Stochastic Models for DNA Tandem Duplication

Farzad Farnoud, with M. Schwartz, J. Bruck

Jan 18, 2016, University of Washington
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**

- **Point Mutation**
  - From TCGATGCA to TGATGCA

- **Deletion**
  - From TGATGCA to TGATGCA

- **Insertion**
  - From TGATGCA to TCATGCA
Types of Mutations

**TGATGCA**
- Point Mutation

**TCATGCA**
- Tandem Duplication
- Deletion
- Insertion
- 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
A tandem repeat region

Ch1: 933,911–935,015
A tandem repeat region

Ch1: 933,911–935,015

repeat unit length

repeat unit
A tandem repeat region

Ch1: 933,911–935,015

Point mutation (PM)
Point mutations are in the same positions
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)
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?
Finding Duplication History

ACGT ACTT GCGT ACTT GCGT ACTT
Finding Duplication History

ACGT  ACTT  GCGT  ACTT  GCGT  ACTT
Finding Duplication History

ACGT ACTT GCGT ACTT GCGT ACTT

ACGT ACTT GCGT ACTT

TD2
Finding Duplication History

ACGT ACTT GCGT ACTT GCGT ACTT

ACGT ACTT GCGT ACTT

PM

TD2
Finding Duplication History

ACGT  ACTT  GCGT  ACTT  GCGT  ACTT

TD2

ACGT  ACTT  GCGT  ACTT

PM

ACGT  ACTT  ACGT  ACTT

TD2

ACGT  ACTT

ACGT  ACTT
Finding Duplication History

ACGT  ACTT  GCGT  ACTT  GCGT  ACTT
  /  /  /  /  /  /
ACGT  ACTT  GCGT  ACTT
  \
ACGT  ACTT  ACGT  ACTT
  /  /  /  /
ACGT  ACTT  ACGT  ACTT
  \
ACGT  ACTT
  /
ACGT  ACTT

TD2
PM
TD2
PM
ACGT  ACGT
Finding Duplication History

ACGT  ACTT  GCGT  ACTT  GCGT  ACTT

ACGT  ACTT  GCGT  ACTT

ACGT  ACTT  ACGT  ACTT

ACGT  ACTT

ACGT  ACGT

ACGT
Finding Duplication History

ACGT  ACTT  GCGT  ACTT  GCGT  ACTT

ACGT  ACTT  GCGT  ACTT

ACGT  ACTT  ACGT  ACTT

ACGT  ACTT

ACGT  ACGT

ACGT

TD1, 2 TD2, 2 PM
Finding Duplication History

ACGT ACTT GCGT ACTT GCGT ACTT

ACGT ACTT GCGT ACTT

ACGT ACTT ACGT ACT

ACGT ACGT

ACGT

Maximum Parsimony
Thought to be NP-hard
[Gascuel et al., 2005]
Given the final sequence, can we efficiently estimate the parameters?
How to extract information from point mutations?

- Autocorrelation function:

  \[ r(\delta) = \text{fraction of symbols at distance } \delta \text{ units that are the same} \]

  \[ r(1) = \frac{11}{20} \]

  \[ r(2) = \frac{15}{16} \]
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:
  
  \( P(PM), P(TD1), P(TD2), \ldots \)

- As \( n \) increases, \( r_n \) tends to a point in the null space of \( A \)
Autocorrelation Limit

\[ r_n(\delta) \]

\[
P(\text{PM}) = 0.250 \\
P(\text{TD1}) = 0.525 \\
P(\text{TD2}) = 0.225
\]
Autocorrelation Limit

\[ r_n(\delta) \]

\[
P(\text{PM}) = 0.250
\]
\[
P(\text{TD1}) = 0.525
\]
\[
P(\text{TD2}) = 0.225
\]
Autocorrelation Limit

\[ r_n(\delta) \]

\[
P(\text{PM}) = 0.250 \\
P(\text{TD1}) = 0.525 \\
P(\text{TD2}) = 0.225
\]
Autocorrelation Limit

\[
P(\text{PM}) = 0.250 \\
P(\text{TD1}) = 0.525 \\
P(\text{TD2}) = 0.225
\]

\[
r_n(\delta)
\]
1. Calculate autocorrelation $r$ of $s$.

2. Find mutation probs such that the $l_2$-norm $\| Ar \|_2$ is minimized.
Estimation Algorithm

1. Calculate autocorrelation $r$ of $s$.

2. Find mutation probs such that the $l_2$-norm $\|Ar\|_2$ is minimized.
Estimation Algorithm

\[ S = \text{GCTCGTTACAGGTGGGCGGGGGAGGCG} \]

\[ \text{GCTCGTTACAGGTGGGCGGGGGAGGCG} \]

\[ \text{GCTCGTTACAGGTGGGCGGGGGAGGCG} \]

\[ \text{GCTCGTTACAGGTGGGCGGGGGAGGCG} \]

\[ \text{GCTCGTTACAGGTGGGCGGGGGAGGCG} \]

\[ \text{GCTCGTTACAGGTGGGCGGGGGAGGCG} \]

\[ \text{GCTCGTTACAGGTGGGCGGGGGAGGCG} \]

\[ \text{GCTCGTTACAGGTGGGCGGGGGAGGCG} \]

\[ \text{GCTCGTTACAGGTGGGCGGGGGAGGCG} \]

\[ \text{GCTCGTTACAGGTGGGCGGGGGAGGCG} \]

\[ \text{GCTCGTTACAGGTGGGCGGGGGAGGCG} \]

\[ \text{GCTCGTTACAGGTGGGCGGGGGAGGCG} \]

\[ \text{GCTCGTTACAGGTGGGCGGGGGAGGCG} \]

\[ \text{GCTCGTTACAGGTGGGCGGGGGAGGCG} \]

\[ \text{GCTCGTTACAGGTGGGCGGGGGAGGCG} \]

\[ \text{GCTCGTTACAGGTGGGCGGGGGAGGCG} \]

\[ \text{GCTCGTTACAGGTGGGCGGGGGAGGCG} \]

\[ \text{GCTCGTTACAGGTGGGCGGGGGAGGCG} \]
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

\[ TGAATGT \]
\[ q = (0.24, 0.33, 0.34, 0.09) \]

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

\[
\min_{\hat{q}} \| Ar \|_2
\]

TGAATGT

\[
q = (0.24, 0.33, 0.34, 0.09)
\]

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.

\[
\min_{\hat{q}} \|Ar - q\|_2
\]

\[
\hat{n} = 161
\]

\[
\text{Total variation between } q \text{ and } \hat{q}
\]

\[
|n - \hat{n}|/n
\]
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.

Data encoded in DNA → Noisy Channel → Noisy Output

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