Stochastic Models for DNA Tandem Duplication

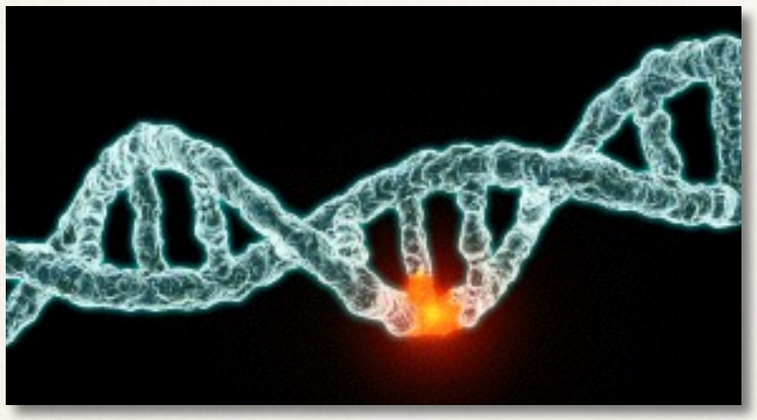
Farzad Farnoud, with M. Schwartz, J. Bruck

Jan 18, 2016, University of Washington

Caltech

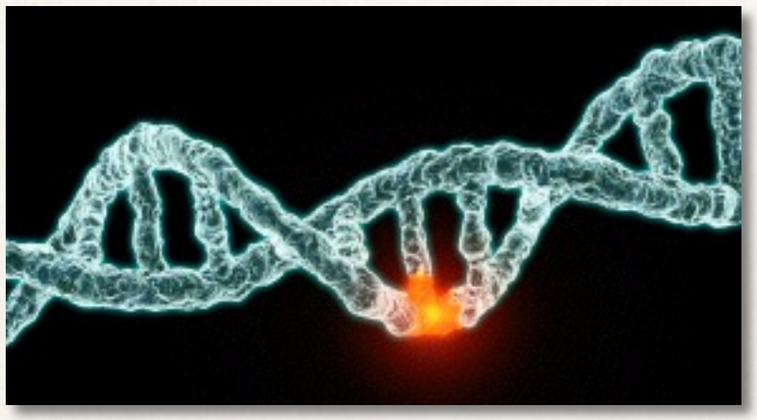
Mutations

Mutations



8.7 million species

Mutations



Data storage in *live DNA*

8.7 million species

Types of Mutations

TGATGCA

↓ Point Mutation

TCATGCA

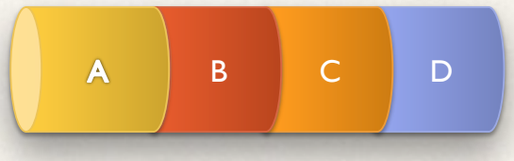
Types of Mutations

TGATGCA

↓ Point Mutation

TCATGCA

↓ Deletion



↓ Insertion

Types of Mutations

TGATGCA

↓ Point Mutation

TCATGCA

↓ Deletion

↓ Insertion

↓ Tandem Duplication

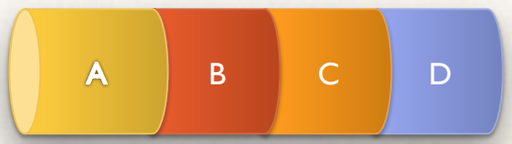
↓ Interspersed Duplication

Types of Mutations

TGATGCA

↓ Point Mutation

TCATGCA



↓ Deletion

↓ Insertion

↓ Tandem Duplication

3% of human genome

↓ Interspersed Duplication

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
GCTCCGTTACAGGTGGGCAGGGGAGGCG GCTCCGTTACAGGTGGGCAGGGGAGGCG
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
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

Point mutations are in the same positions

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

Stochastic Model

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

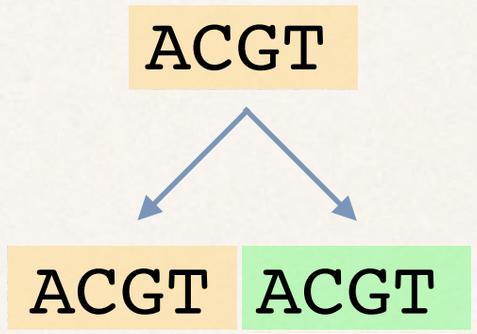
Stochastic Model

ACGT

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

Stochastic Model

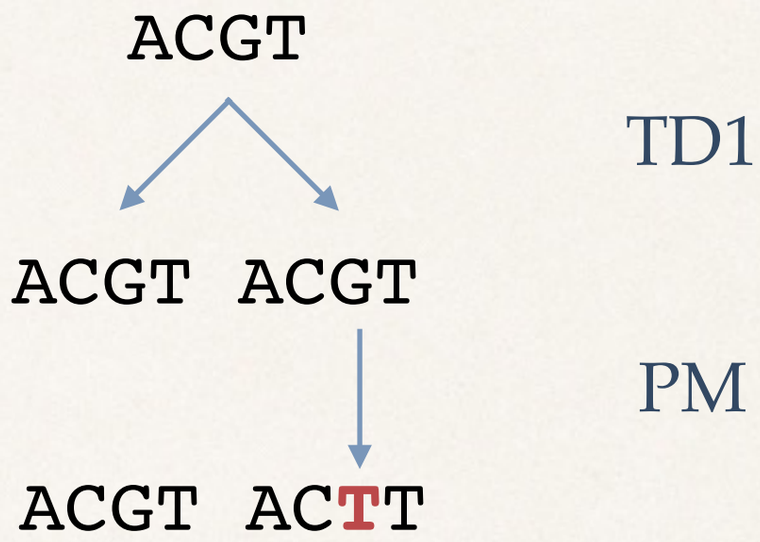
Start from one repeat unit (***seed***).
Random mutations:
- Tandem duplications (TD)
  of one or more repeat units
- Point mutations (PM)



