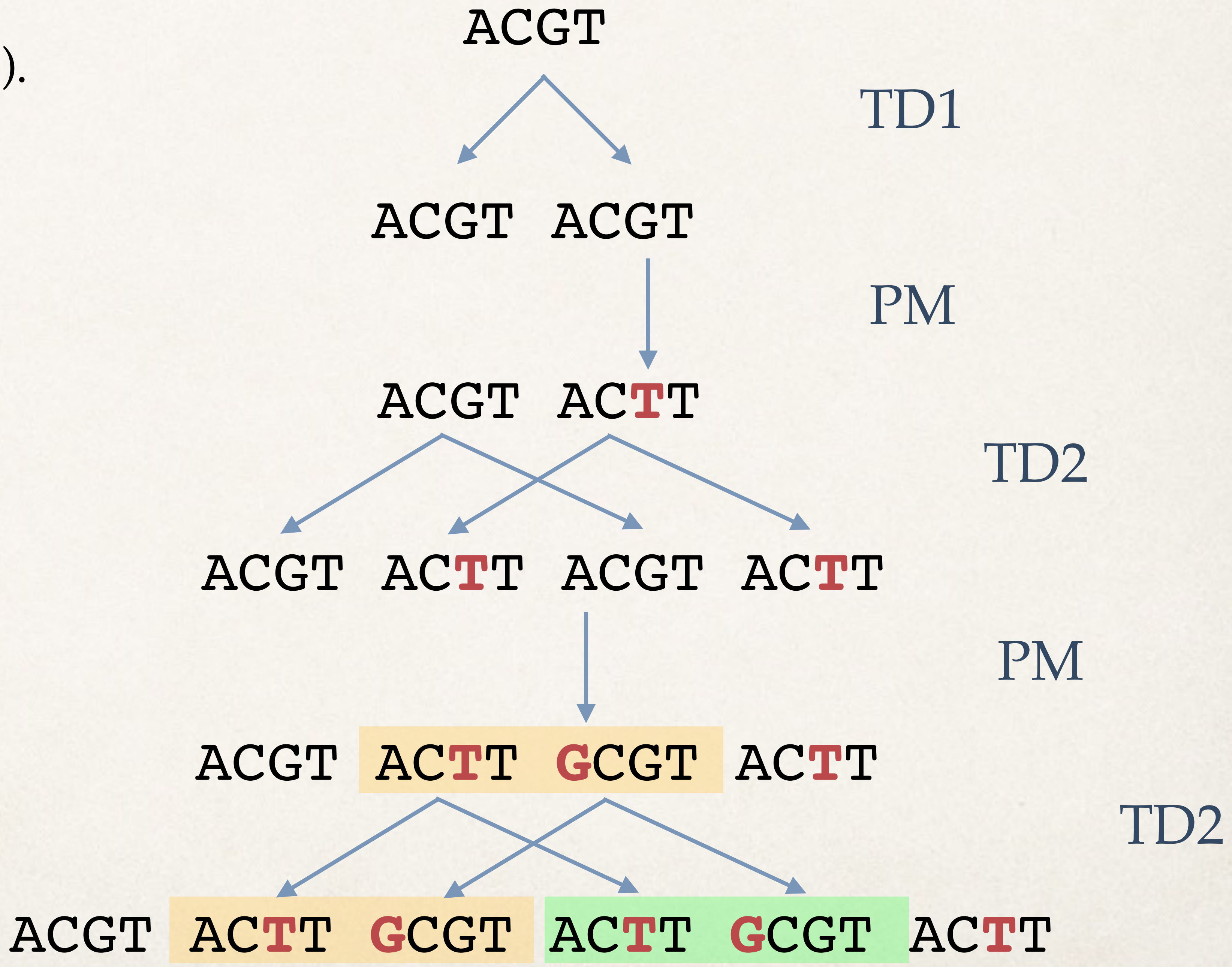
PM

ACGT ACTT GCGT ACTT

Stochastic Model

Start from one repeat unit (*seed*).
Random mutations:

- Tandem duplications (TD) of one or more repeat units
- Point mutations (PM)



ACGT

TD1

ACGT ACGT

PM

ACGT ACTT

TD2

ACGT ACTT ACGT ACTT

PM

ACGT ACTT GCGT ACTT

TD2

ACGT ACTT GCGT ACTT GCGT ACTT

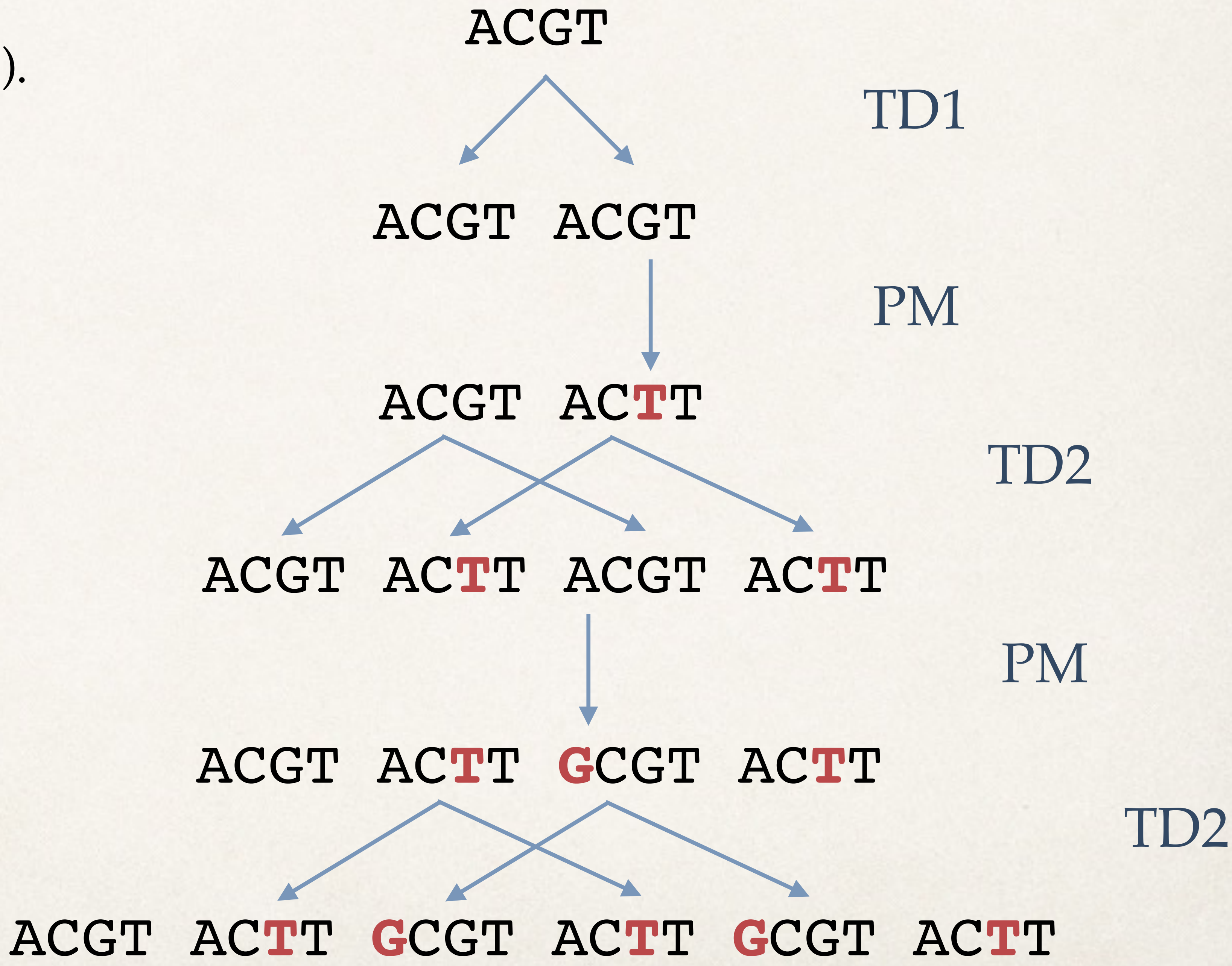
Stochastic Model

Start from one repeat unit (***seed***).
Random mutations:
- Tandem duplications (TD)
  of one or more repeat units
- Point mutations (PM)

Parameters of the model:

- Prob. of PM
- Prob. of TDs
  of different lengths

Can we learn them?



ACGT

TD1

ACGT ACGT

PM

ACGT ACTT

TD2

ACGT ACTT ACGT ACTT

PM

ACGT ACTT GCGT ACTT

TD2

ACGT ACTT GCGT ACTT GCGT ACTT

Stochastic Model

Start from one repeat unit (***seed***).
Random mutations:
- Tandem duplications (TD)
  of one or more repeat units
- Point mutations (PM)

Parameters of the model:
- Prob. of PM
- Prob. of TDs
  of different lengths

Can we learn them?

