Purity Dependant Markov Models for Microsatellite Mutation

Tristan L. Stark

University of Tasmania

tlstark@utas.edu.au

November 5, 2014
Overview

1. Microsatellites
2. Existing Models
3. Purity-dependant Model
4. Applications
Microsatellites

- Repeats of a short motif, e.g. AT repeated 6 times:

```
A T A T A T A T A T A T A
```
Microsatellites

- Repeats of a short motif, e.g. AT repeated 6 times:

\[
\begin{array}{cccccccc}
A & T & A & T & A & T & A & T \\
\end{array}
\]

- Think of microsatellites as repeat units:

\[
\begin{array}{cccc}
AT & AT & AT & AT \\
\end{array}
\]
Microsatellites

- Repeats of a short motif, e.g. AT repeated 6 times:

  \[
  \text{ATATATATATATATAT}
  \]

- Think of microsatellites as repeat units:

  \[
  \text{AT AT AT AT AT AT}
  \]

- Highly polymorphic.
- Abundant in eukaryote genomes.
- Often selectively neutral.
Slipped-strand mispairing

Contraction
During replication, a loop may form in the template strand leading to a decrease in the number of repeats in the new strand.

\[
\begin{array}{c}
A \quad T \\
T \\
A \quad T \\
T \quad A \\
A \quad T \quad A \quad T \quad A \\
T \quad A \quad T \quad A \quad T \\
\end{array}
\]

Loop formed in Template Strand
Template Strand
New Strand
Slipped-strand mispairing

**Expansion**

Alternatively, a loop may form in the new strand, leading to an increase in repeat number relative to the template.

```
Template Strand
ATATATAATAATAATA

New Strand
TATATAATAATAATA

Loop formed in New Strand
```

Tristan L. Stark (UTAS)
e.g. a symmetric random walk:

The main factors accounted for are:
- Length dependence of mutation rate.
- Bias towards contraction or expansion.
- Size of the mutation events.
[Wu and Drummond, 2011] proposed a class of models which captures many of the models in the literature as subclasses. This model allows for:

1. Quadratic functions of repeat number for mutation rate.
2. Length dependent mutational bias.
3. Geometrically distributed slippage event sizes.
[Wu and Drummond, 2011] proposed a class of models which captures many of the models in the literature as subclasses.

This model allows for:

1. Quadratic functions of repeat number for mutation rate.
2. Length dependent mutational bias.
3. Geometrically distributed slippage event sizes.

For the one-phase models (slippage events of size 1 only) model is given by

\[
q_{ij} = \begin{cases} 
\alpha(u_0, u_1, u_2, i)\beta(b_0, b_1, i) & \text{if } i - j = -1 \\
\alpha(u_0, u_1, u_2, i)(1 - \beta(b_0, b_1, i)) & \text{if } i - j = 1 \\
-\sum_{k \neq i} q_{ik} & \text{if } i = j.
\end{cases}
\]
Microsatellites also susceptible to point mutations.

\[
\begin{array}{ccccccc}
\text{AT} & \text{AT} & \text{AT} & \text{AC} & \text{AT} & \text{AT} \\
\end{array}
\]

How to deal with this?

\[
\begin{array}{ccc}
\text{AT} & \text{AT} & \text{AT} \\
\text{AT} & \text{AT} \\
\end{array}
\]
These models lose useful information, and may invalidate IID assumption.

Loop forming around impure repeat.