TD1

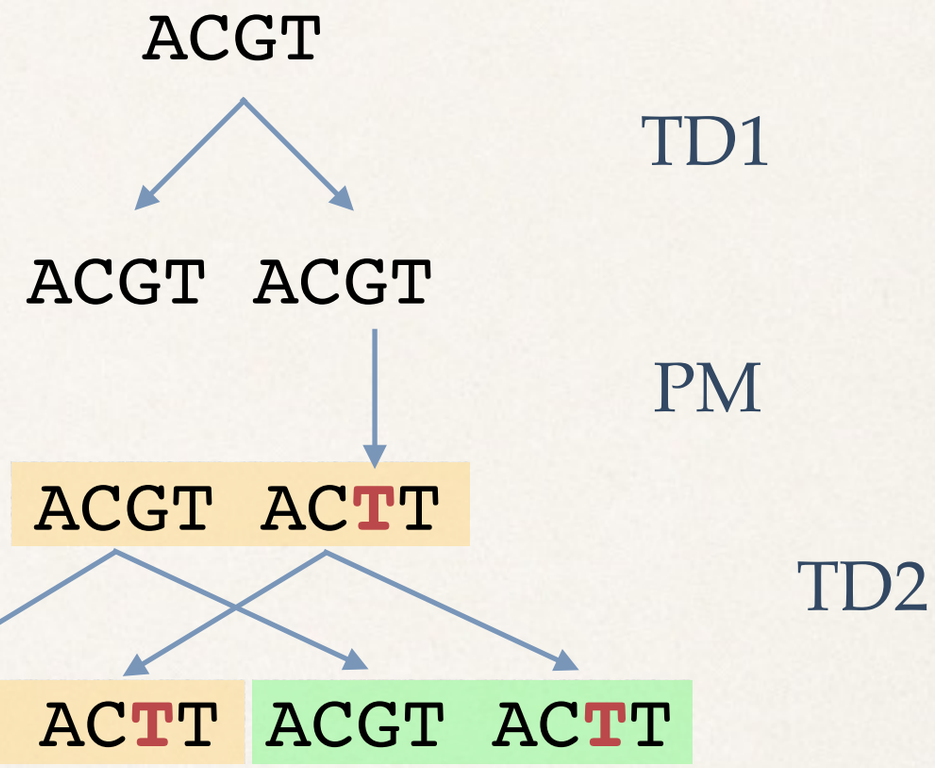
Stochastic Model

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



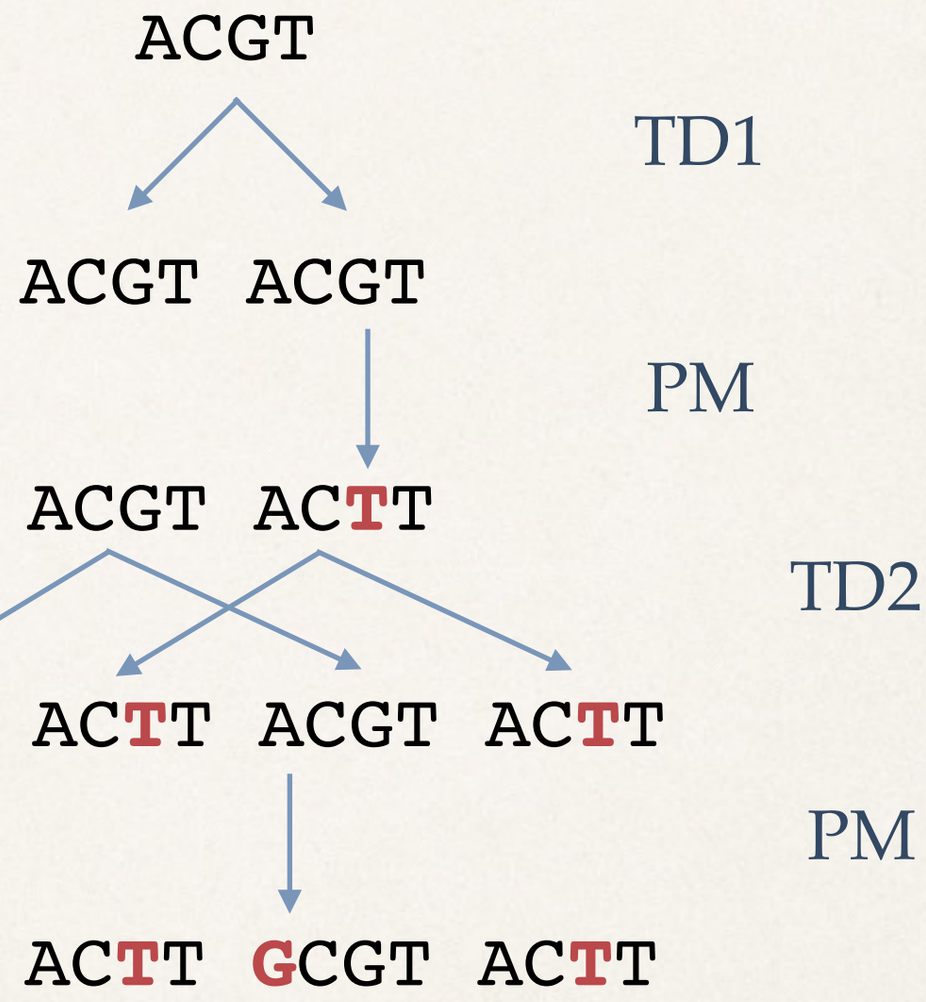
Stochastic Model

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



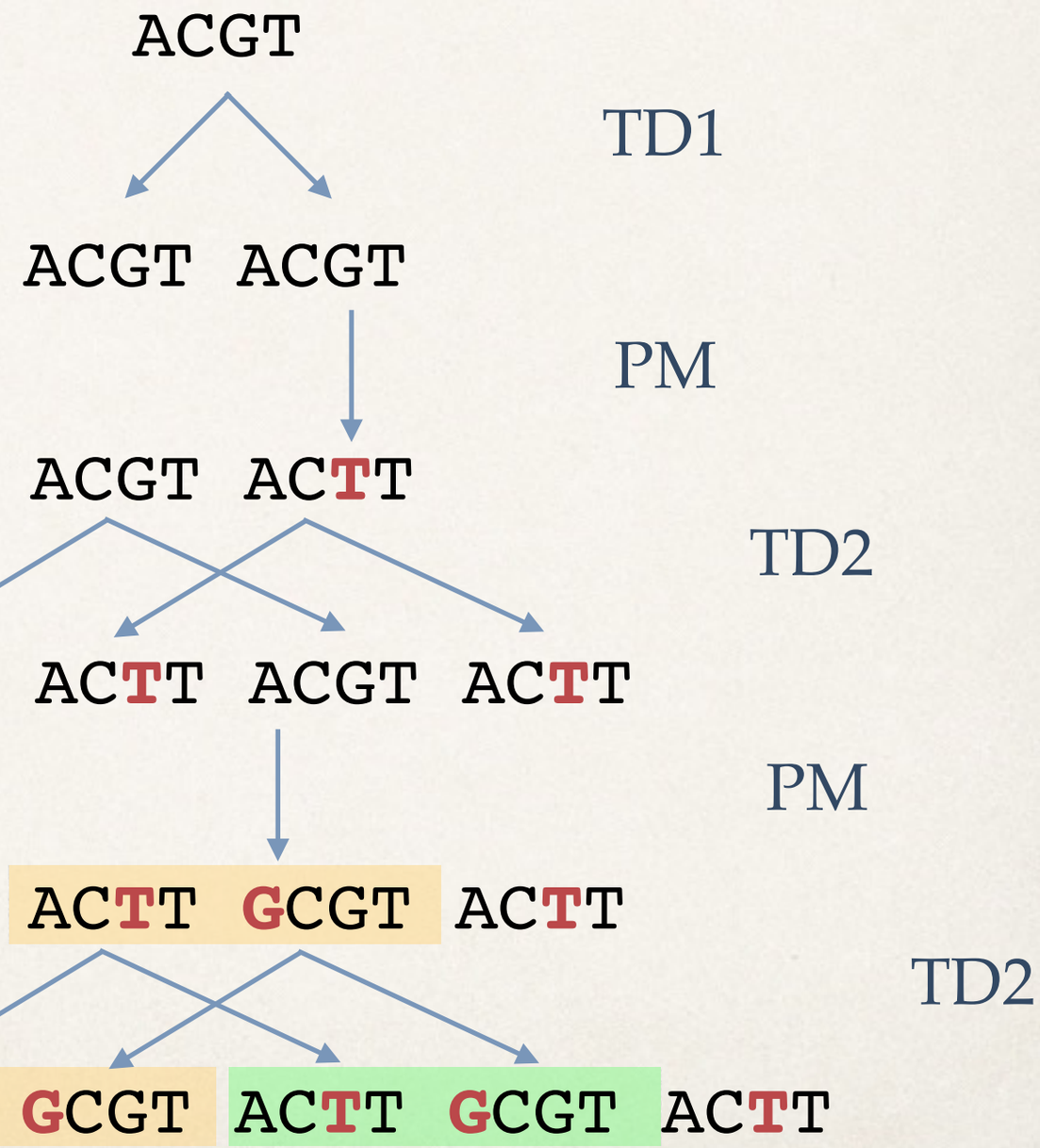
Stochastic Model

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)

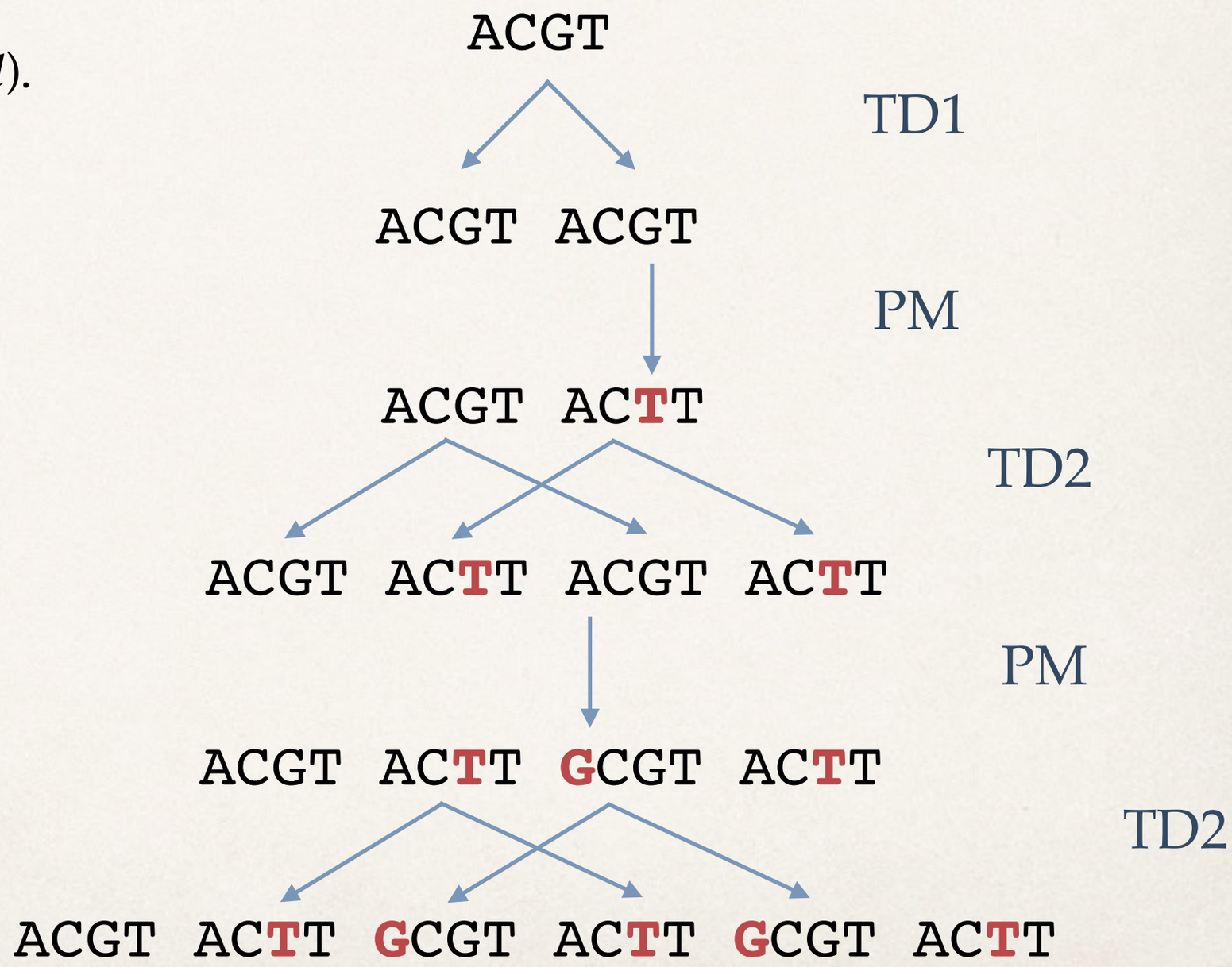


Stochastic Model

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?



