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:

$$\boxed{A}\boxed{T}\boxed{A}\boxed{T}\boxed{A}\boxed{T}\boxed{A}\boxed{T}\boxed{A}\boxed{T}\boxed{A}\boxed{T}$$

Microsatellites

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

$$\boxed{A}\,\boxed{T}\,\boxed{A}\,\boxed{T}\,\boxed{A}\,\boxed{T}\,\boxed{A}\,\boxed{T}\,\boxed{A}\,\boxed{T}\,\boxed{A}\,\boxed{T}$$

- Think of microsatellites as repeat units:

$$\boxed{AT}\;\boxed{AT}\;\boxed{AT}\;\boxed{AT}\;\boxed{AT}\;\boxed{AT}$$

Microsatellites

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

$$\boxed{A}\,\boxed{T}\,\boxed{A}\,\boxed{T}\,\boxed{A}\,\boxed{T}\,\boxed{A}\,\boxed{T}\,\boxed{A}\,\boxed{T}\,\boxed{A}\,\boxed{T}$$

- Think of microsatellites as repeat units:

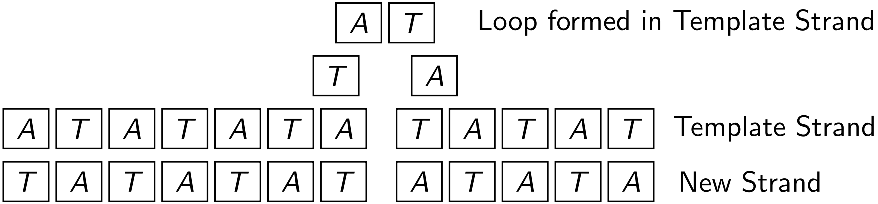
$$\boxed{AT}\quad\boxed{AT}\quad\boxed{AT}\quad\boxed{AT}\quad\boxed{AT}\quad\boxed{AT}$$

- 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.



A T Loop formed in Template Strand

T A

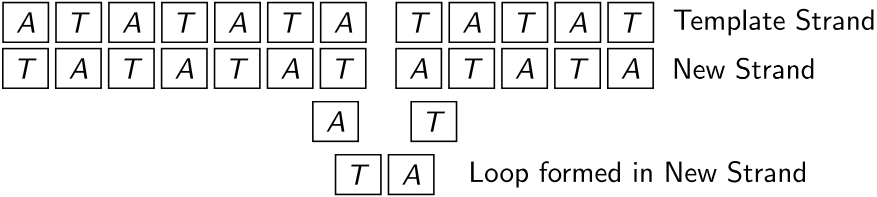
A T A T A T A T A T A T Template Strand

T A T A T A T A T A T A 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.



A T A T A T A T A T A T Template Strand

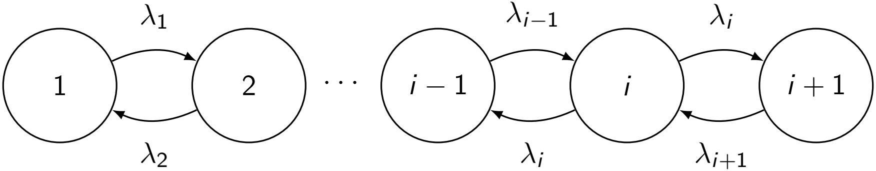
T A T A T A T A T A T A New Strand

A T

T A Loop formed in New Strand

Models for repeat number

- 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.

General one-phase slippage model

- [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.

General one-phase slippage model

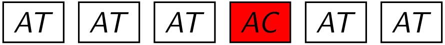
- [Wu and Drummond, 2011] proposed a class of models which captures many of the models in the literature as subclasses.
- This model allows for:
  1. Quadratic functions of repeat number for mutation rate.
  2. Length dependent mutational bias.
  3. Geometrically distributed slippage event sizes.

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

$$q_{ij} = \begin{cases} \alpha(u_0, u_1, u_2, i)\beta(b_0, b_1, i) & \text{if } i - j = -1 \\ \alpha(u_0, u_1, u_2, i)(1 - \beta(b_0, b_1, i)) & \text{if } i - j = 1 \\ -\sum_{k \neq i} q_{ik} & \text{if } i = j. \end{cases}$$

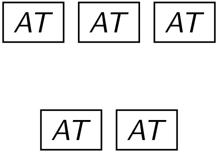
Point mutation

- Microsatellites also susceptible to point mutations.