Template Strand

New Strand
Kruglyak’s proportional slippage model

- [Kruglyak, 1998] proposed a model which included point mutation.
- They assumed slippage was linearly proportional to repeat number,
- and that point mutation would occur in any repeat at a constant rate $a$.

$$q_{ij} = \begin{cases} 
  c & \text{for } i = 1, j = 2 \\
  (i - 1)b & \text{for } i > 1, j = i + 1 \\
  (i - 1)b + a & \text{for } i > 1, j = i - 1 \\
  a & \text{for } i > 1, j < i - 1 \\
  0 & \text{otherwise.}
\end{cases}$$
Kruglyak’s proportional slippage model

- Kruglyak and Durrett proved in a later paper [Durrett, 1999] that the stationary distribution exists.
- Stationary distribution can be shown to satisfy

\[c \pi_1 = b \pi_2 + a \sum_{j=2}^{\infty} \pi(j),\]

\[b(i - 1)\pi_i = b i \pi_{i+1} + ia \sum_{i=i+1}^{\infty} \pi_j \text{ for } i \geq 2.\]
We move up a dimension in the state space.

\[(i, j)\]

\# repeats \# interruptions

\[
\begin{array}{cccc}
  AT & AT & AT & AC & AT & AT \\
\end{array}
\]

\[= (6, 1)\]
Key Assumptions

- Effect of impurity is independent of location.

\[
\begin{array}{cccc}
AT & AT & AT & AC & AT \\
\end{array}
= \begin{array}{cccc}
AT & AC & AT & AT & AT \\
\end{array}
\]

- Each base pair is either ‘correct’ or ‘incorrect’.

\[
\begin{array}{cc}
AT \\n\neq \begin{array}{cc}
AC \\
\end{array} \\
= \begin{array}{cc}
AG \\
\end{array} \\
= \begin{array}{cc}
AA \\
\end{array}
\end{array}
\]
A repeat unit is either pure or impure - binary.

\[ AT \neq AX = YT = YX \]

Slippage events of length 1 only.
Slipped-strand mispairing

- Process may transition from a state \((i, j)\) to \((i + 1, j)\) at a rate given by \(r_s(i, j)\).

Point mutation

- Process may transition from a state \((i, j)\) to \((i, j + 1)\) at a rate given by \(r_m(i, j)\).
Slipped-strand mispairing

- Process may transition from a state \((i, j)\) to \((i + 1, j)\) at a rate given by \(r_s(i, j)\).
- Process may transition from a state \((i, j)\) to \((i - 1, j)\) at a rate given by \(r_s(i, j) \frac{(i-j)}{i}\).
Slipped-strand mispairing

- Process may transition from a state \((i, j)\) to \((i + 1, j)\) at a rate given by \(r_s(i, j)\).
- Process may transition from a state \((i, j)\) to \((i - 1, j)\) at a rate given by \(r_s(i, j)\frac{(i-j)}{i}\).
- Process may transition from a state \((i, j)\) to \((i - 1, j - 1)\) at a rate given by \(r_s(i, j)\frac{j}{i}\).
Slipped-strand mispairing

- Process may transition from a state \((i, j)\) to \((i + 1, j)\) at a rate given by \(r_s(i, j)\).
- Process may transition from a state \((i, j)\) to \((i - 1, j)\) at a rate given by \(r_s(i, j)\frac{(i-j)}{i}\).
- Process may transition from a state \((i, j)\) to \((i - 1, j - 1)\) at a rate given by \(r_s(i, j)\frac{j}{i}\).

Point mutation

- Process may transition from a state \((i, j)\) to \((i, j + 1)\) at a rate given by \(r_m(i, j)\).
The General Purity-Dependant Model

In its most general form, our model is given by generator $\mathbf{Q} = [q_{ij}]$ where

$$q(i,j)(k,l) = \begin{cases} 
    r_s(i,j) \beta(i) & \text{for } k = i + 1, \ l = j \\
    r_s(i,j)(1 - \beta(i)) \frac{(i-j)}{i} & \text{for } k = i - 1, \ l = j \\
    r_s(i,j)(1 - \beta(i)) \frac{i}{i} & \text{for } k = i - 1, \ l = j - 1 \\
    r_m(i,j) & \text{for } k = i, \ l = j + 1.
\end{cases}$$
The General Purity-Dependant Model

\[
\begin{align*}
\text{rs}(1,0) & \quad \beta(1) \\
\text{rs}(2,0) & \quad \beta(2) \\
\text{rs}(3,0) & \quad \beta(3) \\
\text{rs}(4,0) & \quad \beta(4) \\
\text{rm}(2,0) & \quad \beta(2) \\
\text{rm}(3,0) & \quad \beta(3) \\
\text{rm}(4,0) & \quad \beta(4) \\
\text{rs}(2,1) & \quad \beta(2) \\
\text{rs}(3,1) & \quad \beta(3) \\
\text{rs}(4,1) & \quad \beta(4) \\
\end{align*}
\]
Some Restrictions

By making some restrictions we can judge the benefits of modeling point mutation/purity.