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

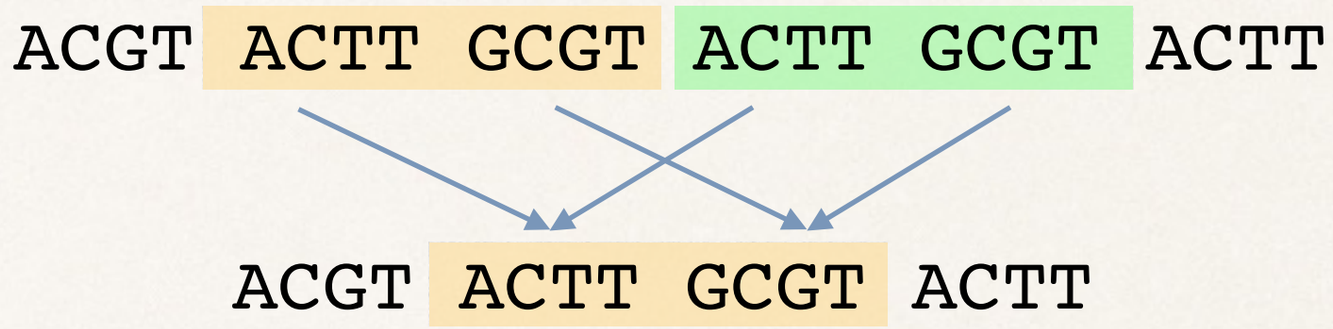
Finding Duplication History

ACGT ACTT GCGT ACTT GCGT ACTT

Finding Duplication History

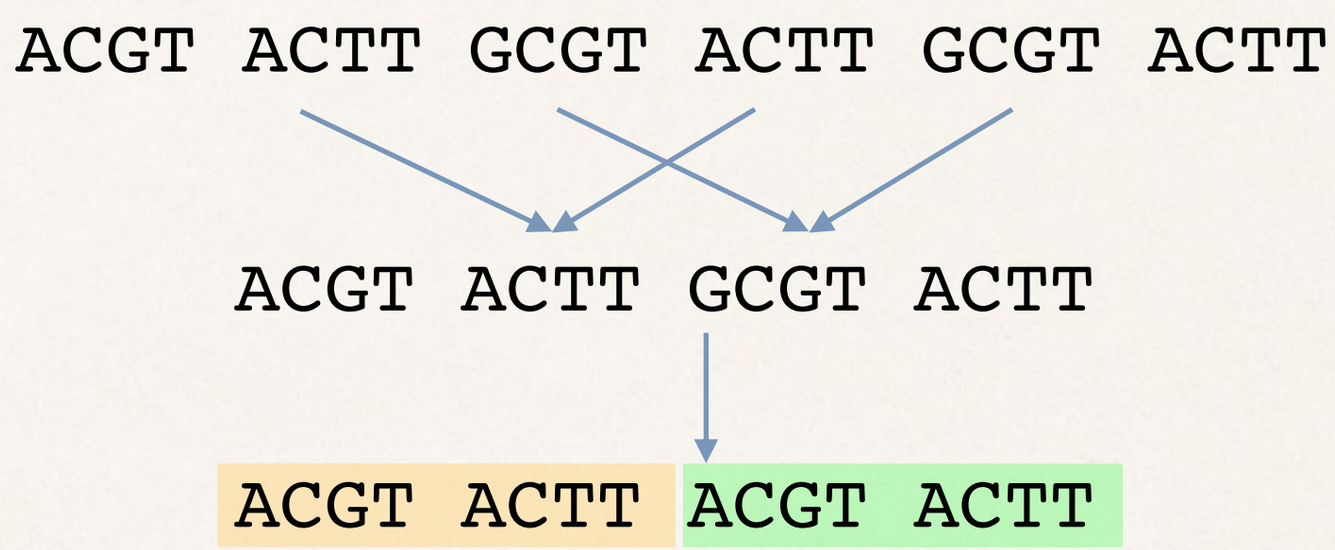
ACGT ACTT GCGT ACTT GCGT ACTT

Finding Duplication History



TD2

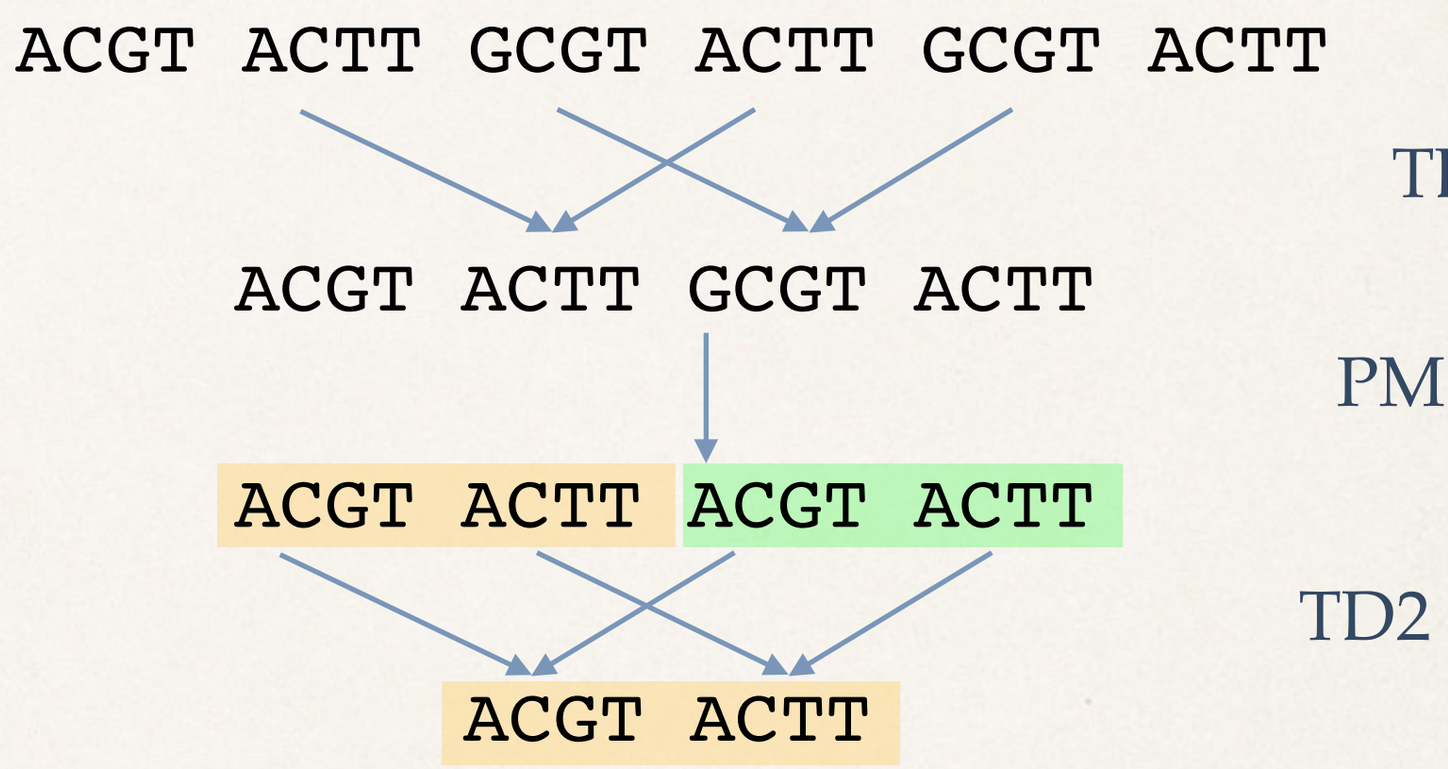
Finding Duplication History



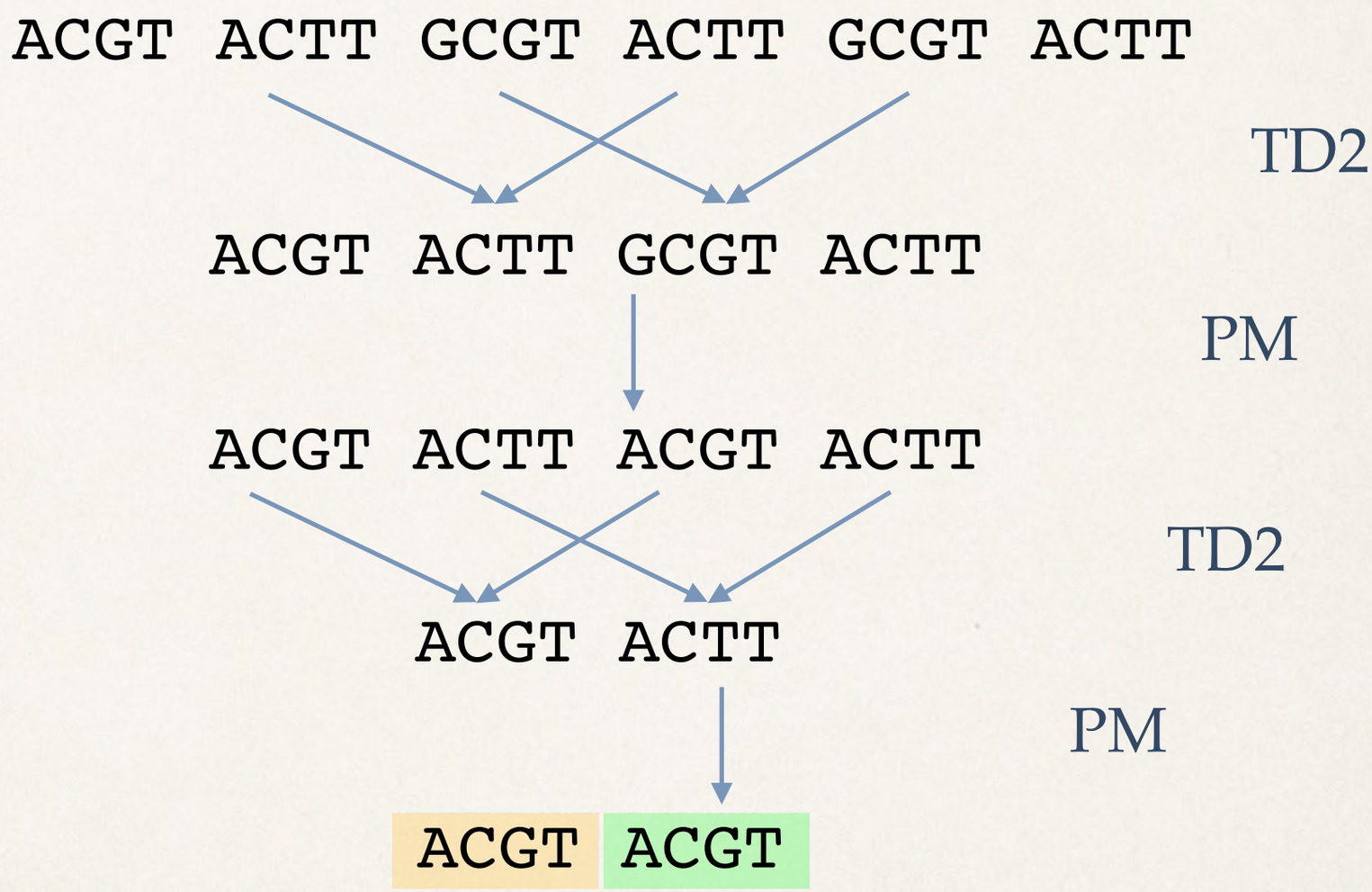