ACGT ACTT GCGT ACTT GCGT ACTT

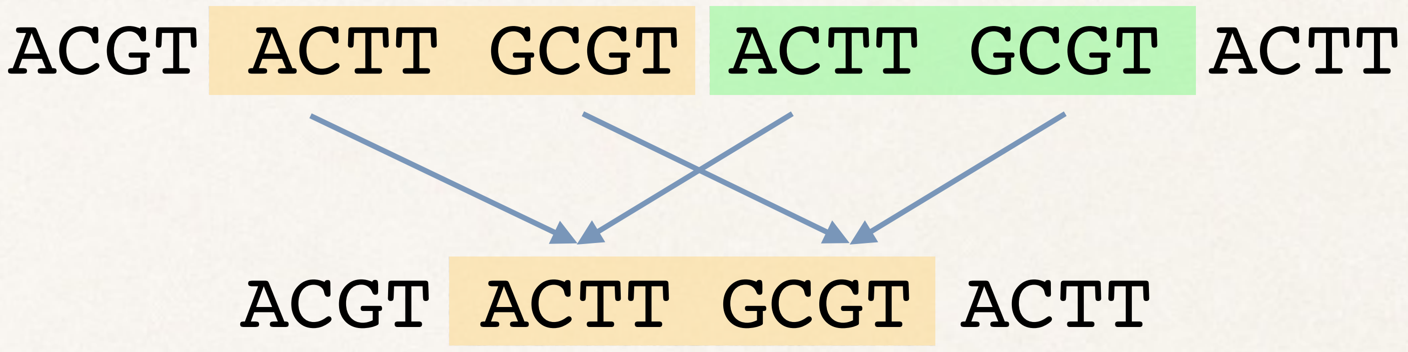
Finding Duplication History

`ACGT ACTT GCGT ACTT GCGT ACTT`

Finding Duplication History

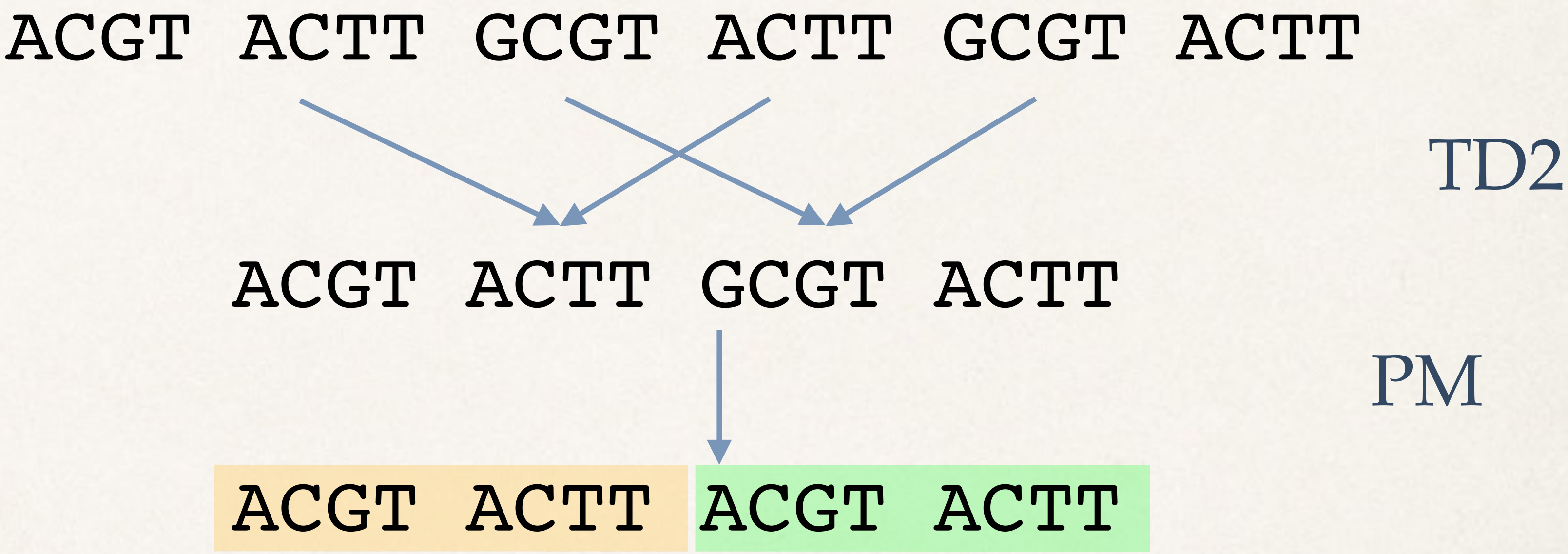
`ACGT ACTT GCGT ACTT GCGT ACTT`

Finding Duplication History



TD2

Finding Duplication History



ACGT ACTT GCGT ACTT GCGT ACTT

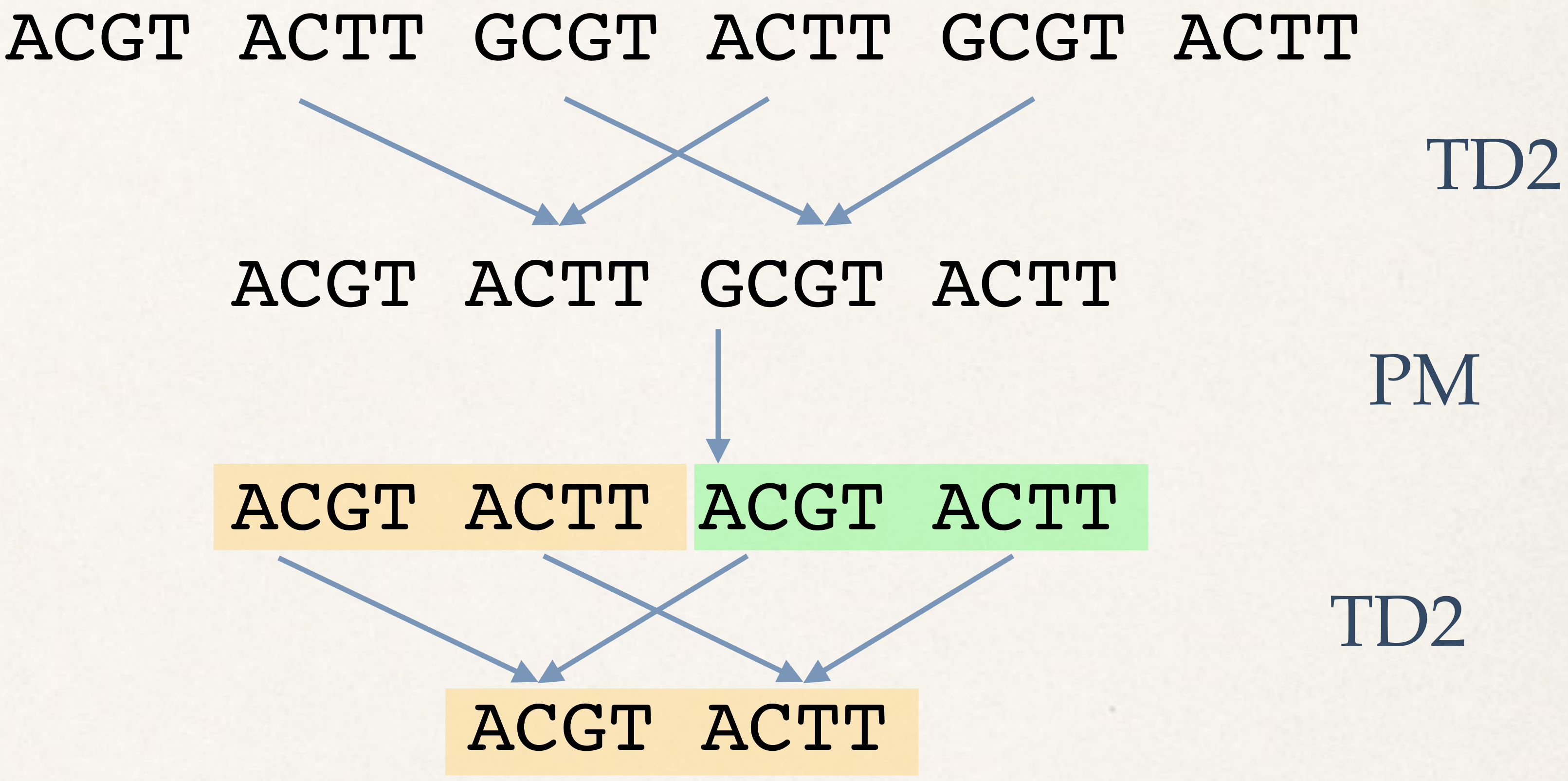
TD2

ACGT ACTT GCGT ACTT

PM

ACGT ACTT ACGT ACTT

Finding Duplication History



ACGT ACTT GCGT ACTT GCGT ACTT

TD2

ACGT ACTT GCGT ACTT

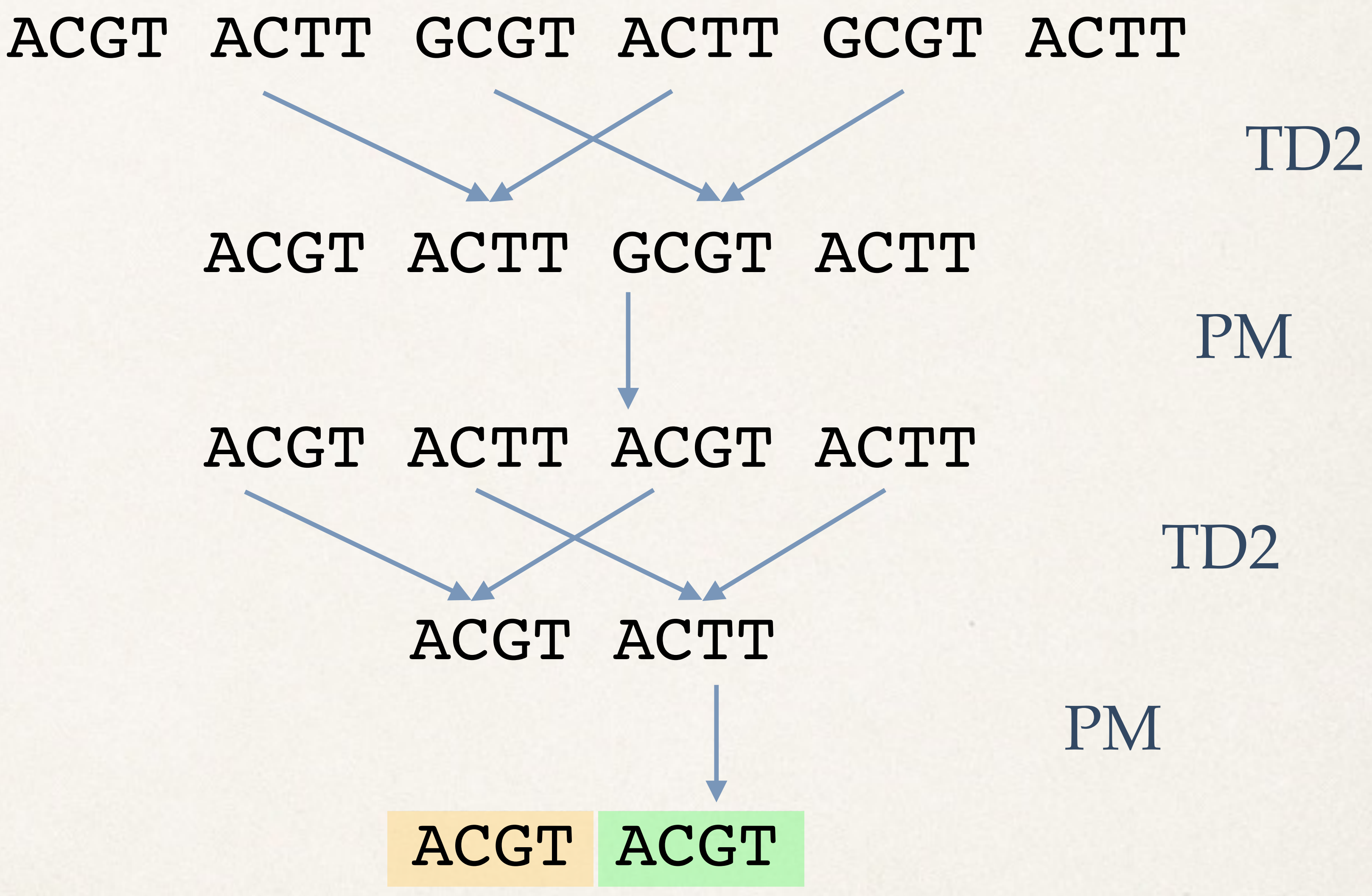
PM

ACGT ACTT ACGT ACTT

TD2

ACGT ACTT

Finding Duplication History



ACGT ACTT GCGT ACTT GCGT ACTT

TD2

ACGT ACTT GCGT ACTT

PM

ACGT ACTT ACGT ACTT

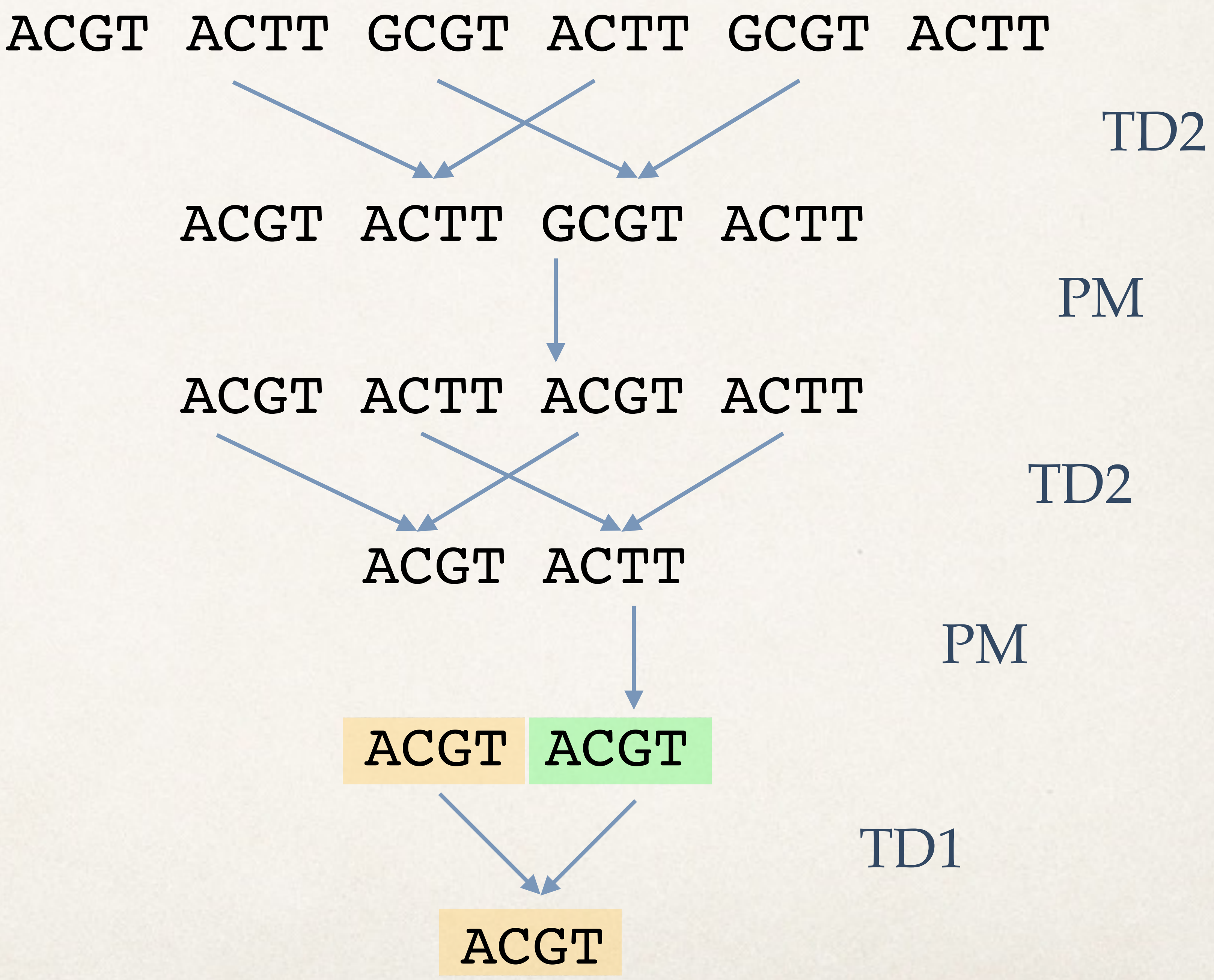
TD2

ACGT ACTT

PM

ACGT ACGT

Finding Duplication History



ACGT ACTT GCGT ACTT GCGT ACTT

TD2