AT AT AT AC AT AT

- How to deal with this?

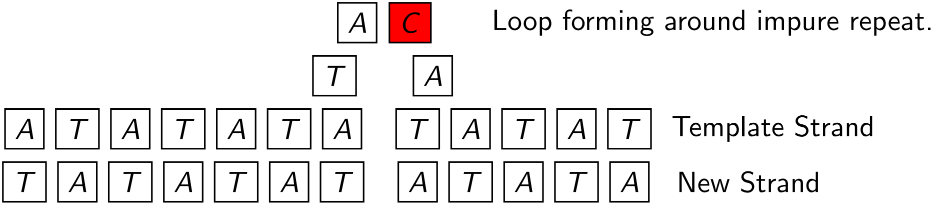


AT AT AT

AT AT

Point mutation

- These models lose useful information, and may invalidate IID assumption.



A C Loop forming around impure repeat.

T A

A T A T A T A T A T A T Template Strand

T A T A T A T A T A T A 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

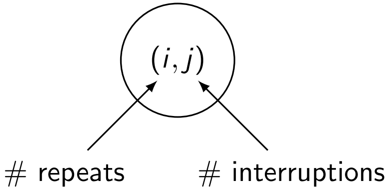
- Kruglyak and Durrett proved in a later paper [Durret, 1999] that the stationary distribution exists.
- Stationary distribution can be shown to satisfy

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

$$b(i-1)\pi_i = bi\pi_{i+1} + ia\sum_{i=i+1}^{\infty}\pi_j \text{ for } i \geq 2.$$

Purity-dependant Model

- We move up a dimension in the state space.



(i,j)

# repeats

interruptions

$$\boxed{AT}\ \boxed{AT}\ \boxed{AT}\ \boxed{AC}\ \boxed{AT}\ \boxed{AT}$$

$$= (6,1)$$

Key Assumptions

- Effect of impurity is independent of location.

$$\boxed{AT}\ \boxed{AT}\ \boxed{AT}\ \fcolorbox{black}{red}{AC}\ \boxed{AT} \quad = \quad \boxed{AT}\ \fcolorbox{black}{red}{AC}\ \boxed{AT}\ \boxed{AT}\ \boxed{AT}$$

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

$$\boxed{A}\boxed{T} \neq \boxed{A}\boxed{C} = \boxed{A}\boxed{G} = \boxed{A}\boxed{A}$$

Extra Assumptions

- A repeat *unit* is either pure or impure - binary.

$$\boxed{AT} \neq \boxed{AX} = \boxed{YT} = \boxed{YX}$$

- Slippage events of length 1 only.

Heuristic model development

Slipped-strand mispairing

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

Heuristic model development

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 $r_s(i, j)\frac{(i-j)}{i}$.

Heuristic model development

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 $r_s(i, j)\frac{j}{i}$.

Heuristic model development

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

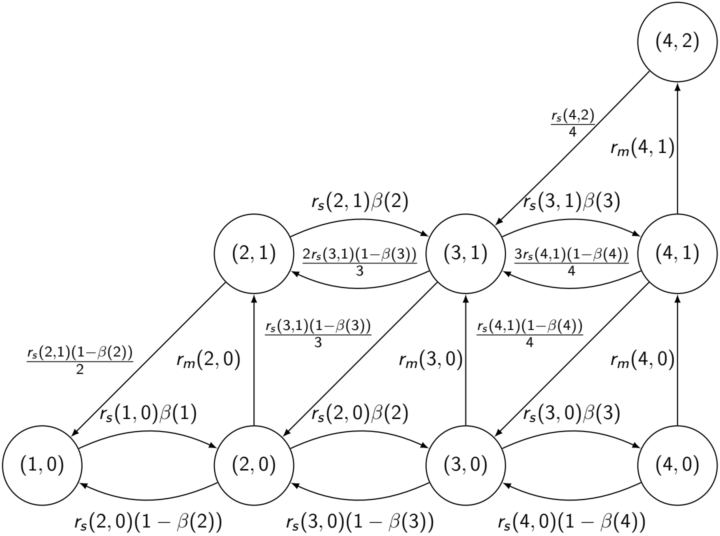
- Process may transition from a state (i, j) to $(i, j + 1)$ at a rate given by $r_m(i, j)$.