TD2

PM

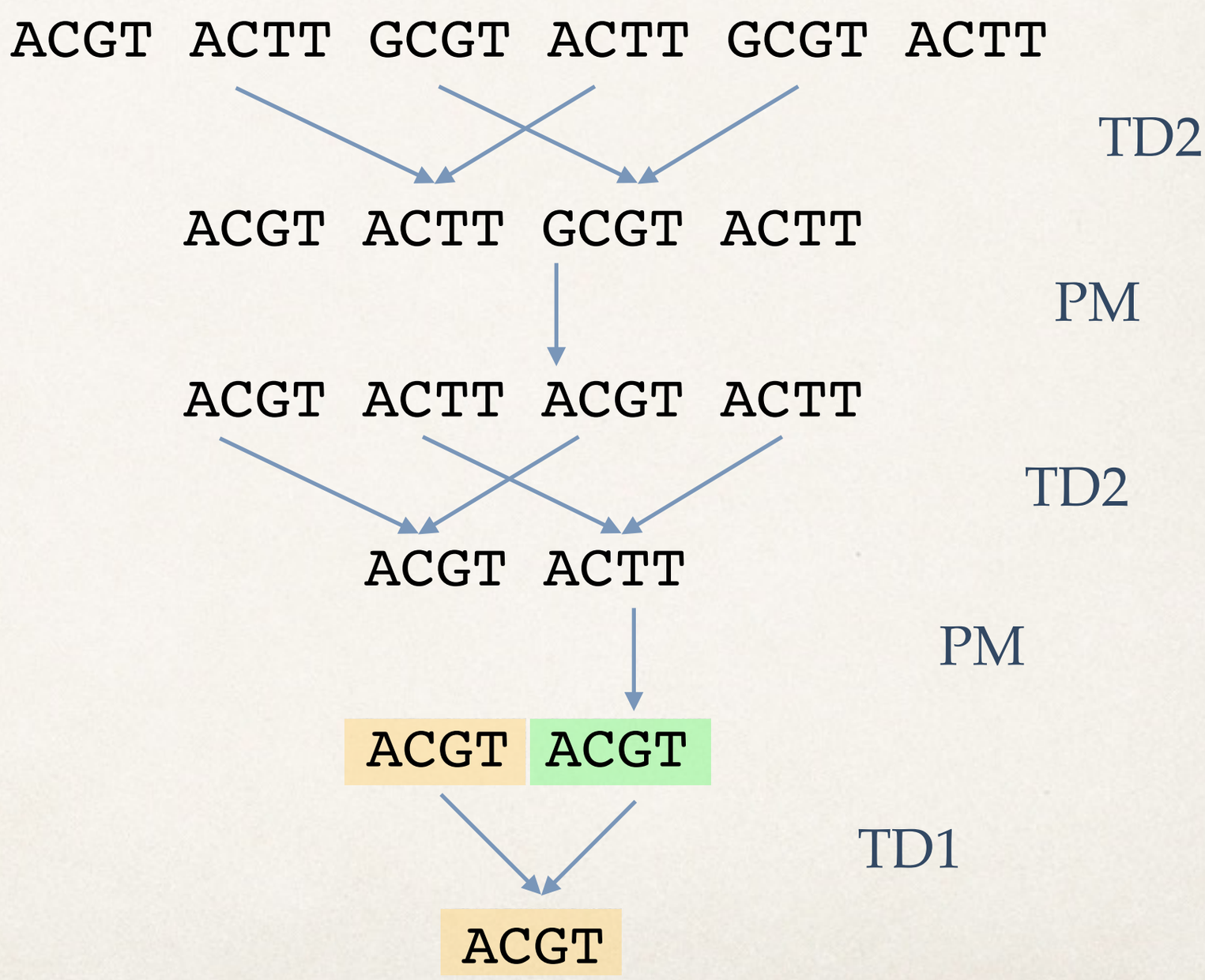
Finding Duplication History



Finding Duplication History

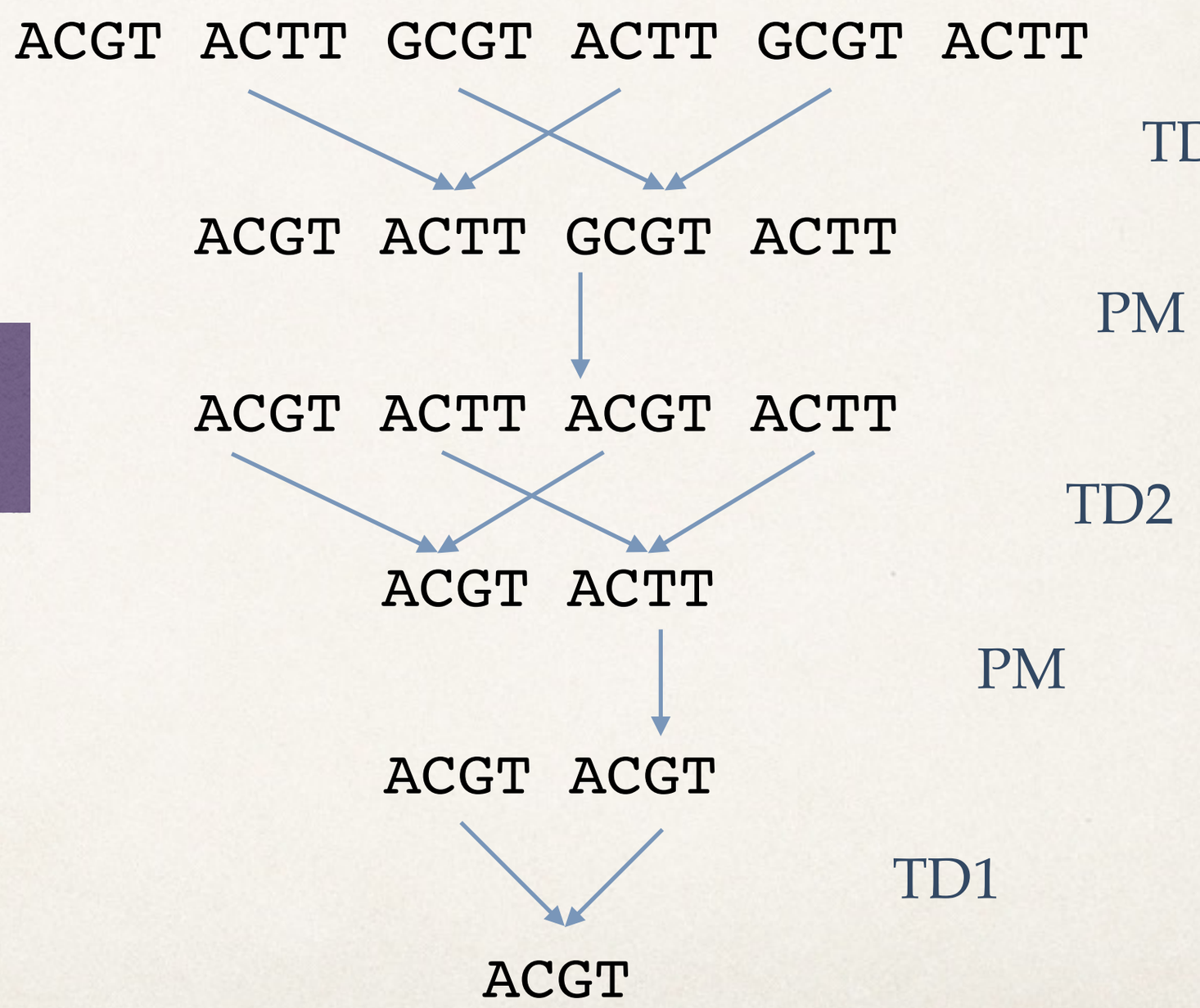


Finding Duplication History



Finding Duplication History

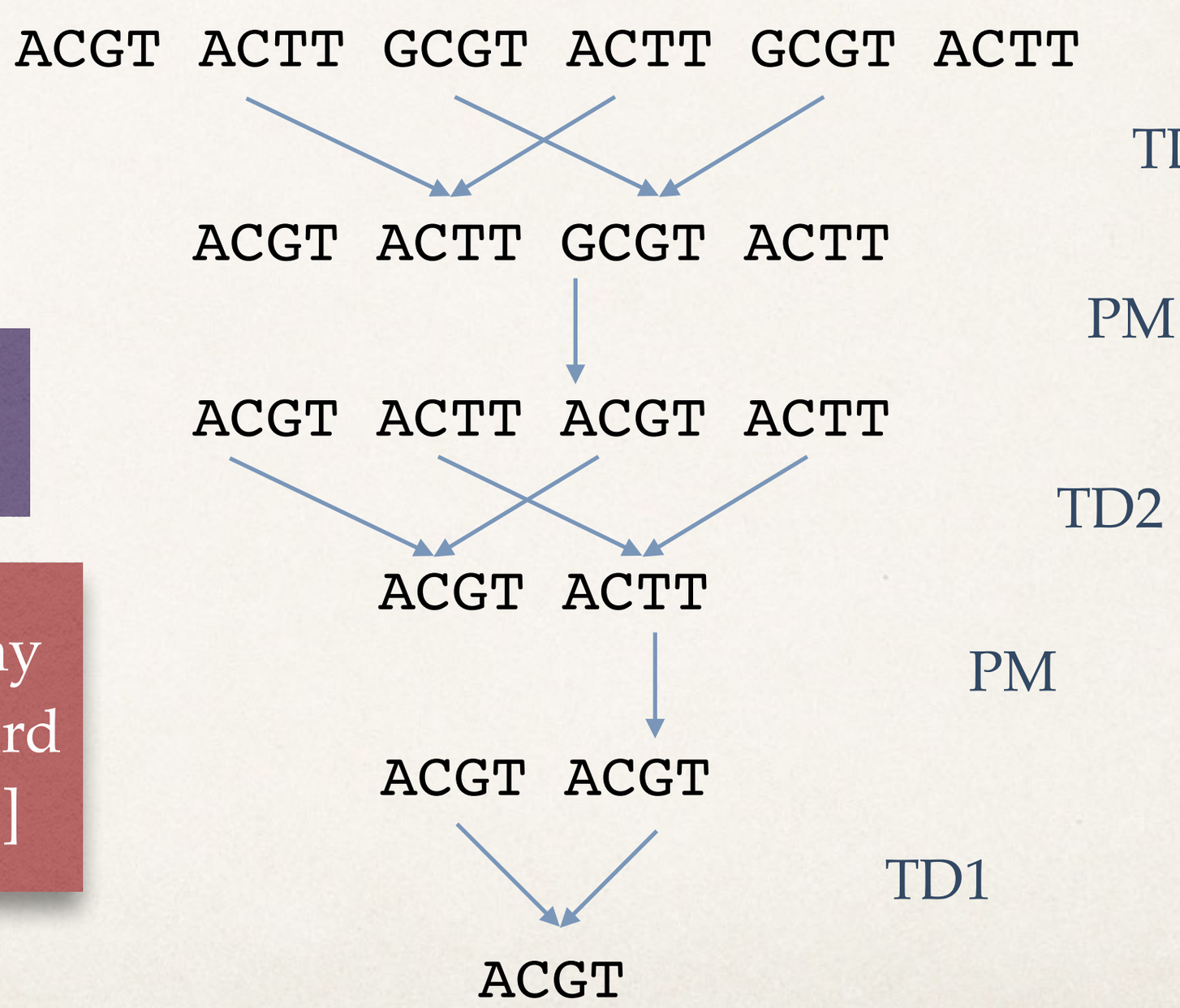
TD1, 2 TD2, 2 PM



Finding Duplication History

TD1, 2 TD2, 2 PM

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