ACGT ACTT GCGT ACTT

PM

ACGT ACTT ACGT ACTT

TD2

ACGT ACTT

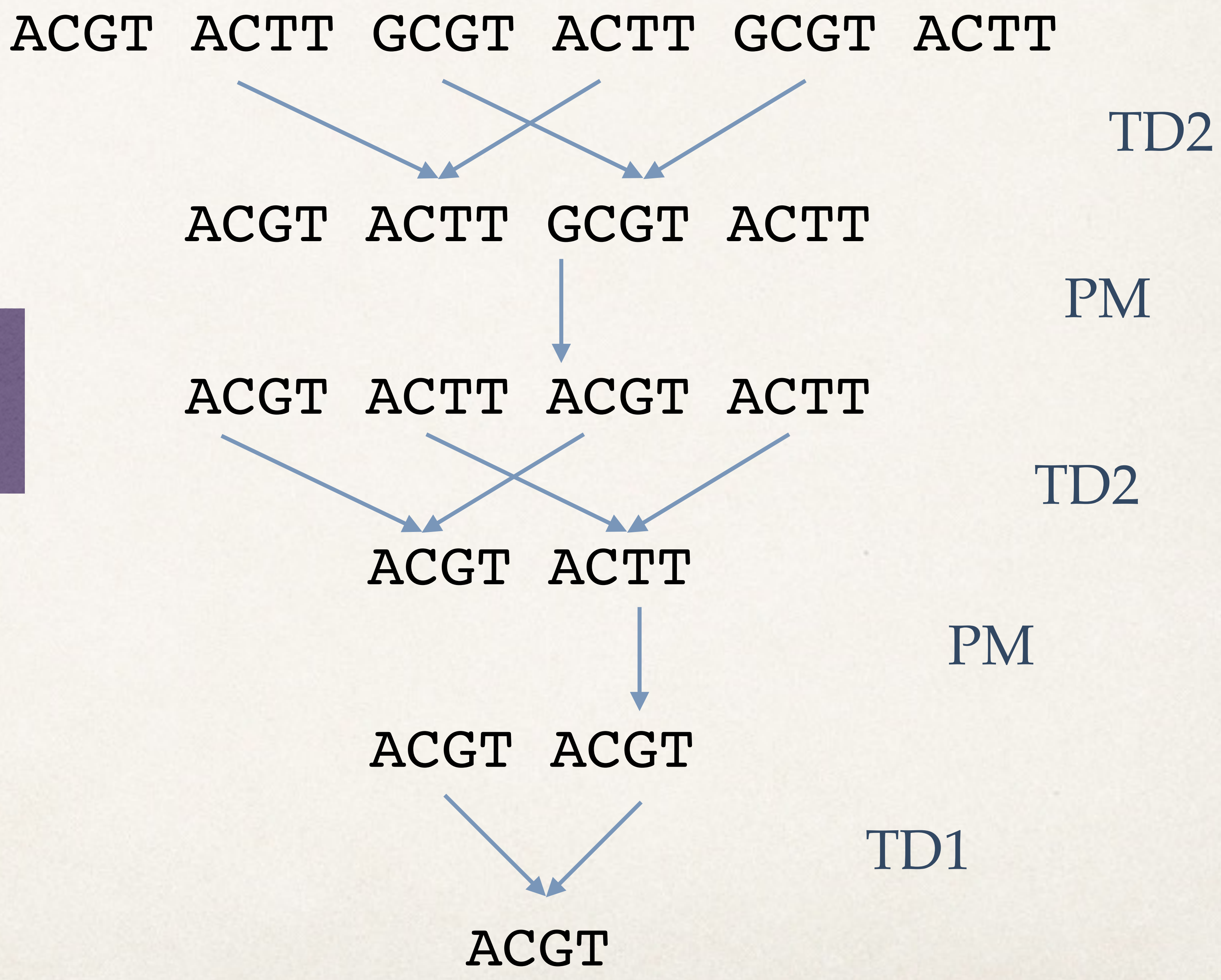
PM

ACGT ACGT

TD1

ACGT

Finding Duplication History



ACGT ACTT GCGT ACTT GCGT ACTT

TD2

ACGT ACTT GCGT ACTT

PM

ACGT ACTT ACGT ACTT

TD2

ACGT ACTT

PM

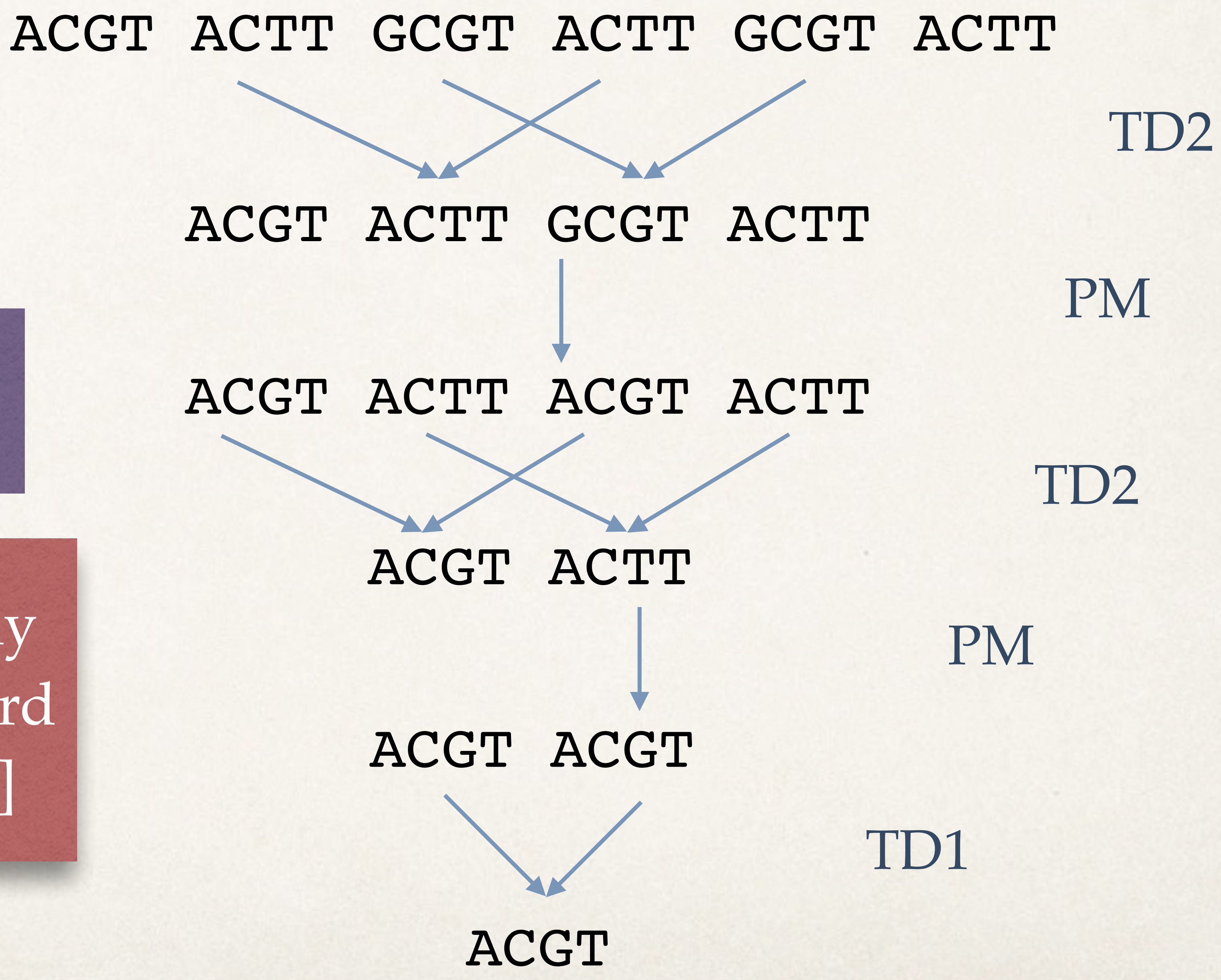
ACGT ACGT

TD1

ACGT

TD1, 2 TD2, 2 PM

Finding Duplication History



`ACGT ACTT GCGT ACTT GCGT ACTT`

TD2

`ACGT ACTT GCGT ACTT`

PM

`ACGT ACTT ACGT ACTT`

TD2

`ACGT ACTT`

PM

`ACGT ACGT`

TD1

`ACGT`

TD1, 2 TD2, 2 PM

Maximum Parsimony
Thought to be NP-hard
[Gascuel et al., 2005]

Given the final sequence, can we efficiently estimate the parameters?

```
GCTCCGTTACAGGTGGGCAGGGGAGGCG GCTGCGTTACAGGTGGGCAGGGGAGGCG
GCTGCGTTACAGGTGGGCAGGGGAGGCG GCTGCGTTACAGGTGGGCAGGGGAGGCG
GCTGCGTTACAGGTGGGCAGGGGAGGCG GCTCCGTTACAGGTGGGCAGGGGAGGCG