**Purity-independant model**

Set $r_s(i,j) \equiv r_s(i)$.

- Models point mutation.
- Purity has no effect on mutation rates.
Some Restrictions

By making some restrictions we can judge the benefits of modeling point mutation/purity.

### Purity-independant model

Set $r_s(i, j) \equiv r_s(i)$.
- Models point mutation.
- Purity has no effect on mutation rates.

### One-dimensional model

Set $r_m(i, j) \equiv 0$ (and fix $j = 0$)
- No point mutation.
- No purity dependance
- Reduced to 1D, one-phase model.
Applications

We choose some specific functions $r_s, \beta, r_m$

- $r_s(i, j) = (u_0 + u_1(i - 1))c^{-j}$,
- $\beta(i) = \frac{1}{1 + e^{-(b_0 + (i - 1)b_1)}}$,
- $r_m(i, j) = d(i - j)$. 
We choose some specific functions $r_s, \beta, r_m$

- $r_s(i, j) = (u_0 + u_1(i - 1))c^{-j}$,
- $\beta(i) = \frac{1}{1 + e^{-(b_0 + (i - 1)b_1)}}$,
- $r_m(i, j) = d(i - j)$.

If we set $c = 1$ then $r_s(i, j) = r_s(i)$. 
We choose some specific functions $r_s, \beta, r_m$

- $r_s(i,j) = (u_0 + u_1(i - 1))c^{-j}$,
- $\beta(i) = \frac{1}{1 + e^{-(b_0 + (i-1)b_1)}}$,
- $r_m(i,j) = d(i - j)$.

If we set $c = 1$ then $r_s(i,j) = r_s(i)$.

If we set $r_m = 0$ then we have Wu and Drummond’s one-phase linear-rate logistic bias model.
Simulation (Purity-dependant Model)

Parameters and their true values

- $u_0 (0.85)$
- $u_1 (0.085)$
- $b_0 (2)$
- $b_1 (-1)$
- $c (0.25)$
- $d (0.06)$

Difference from true value

Tristan L. Stark (UTAS)
Simulation (Purity-independent model)

<table>
<thead>
<tr>
<th>Parameters</th>
<th>True Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>$u_0$</td>
<td>0.85</td>
</tr>
<tr>
<td>$u_1$</td>
<td>0.085</td>
</tr>
<tr>
<td>$b_0$</td>
<td>2</td>
</tr>
<tr>
<td>$b_1$</td>
<td>-1</td>
</tr>
<tr>
<td>$d$</td>
<td>0.06</td>
</tr>
</tbody>
</table>

Parameters and their true values

Tristan L. Stark (UTAS)
Microsatellite Models
November 5, 2014 21 / 24
Acknowledgements

Supervisors

- Dr Małgorzata O’Reilly
- Dr Barbara Holland
- Dr Bennet McComish
Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations
_Molecular Biology and Evolution_

A new stochastic model of microsatellite evolution
_Applied Probability Trust_

Wu, C. and Drummond, A. (2011)
Joint inference of microsatellite mutation models, population history and genealogies using transdimensional Markov Chain Monte Carlo
_Genetics Soc America_

Walsh, J. (1987)
Persistence of tandem arrays: implications for satellite and simple-sequence DNAs
_Genetics Soc America_
References II

A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population
Genetical research

Microsatellite mutation models insights from a comparison of humans and chimpanzees
Genetics Soc America