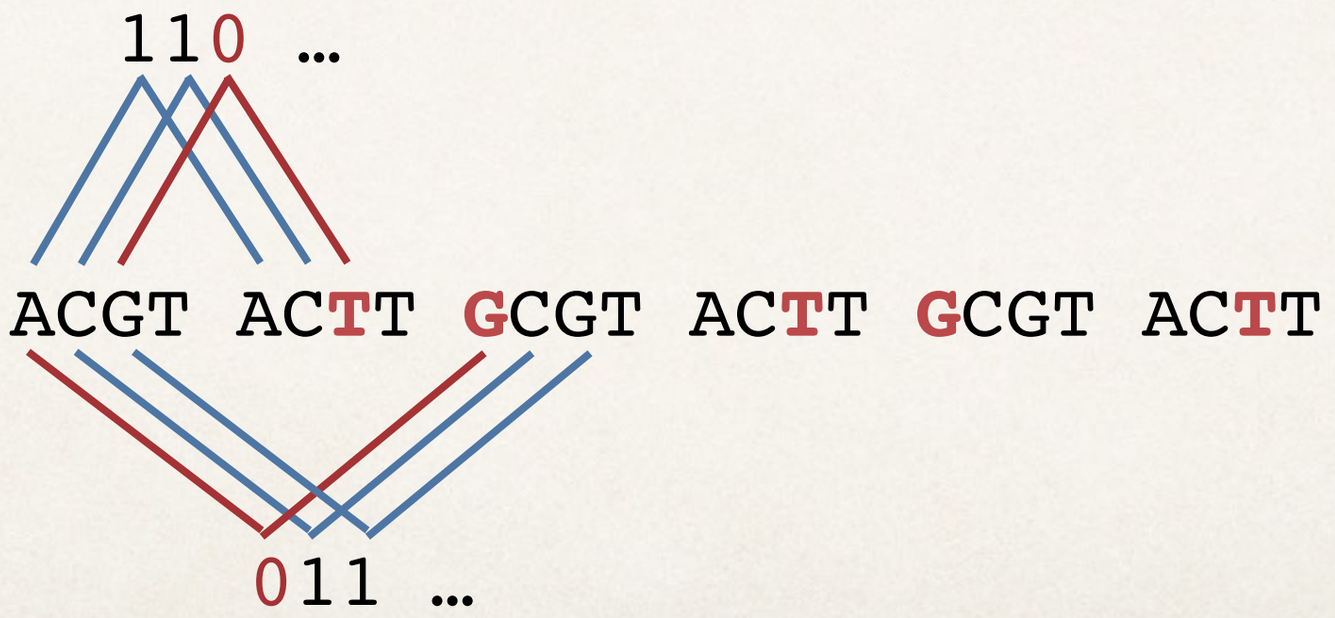


Given the final sequence, can we efficiently estimate the parameters?

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

How to extract information from point mutations?

- Autocorrelation function:
  - ★ $r(\delta)$ = *fraction of symbols at distance δ units that are the same*



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:

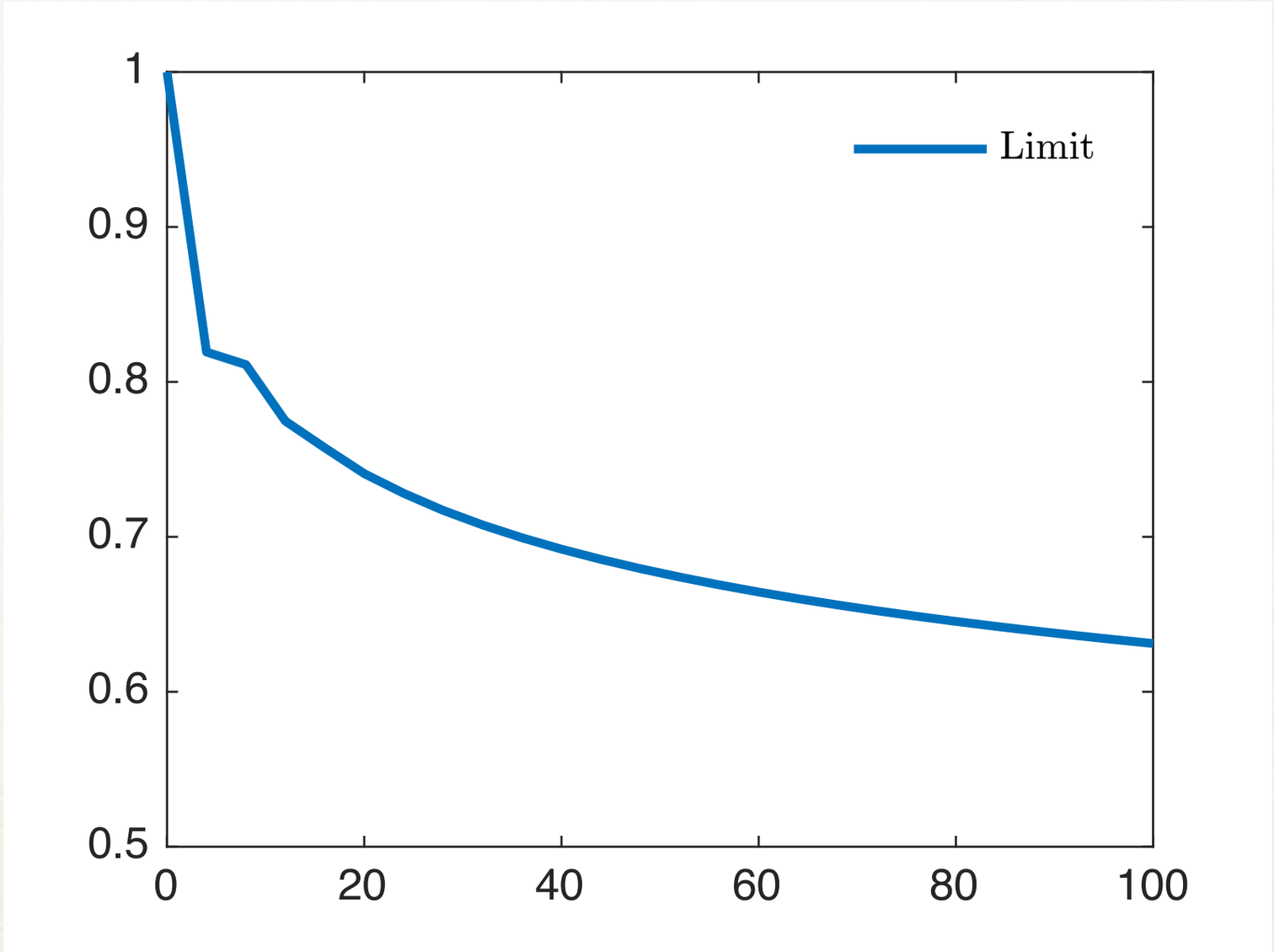
$$\frac{d}{dt} r_t = A r_t$$

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*

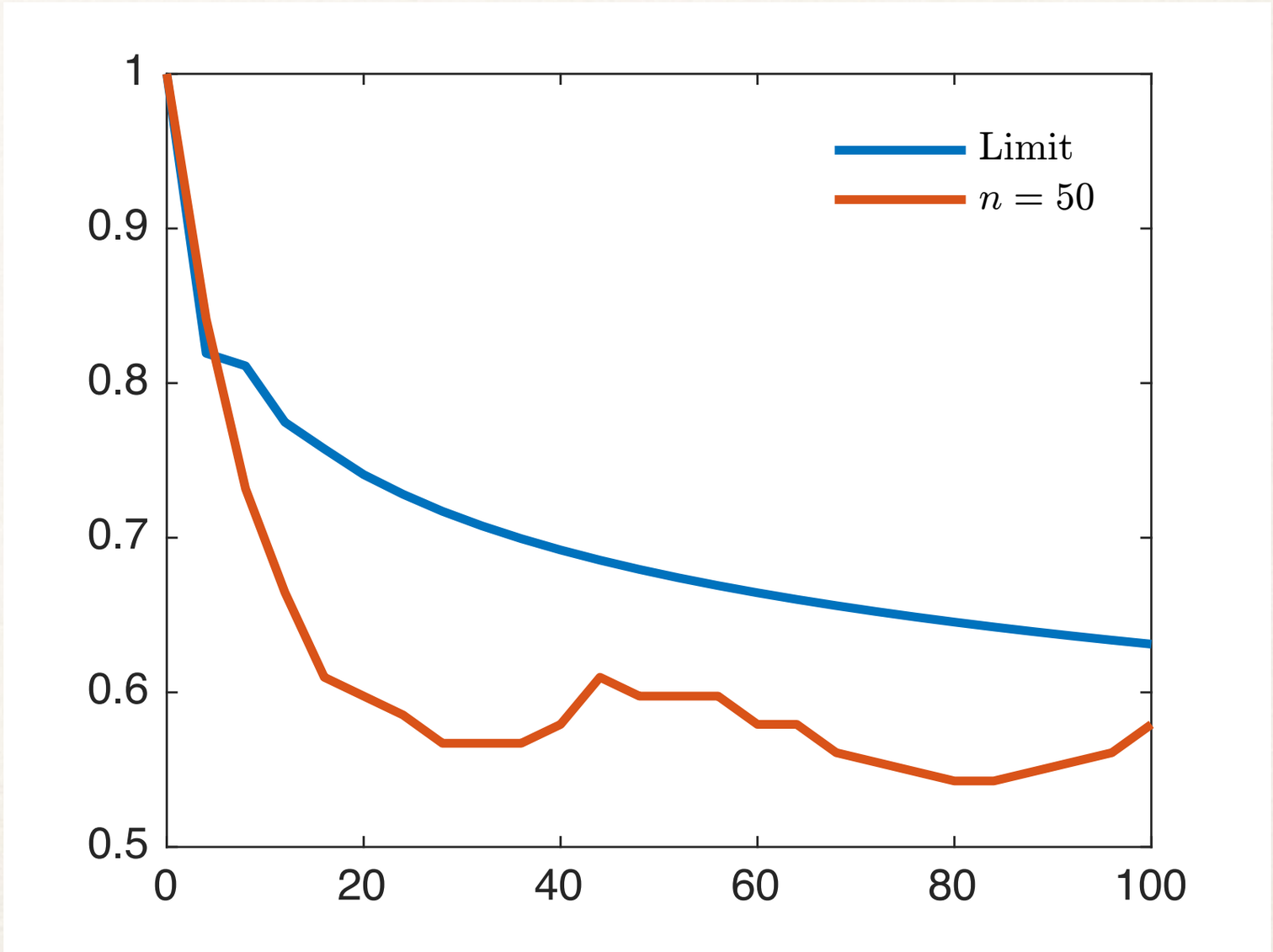
Autocorrelation Limit

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



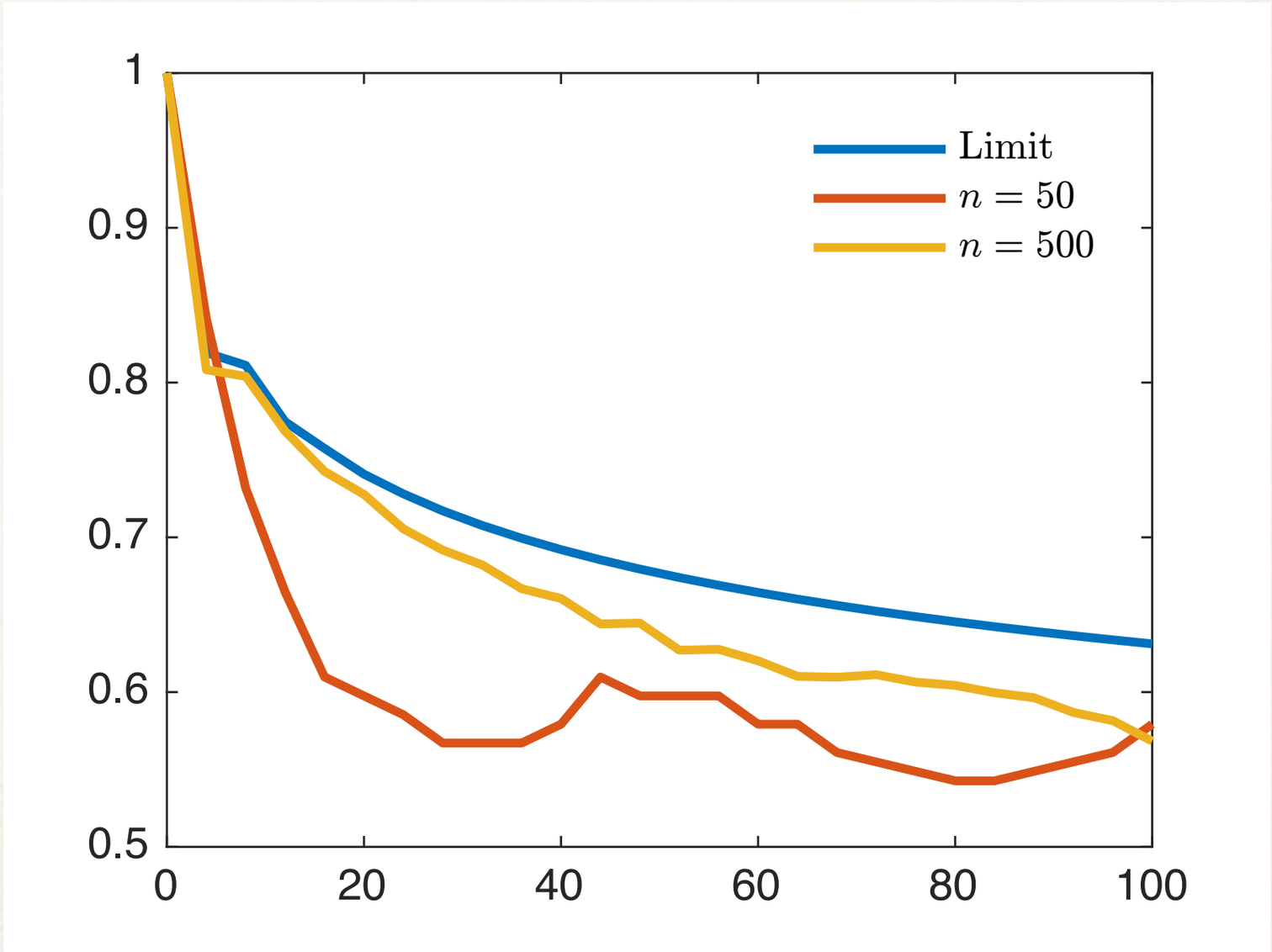
Autocorrelation Limit

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



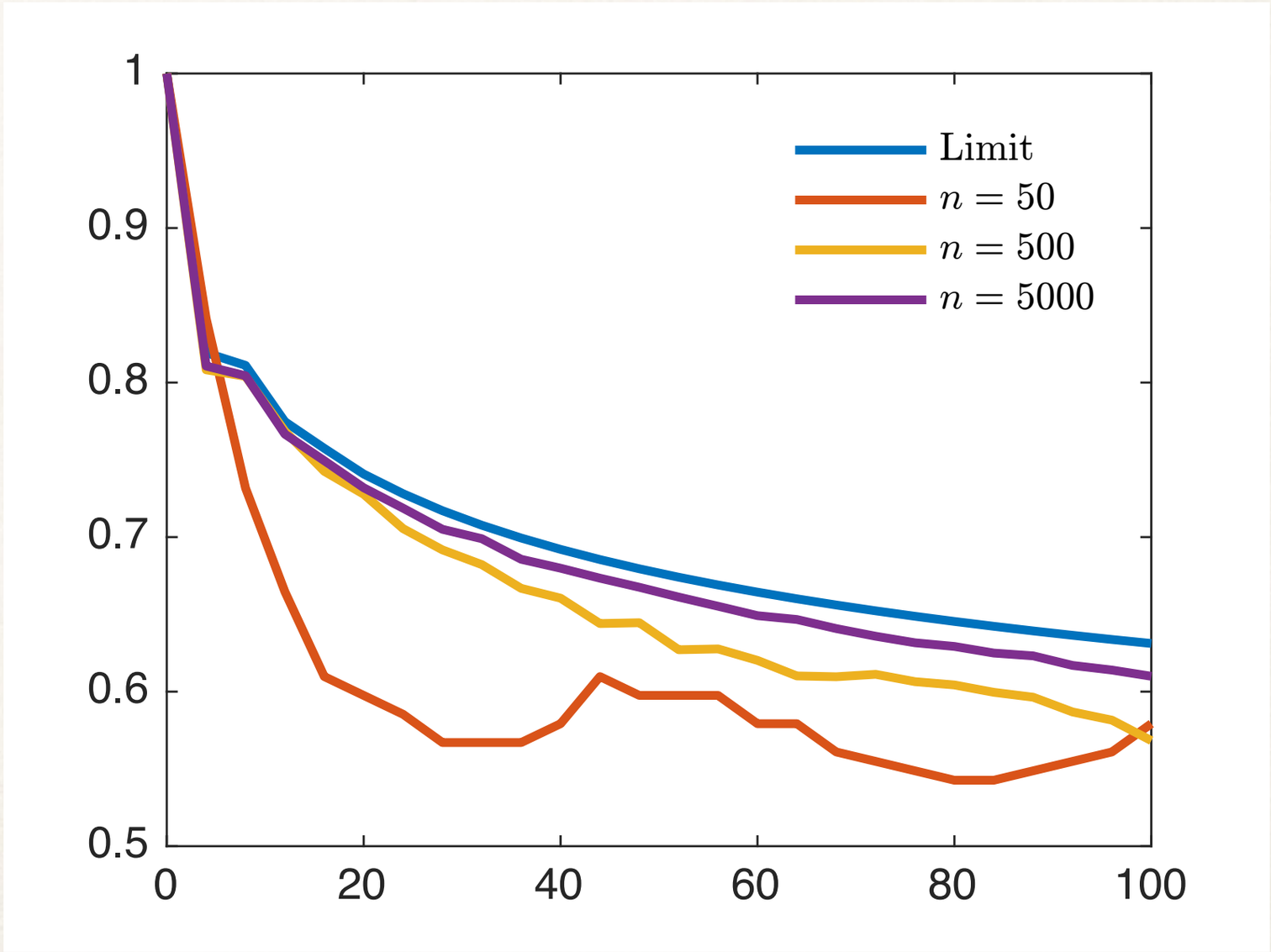
Autocorrelation Limit

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



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
GCTCCGTTACAGGTGGGCGGGGGAGGCT GCTCCGTTACAGGTGGGCGGGGGAGGCT