GCTGCGTTACAGGTGGGCAGGGGAGGCG GCTGCGTTACAGGTGGGCAGGGGAGGCG
GCTGCGTTACAGGTGGGCAGGGGAGGCG GCTCCGTTACAGGTGGGCAGGGGAGGCG
GCTCCGTTACAGGTGGGCAGGGGAGGCG GCTGCGTTACAGGTGGGCAGGGGAGGCG
GCTGCGTTACAGGTGGGCGGGGGAGGCG GCTGCGTTACAGGTGGGCGGGGGAGGCG
GCTGCGTTACAGGTGGGCAGGGGGGGCG GCTGCGTTACAGGTGGGCGGGGGAGGCG
GCTGCGTTACAGGTGGGCGGGGGAGGCG GCTCCGTTACAGGTGGGCGGGGGAGGCG
GCTGCGTTACAGGTGGGCGGGGGAGGCG GCTGCGTTACAGGTGGGCGGGGGGGGCG
GCTGCGTTACAGGTGGGCGGGGGAGGCT GCTCCGTTACAGGTGGGCGGGGGAGGCT
GCTCCGTTACAGGTGGGCGGGGGGGGCG GCTGCGTTACAGGTGGGCGGGGGGGGCG
GCTGCGTTACAGGTGGGCGGGGGAGGCG GCTGCGTTACAGGTGGGCGGGGGAGGCG
GCTCCGTTACAGGTGGGCGGGGGAGGCG GCTGCGTTACAGGTGGGCGGGGGAGGCG
GCTGCGTTACAGGTGGGCAGGGGAGGCG GCTGCGTTACAGGTGGGCAGGGGAGGCG
GCTGCGTTACAGGTGGGCGGGGGAGGCG GCTCCGTTACAGGTGGGCGGGGGAGGCG
GCTGCGTTACAGGTGGGCGGGGGAGGCG GCTGCGTTACAGGTGGGCGGGGGAGGCG
```

How to extract information from point mutations?