GCTCCGTTACAGGTGGGCGGGGGGGGCG GCTGCGTTACAGGTGGGCGGGGGGGGCG
GCTGCGTTACAGGTGGGCGGGGGAGGCG GCTCCGTTACAGGTGGGCGGGGGAGGCG
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

$s =$

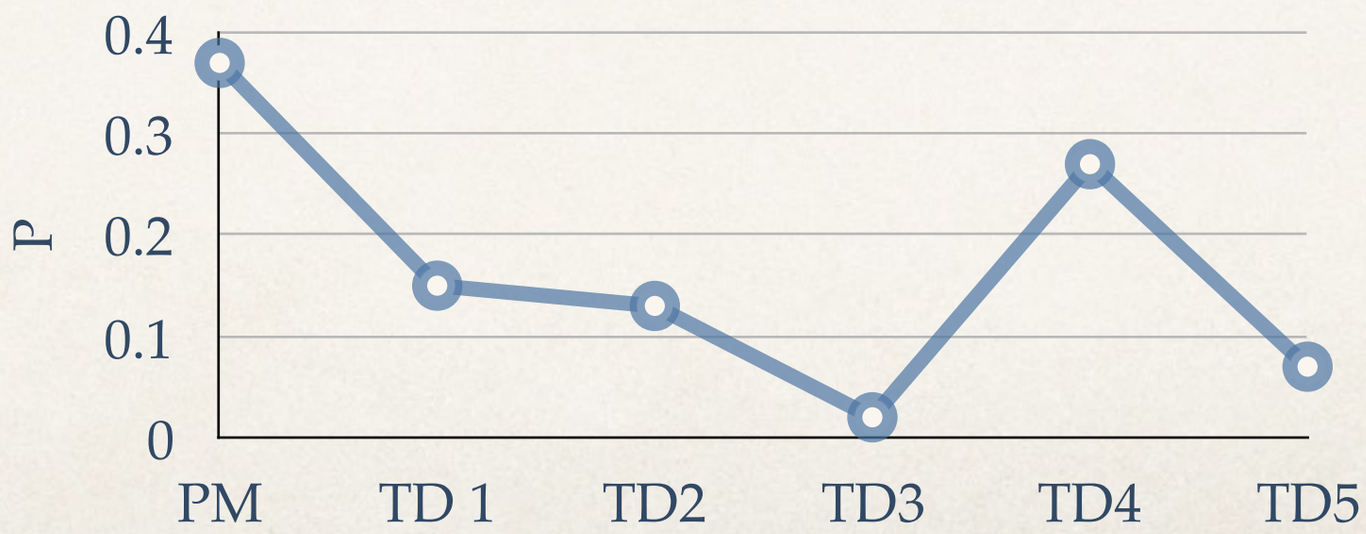
GCT**C**CGTTACAGGTGGGCAGGGGAGGCG GCT**G**CGTTACAGGTGGGCAGGGGAGGCG
GCTGCGTTACAGGTGGGCAGGGGAGGCG GCTGCGTTACAGGTGGGCAGGGGAGGCG
GCTGCGTTACAGGTGGGCAGGGGAGGCG GCTCCGTTACAGGTGGGCAGGGGAGGCG
GCTGCGTTACAGGTGGGCAGGGGAGGCG GCTGCGTTACAGGTGGGCAGGGGAGGCG
GCTGCGTTACAGGTGGGCAGGGGAGGCG GCTCCGTTACAGGTGGGCAGGGGAGGCG
GCTCCGTTACAGGTGGGCAGGGGAGGCG 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