- Autocorrelation function:
  - $r(\delta) = \textit{fraction of symbols at distance } \delta \textit{ units that are the same}$

$r(1)=11/20$

$r(2)=15/16$

Stochastic Approximation

- Suppose a discrete random process *x* satisfies:

  $$x_{n+1} - x_n = \frac{1}{n}(h(x_n) + M_{n+1})$$

  for a Lipschitz function *h*, and martingale difference *M*.

- Then x_n converges almost surely to a compact connected internally chain transitive invariant set of the ode

  $$\dot{x}_t = h(x_t).$$

Stochastic Approximation for Autocorrelation

Stochastic Approximation for Autocorrelation

- r_n: autocorr. after n mutations

- The stochastic approximation equation for r_n:

$$\frac{d}{dt} r_t = A r_t$$

A: a matrix that depends on the parameters:

P(PM), P(TD1), P(TD2), …

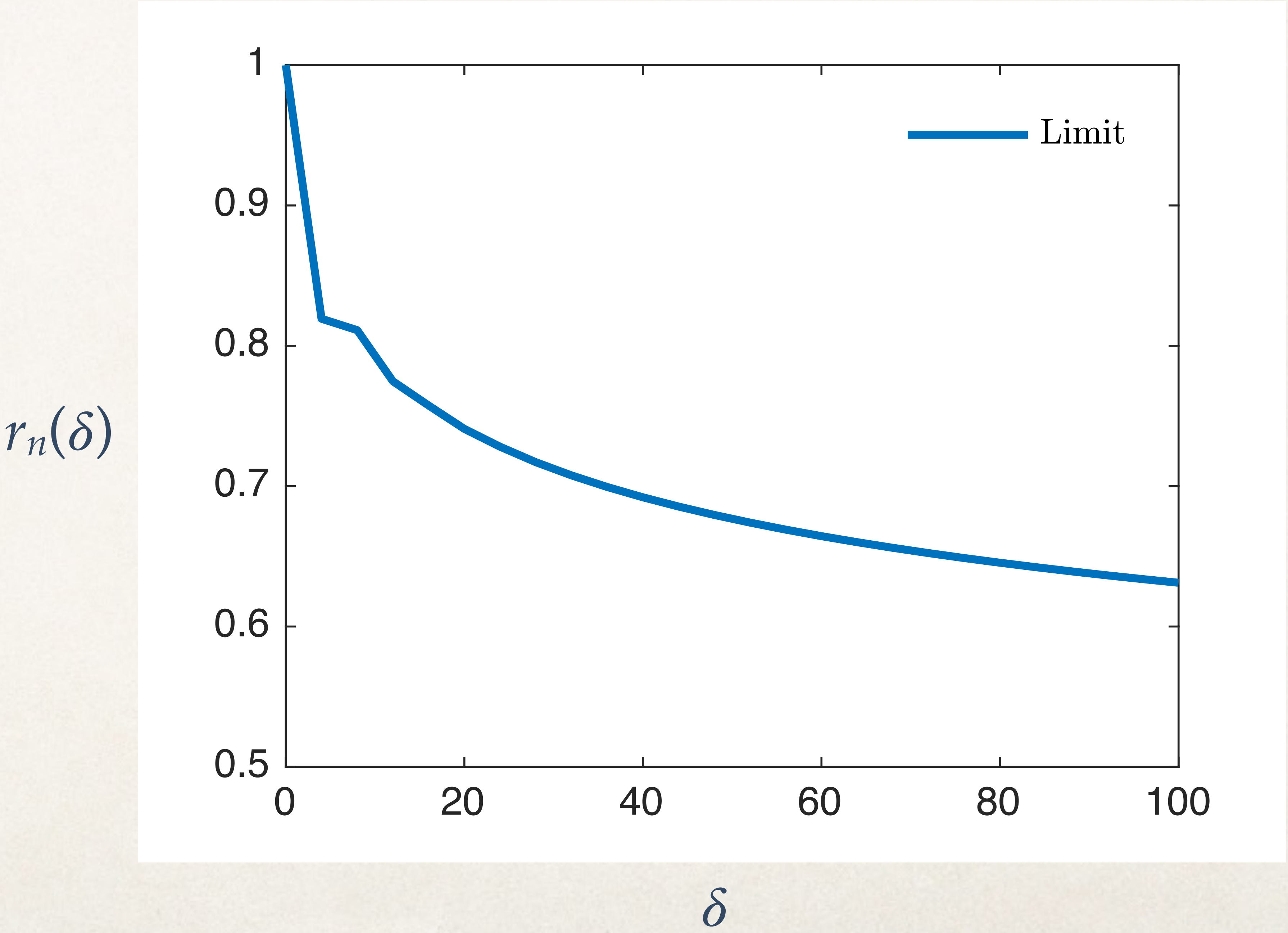
- *As n increases, r_n tends to a point in the null space of A*