$$s = \begin{array}{ll}
\texttt{GCTCCGTTACAGGTGGGCAGGGGAGGCG} & \texttt{GCTGCGTTACAGGTGGGCAGGGGAGGCG} \\
\texttt{GCTGCGTTACAGGTGGGCAGGGGAGGCG} & \texttt{GCTGCGTTACAGGTGGGCAGGGGAGGCG} \\
\texttt{GCTGCGTTACAGGTGGGCAGGGGAGGCG} & \texttt{GCTCCGTTACAGGTGGGCAGGGGAGGCG} \\
\texttt{GCTGCGTTACAGGTGGGCAGGGGAGGCG} & \texttt{GCTGCGTTACAGGTGGGCAGGGGAGGCG} \\
\texttt{GCTGCGTTACAGGTGGGCAGGGGAGGCG} & \texttt{GCTCCGTTACAGGTGGGCAGGGGAGGCG} \\
\texttt{GCTCCGTTACAGGTGGGCAGGGGAGGCG} & \texttt{GCTGCGTTACAGGTGGGCAGGGGAGGCG} \\
\texttt{GCTGCGTTACAGGTGGGCGGGGGAGGCG} & \texttt{GCTGCGTTACAGGTGGGCGGGGGAGGCG} \\
\texttt{GCTGCGTTACAGGTGGGCAGGGGGGGCG} & \texttt{GCTGCGTTACAGGTGGGCGGGGGAGGCG} \\
\texttt{GCTGCGTTACAGGTGGGCGGGGGAGGCG} & \texttt{GCTCCGTTACAGGTGGGCGGGGGAGGCG} \\
\texttt{GCTGCGTTACAGGTGGGCGGGGGAGGCG} & \texttt{GCTGCGTTACAGGTGGGCGGGGGGGGCG} \\
\texttt{GCTCCGTTACAGGTGGGCGGGGGAGGCT} & \texttt{GCTCCGTTACAGGTGGGCGGGGGAGGCT} \\
\texttt{GCTCCGTTACAGGTGGGCGGGGGGGGCG} & \texttt{GCTGCGTTACAGGTGGGCGGGGGGGGCG} \\
\texttt{GCTGCGTTACAGGTGGGCGGGGGAGGCG} & \texttt{GCTGCGTTACAGGTGGGCGGGGGAGGCG} \\
\texttt{GCTCCGTTACAGGTGGGCGGGGGAGGCG} & \texttt{GCTGCGTTACAGGTGGGCGGGGGAGGCG} \\
\texttt{GCTGCGTTACAGGTGGGCAGGGGAGGCG} & \texttt{GCTGCGTTACAGGTGGGCAGGGGAGGCG} \\
\texttt{GCTGCGTTACAGGTGGGCGGGGGAGGCG} & \texttt{GCTCCGTTACAGGTGGGCGGGGGAGGCG} \\
\texttt{GCTGCGTTACAGGTGGGCGGGGGAGGCG} & \texttt{GCTGCGTTACAGGTGGGCGGGGGAGGCG}
\end{array}$$



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 **q** = (*PM1, TD1, TD2, TD3*) randomly

TGAATGT

q = (*0.24, 0.33, 0.34, 0.09*)

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

TGAATGT

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

200 mutations: TGAATGTGCGT…

Simulation

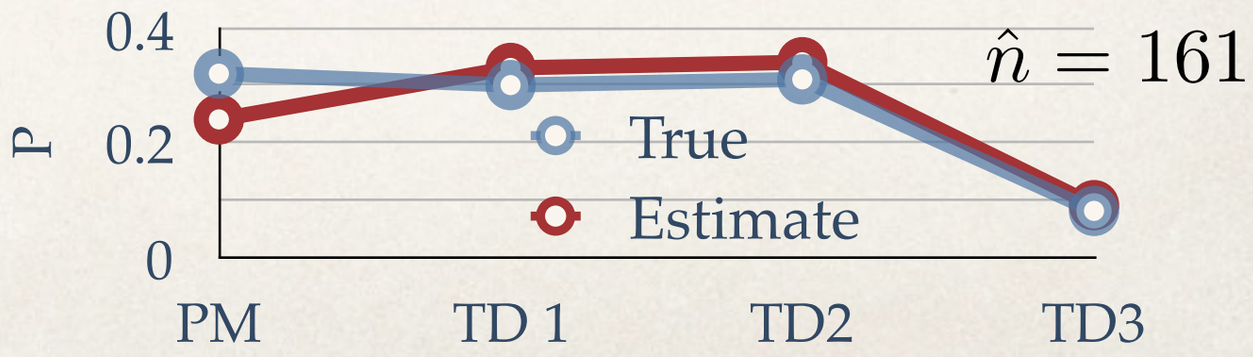
- Start with a short random seed over {A,C,G,T}
- Choose the parameters $\mathbf{q} = (PM1, TD1, TD2, TD3)$ randomly
- Apply n random mutations
- Estimate the parameters

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

TGAATGT

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

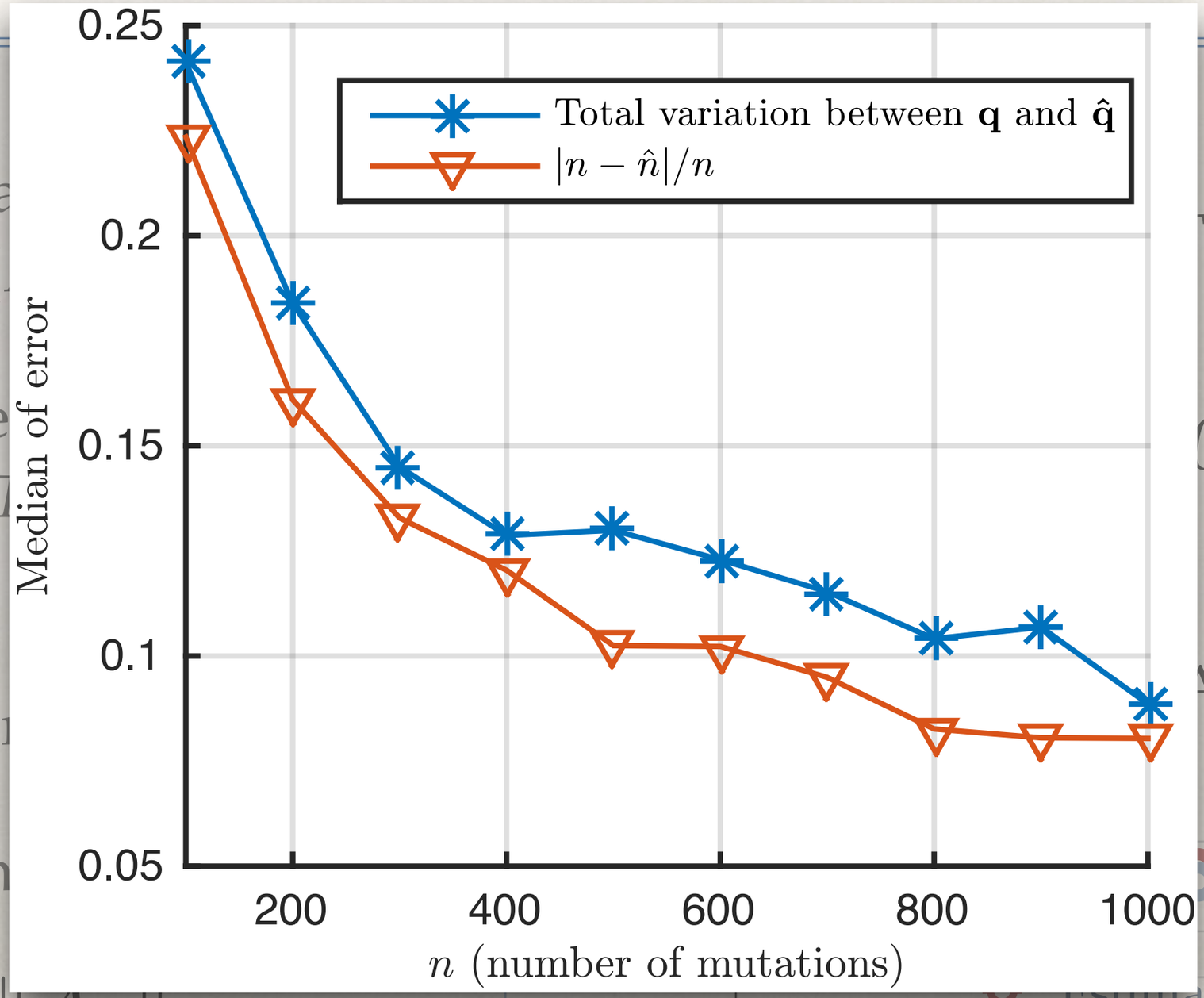
200 mutations: TGAATGTGCGT…



Simulation

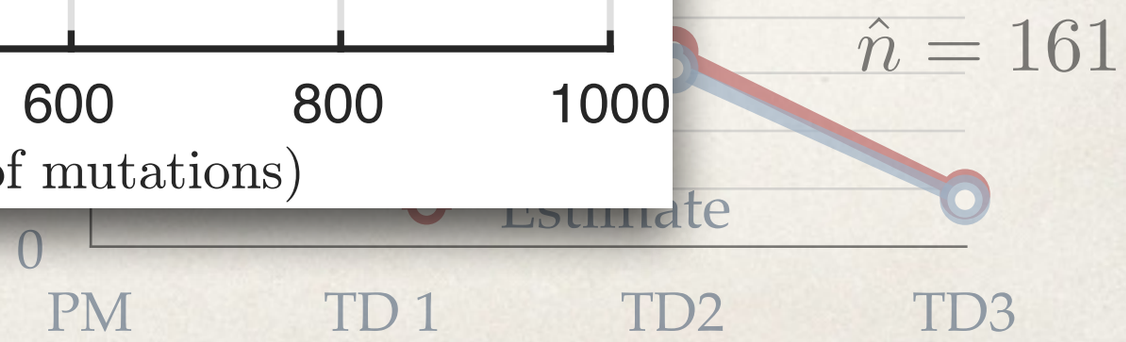
- Start with a
  seed over {
- Choose the
  $\mathbf{q} = (PM1, T$
  randomly
- Apply n ra
- Estimate th

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



0.34, 0.09)

ATGTGCGT…



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.

- 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.

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