Autocorrelation Limit

$P(PM) = 0.250$

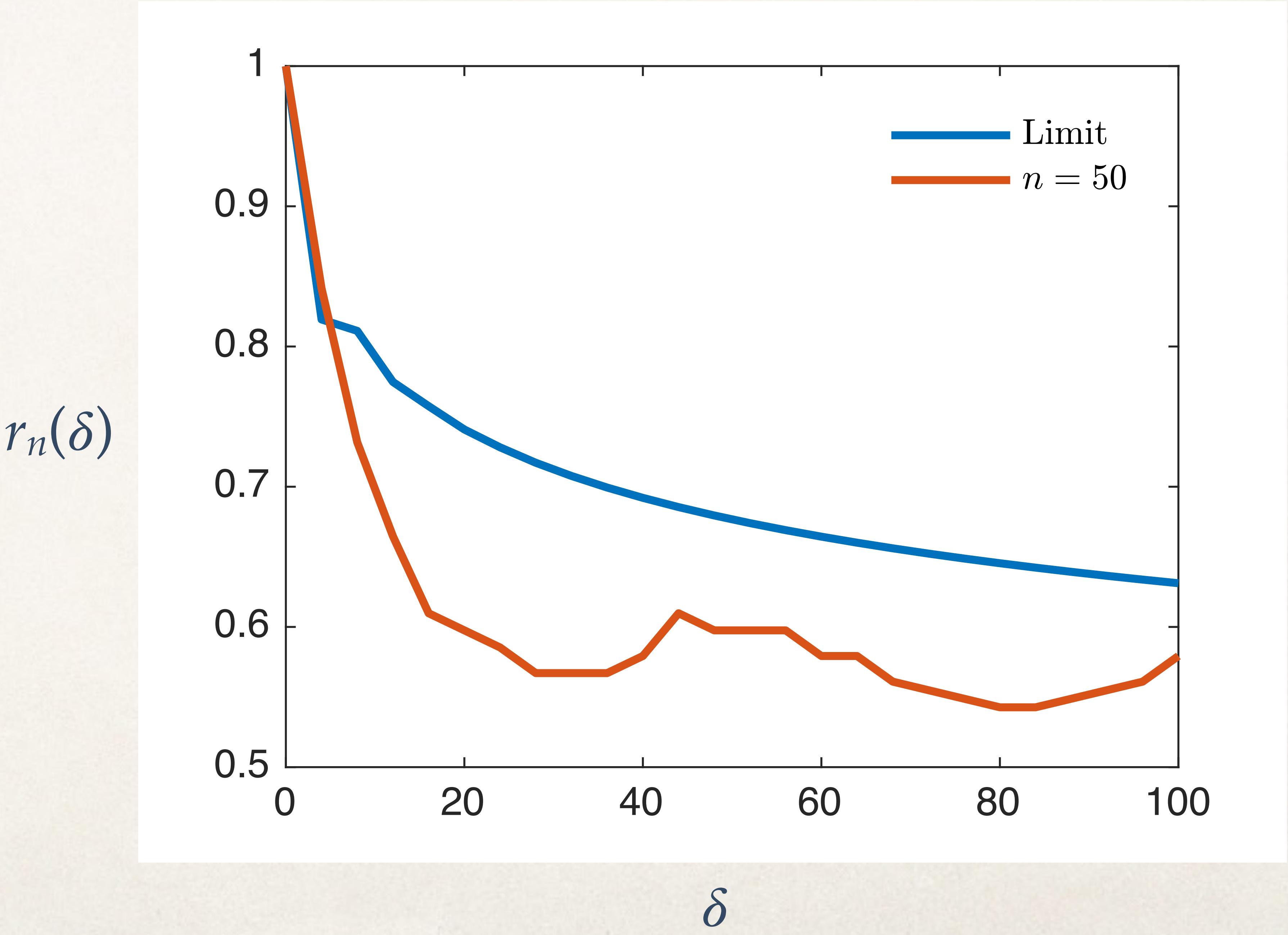
$P(TD1) = 0.525$

$P(TD2) = 0.225$



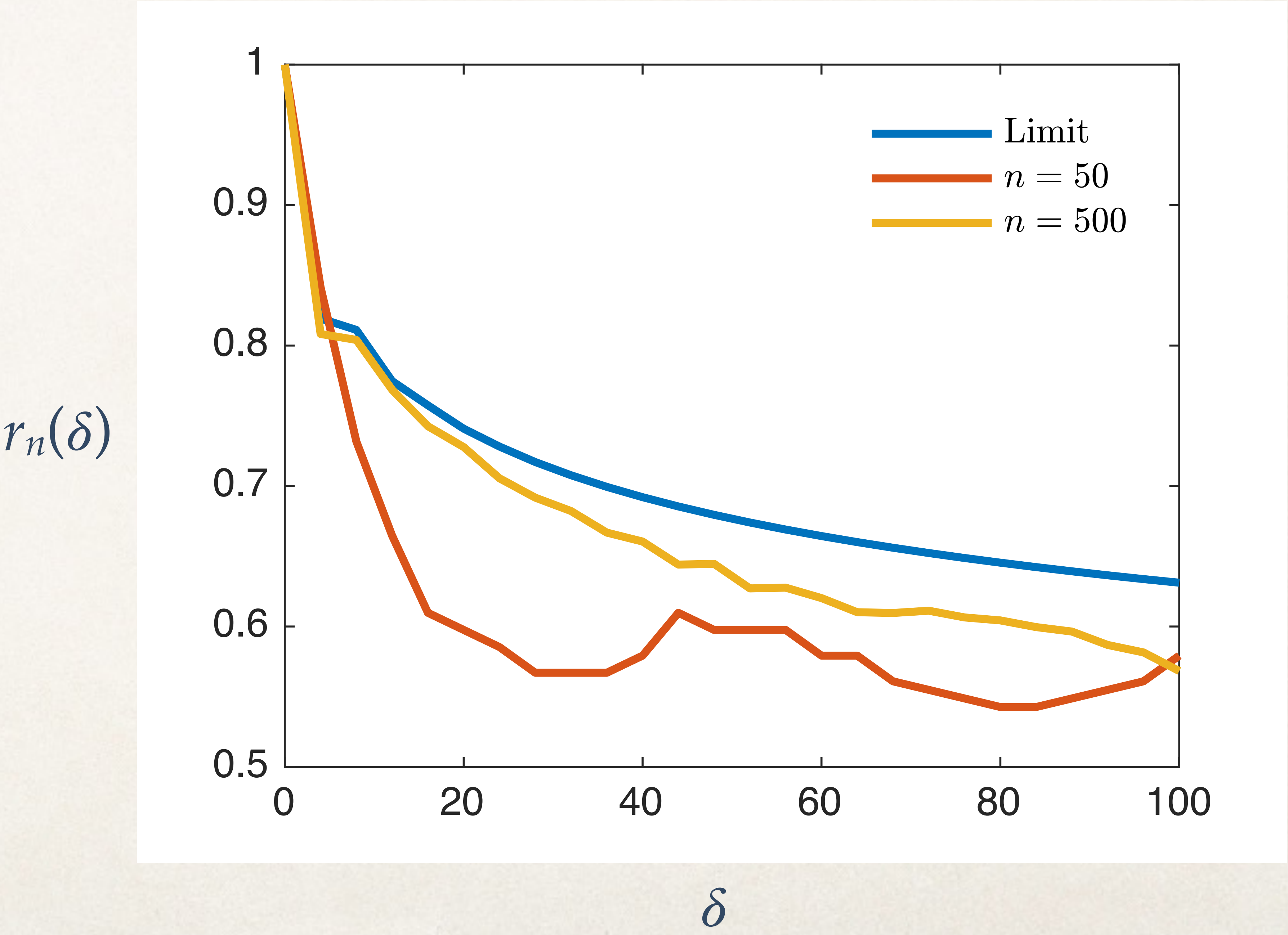
Autocorrelation Limit

P(PM) = 0.250
P(TD1) = 0.525
P(TD2) = 0.225



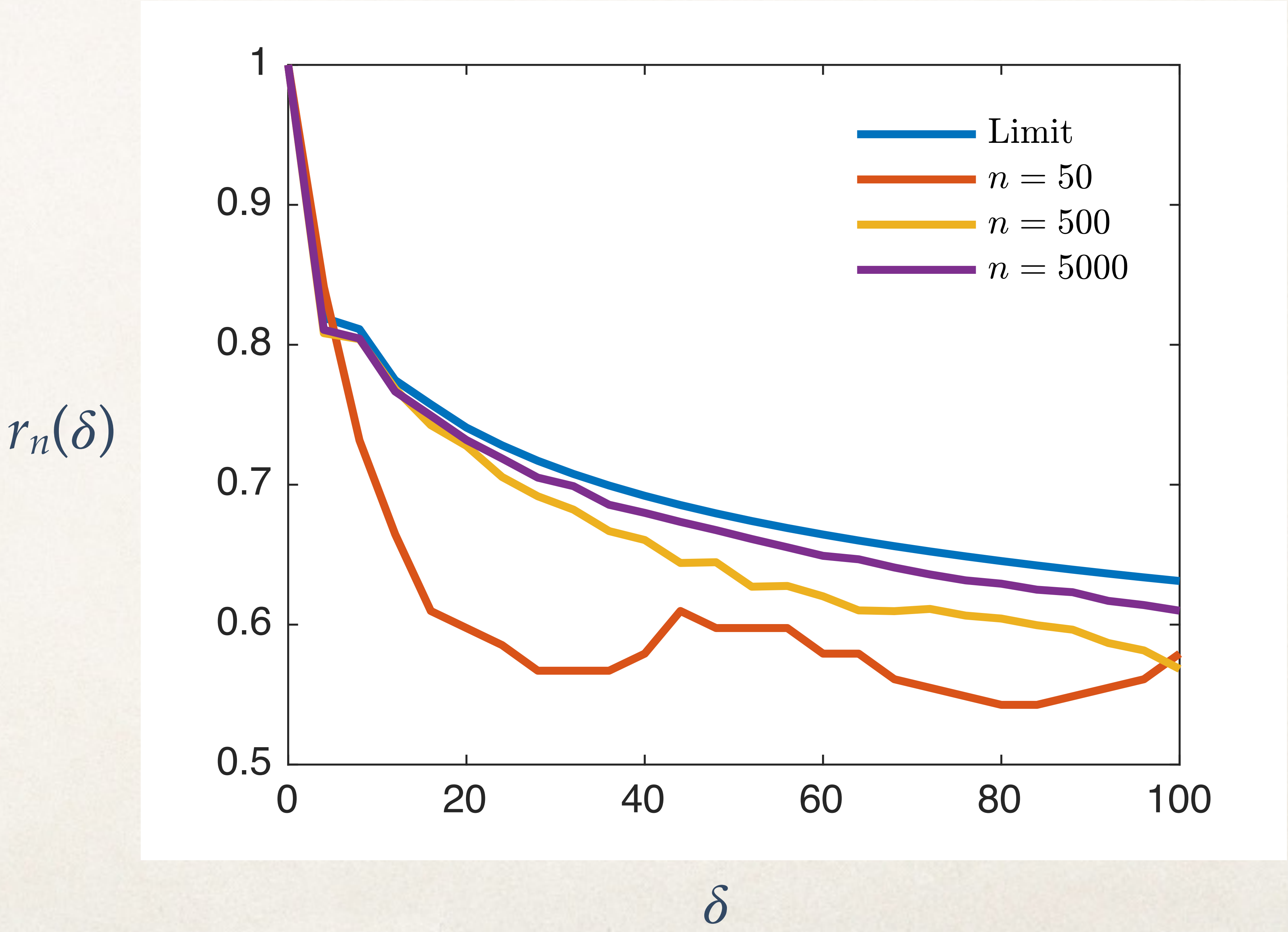
Autocorrelation Limit

P(PM) $= 0.250$
P(TD1) $= 0.525$
P(TD2) $= 0.225$



Autocorrelation Limit

P(PM) = 0.250
P(TD1) = 0.525
P(TD2) = 0.225



Estimation Algorithm

$s =$

```
GCTCCGTTACAGGTGGGCAGGGGAGGCG GCTGCGTTACAGGTGGGCAGGGGAGGCG
GCTGCGTTACAGGTGGGCAGGGGAGGCG GCTGCGTTACAGGTGGGCAGGGGAGGCG
GCTGCGTTACAGGTGGGCAGGGGAGGCG GCTCCGTTACAGGTGGGCAGGGGAGGCG
GCTGCGTTACAGGTGGGCAGGGGAGGCG GCTGCGTTACAGGTGGGCAGGGGAGGCG
GCTGCGTTACAGGTGGGCAGGGGAGGCG GCTCCGTTACAGGTGGGCAGGGGAGGCG
GCTCCGTTACAGGTGGGCAGGGGAGGCG GCTGCGTTACAGGTGGGCAGGGGAGGCG
GCTGCGTTACAGGTGGGCGGGGGAGGCG GCTGCGTTACAGGTGGGCGGGGGAGGCG
GCTGCGTTACAGGTGGGCAGGGGGGGCG GCTGCGTTACAGGTGGGCGGGGGAGGCG
GCTGCGTTACAGGTGGGCGGGGGAGGCG GCTCCGTTACAGGTGGGCGGGGGAGGCG
GCTGCGTTACAGGTGGGCGGGGGAGGCG GCTGCGTTACAGGTGGGCGGGGGGGGCG
GCTGCGTTACAGGTGGGCGGGGGAGGCT GCTCCGTTACAGGTGGGCGGGGGAGGCT
GCTCCGTTACAGGTGGGCGGGGGGGGCG GCTGCGTTACAGGTGGGCGGGGGGGGCG
GCTGCGTTACAGGTGGGCGGGGGAGGCG GCTGCGTTACAGGTGGGCGGGGGAGGCG
GCTCCGTTACAGGTGGGCGGGGGAGGCG GCTGCGTTACAGGTGGGCGGGGGAGGCG
GCTGCGTTACAGGTGGGCAGGGGAGGCG GCTGCGTTACAGGTGGGCAGGGGAGGCG
GCTGCGTTACAGGTGGGCGGGGGAGGCG GCTCCGTTACAGGTGGGCGGGGGAGGCG
GCTGCGTTACAGGTGGGCGGGGGAGGCG GCTGCGTTACAGGTGGGCGGGGGAGGCG
```

1. Calculate autocorrelation r of s.
2. Find mutation probs such that the *l_2-norm* $\|Ar\|_2$ is minimized.

Estimation Algorithm

$s =$

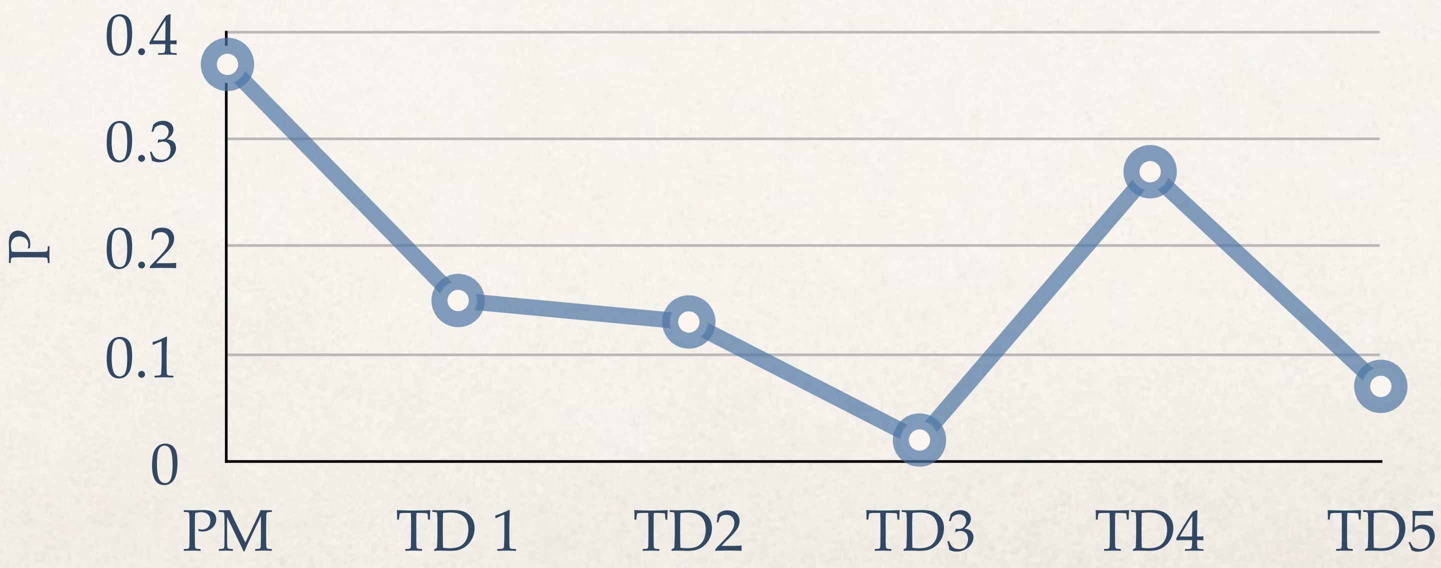
```
GCTCCGTTACAGGTGGGCAGGGGAGGCG GCTGCGTTACAGGTGGGCAGGGGAGGCG
GCTGCGTTACAGGTGGGCAGGGGAGGCG GCTGCGTTACAGGTGGGCAGGGGAGGCG
GCTGCGTTACAGGTGGGCAGGGGAGGCG GCTCCGTTACAGGTGGGCAGGGGAGGCG
GCTGCGTTACAGGTGGGCAGGGGAGGCG GCTGCGTTACAGGTGGGCAGGGGAGGCG
GCTGCGTTACAGGTGGGCAGGGGAGGCG GCTCCGTTACAGGTGGGCAGGGGAGGCG
GCTCCGTTACAGGTGGGCAGGGGAGGCG GCTGCGTTACAGGTGGGCAGGGGAGGCG
GCTGCGTTACAGGTGGGCGGGGGAGGCG GCTGCGTTACAGGTGGGCGGGGGAGGCG
GCTGCGTTACAGGTGGGCAGGGGGGGCG GCTGCGTTACAGGTGGGCGGGGGAGGCG
GCTGCGTTACAGGTGGGCGGGGGAGGCG GCTCCGTTACAGGTGGGCGGGGGAGGCG
GCTGCGTTACAGGTGGGCGGGGGAGGCG GCTGCGTTACAGGTGGGCGGGGGGGGCG
GCTGCGTTACAGGTGGGCGGGGGAGGCT GCTCCGTTACAGGTGGGCGGGGGAGGCT
GCTCCGTTACAGGTGGGCGGGGGGGGCG GCTGCGTTACAGGTGGGCGGGGGGGGCG
GCTGCGTTACAGGTGGGCGGGGGAGGCG GCTGCGTTACAGGTGGGCGGGGGAGGCG
GCTCCGTTACAGGTGGGCGGGGGAGGCG GCTGCGTTACAGGTGGGCGGGGGAGGCG
GCTGCGTTACAGGTGGGCAGGGGAGGCG GCTGCGTTACAGGTGGGCAGGGGAGGCG
GCTGCGTTACAGGTGGGCGGGGGAGGCG GCTCCGTTACAGGTGGGCGGGGGAGGCG
GCTGCGTTACAGGTGGGCGGGGGAGGCG GCTGCGTTACAGGTGGGCGGGGGAGGCG
```

1. Calculate autocorrelation r of s.
2. Find mutation probs such that the l_2-*norm* $||Ar||_2$ is minimized.

Estimation Algorithm

$s =$

```
GCTCCGTTACAGGTGGGCAGGGGAGGCG GCTGCGTTACAGGTGGGCAGGGGAGGCG
GCTGCGTTACAGGTGGGCAGGGGAGGCG GCTGCGTTACAGGTGGGCAGGGGAGGCG
GCTGCGTTACAGGTGGGCAGGGGAGGCG GCTCCGTTACAGGTGGGCAGGGGAGGCG
GCTGCGTTACAGGTGGGCAGGGGAGGCG GCTGCGTTACAGGTGGGCAGGGGAGGCG
GCTGCGTTACAGGTGGGCAGGGGAGGCG GCTCCGTTACAGGTGGGCAGGGGAGGCG
GCTCCGTTACAGGTGGGCAGGGGAGGCG GCTGCGTTACAGGTGGGCAGGGGAGGCG
GCTGCGTTACAGGTGGGCGGGGGAGGCG GCTGCGTTACAGGTGGGCGGGGGAGGCG
GCTGCGTTACAGGTGGGCAGGGGGGGCG GCTGCGTTACAGGTGGGCGGGGGAGGCG
GCTGCGTTACAGGTGGGCGGGGGAGGCG GCTCCGTTACAGGTGGGCGGGGGAGGCG
GCTGCGTTACAGGTGGGCGGGGGAGGCG GCTGCGTTACAGGTGGGCGGGGGGGGCG
GCTGCGTTACAGGTGGGCGGGGGAGGCT GCTCCGTTACAGGTGGGCGGGGGAGGCT
GCTCCGTTACAGGTGGGCGGGGGGGGCG GCTGCGTTACAGGTGGGCGGGGGGGGCG
GCTGCGTTACAGGTGGGCGGGGGAGGCG GCTGCGTTACAGGTGGGCGGGGGAGGCG
GCTCCGTTACAGGTGGGCGGGGGAGGCG GCTGCGTTACAGGTGGGCGGGGGAGGCG
GCTGCGTTACAGGTGGGCAGGGGAGGCG GCTGCGTTACAGGTGGGCAGGGGAGGCG
GCTGCGTTACAGGTGGGCGGGGGAGGCG GCTCCGTTACAGGTGGGCGGGGGAGGCG
GCTGCGTTACAGGTGGGCGGGGGAGGCG GCTGCGTTACAGGTGGGCGGGGGAGGCG
```



Simulation

Simulation

- Start with a short random seed over {A,C,G,T}

TGAATGT

Simulation

- Start with a short random seed over {A,C,G,T}
- Choose the parameters $\mathbf{q} = (PM1, TD1, TD2, TD3)$ randomly

TGAATGT

$$\mathbf{q} = (0.24, 0.33, 0.34, 0.09)$$

Simulation

- Start with a short random seed over {A,C,G,T}
- Choose the parameters $\mathbf{q} = (PM1, TD1, TD2, TD3)$ randomly
- Apply n random mutations

TGAATGT

$$\mathbf{q} = (0.24, 0.33, 0.34, 0.09)$$

200 mutations: TGAATGTGCGT…

Simulation

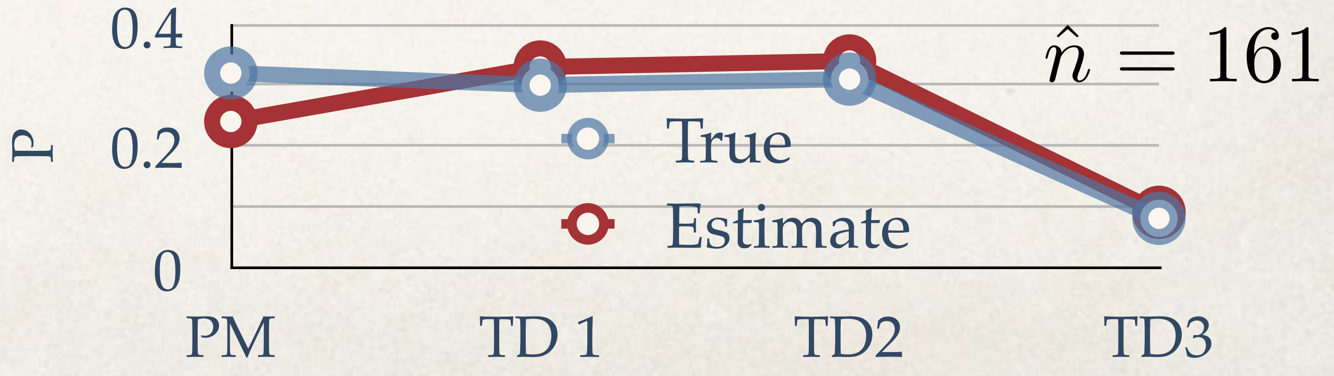
- Start with a short random seed over {A,C,G,T}
- Choose the parameters **q** = (*PM1, TD1, TD2, TD3)* randomly
- Apply *n* random mutations
- Estimate the parameters

$$\min_{\hat{\mathbf{q}}} \|Ar\|_2$$

TGAATGT

$\mathbf{q}$ = (*0.24, 0.33, 0.34, 0.09*)

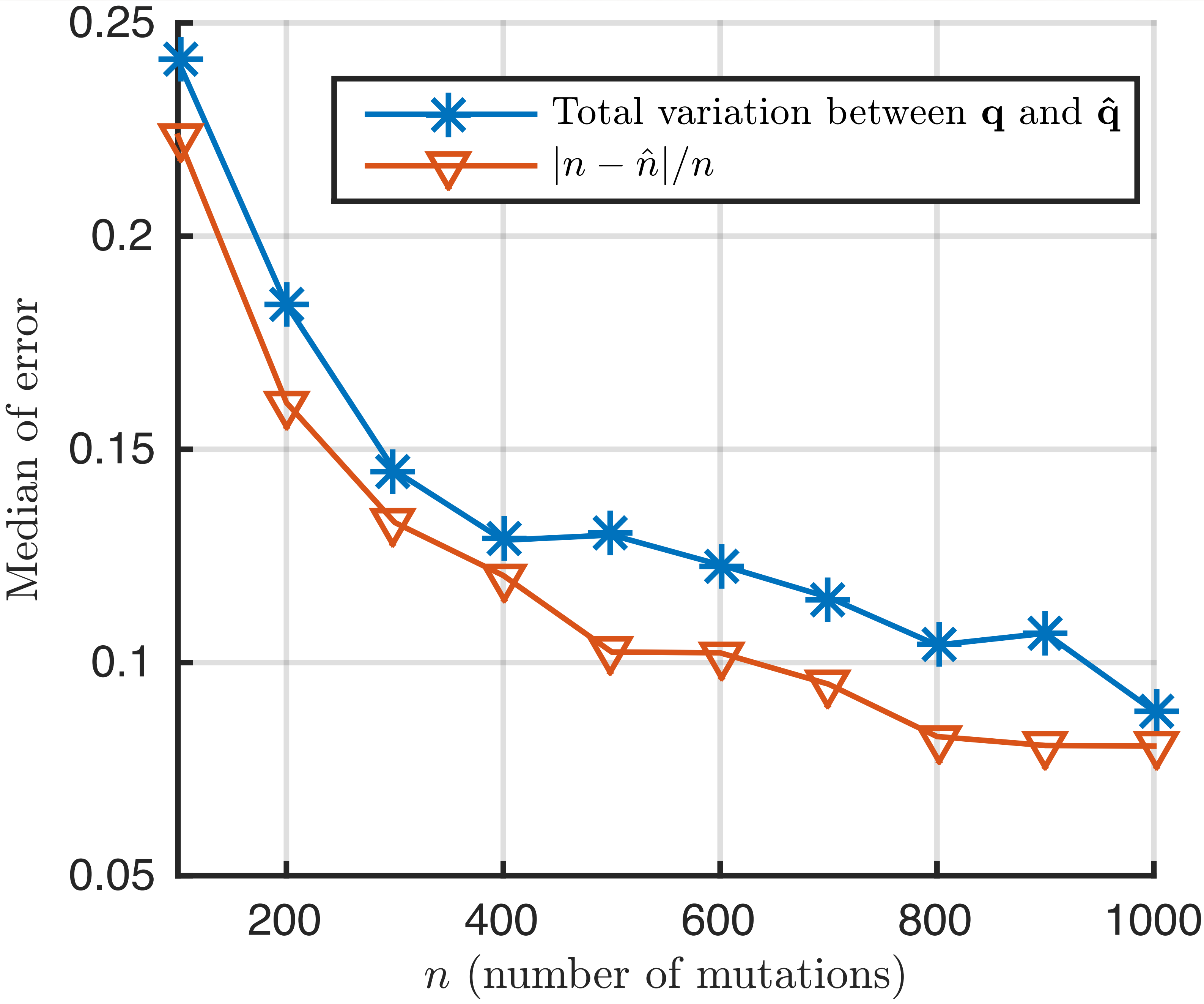
200 mutations: TGAATGTGCGT…

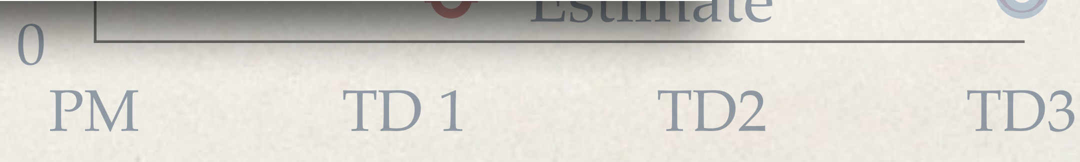


Simulation

- Start with a
  seed over {
- Choose the … 0.34, 0.09)
  q = (*PM1, T*
  randomly
- Apply *n* ra … ATGTGCGT…
- Estimate th

$$\min_{\hat{\mathbf{q}}} ||Ar||_2$$





Summary and Next Steps

Summary and Next Steps

- Stochastic estimation algorithm (NP-Hard(?) combinatorial problem).

Summary and Next Steps

- Stochastic estimation algorithm (NP-Hard(?) combinatorial problem).
- Point mutation enables estimation of duplication lengths.

Summary and Next Steps

- Stochastic estimation algorithm (NP-Hard(?) combinatorial problem).
- Point mutation enables estimation of duplication lengths.

CAATGTAC… → Noisy Channel → CAATATGGAC…

Data encoded in DNA

Noisy Output

- Capacity? Error-correcting codes?
  Synthetic biology algorithms using storage, performing encoding/decoding?

Summary and Next Steps

- Stochastic estimation algorithm (NP-Hard(?) combinatorial problem).
- Point mutation enables estimation of duplication lengths.

CAATGTAC…

Data encoded in DNA

Noisy Channel

CAATATGGAC…

Noisy Output

- Capacity? Error-correcting codes?
  Synthetic biology algorithms using storage, performing encoding/decoding?
- Application to phylogenetics (estimation of # mutations of each type)