The General Purity-Dependant Model

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

$$q_{(i,j)(k,l)} = \begin{cases} r_s(i,j)\beta(i) & \text{for } k = i+1, l = j \\ r_s(i,j)(1-\beta(i))\frac{(i-j)}{i} & \text{for } k = i-1, l = j \\ r_s(i,j)(1-\beta(i))\frac{j}{i} & \text{for } k = i-1, l = j-1 \\ r_m(i,j) & \text{for } k = i, l = j+1. \end{cases}$$

The General Purity-Dependant Model



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, β, r_m

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

Applications

We choose some specific functions r_s, β, 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 $r_s(i,j) = r_s(i)$.

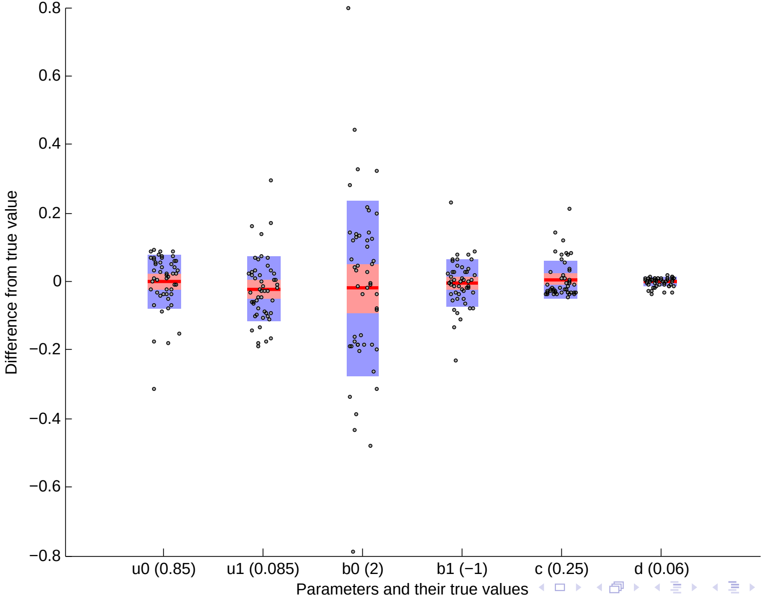
Applications

We choose some specific functions r_s, β, r_m

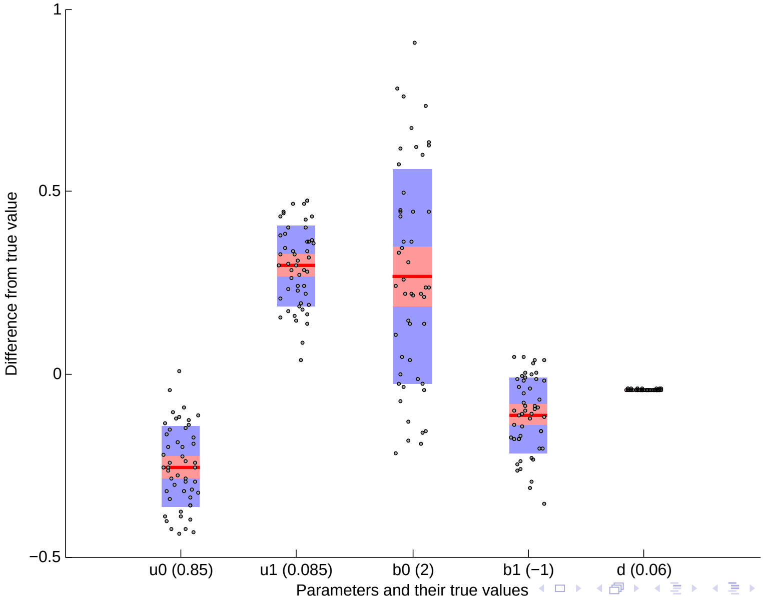
- $r_s(i,j) = (u_0 + u_1(i-1))c^{-j}$,
- $\beta(i) = \frac{1}{1+e^{-(b_0+(i-1)b_1}}$,
- $r_m(i,j) = d(i-j)$.

- 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)



Simulation (Purity-independant model)



Acknowledgements

Supervisors

- Dr Małgorzata O'Reilly
- Dr Barbara Holland

- Dr Bennet McComish

References I

Kruglyak, S. and Durrett, R. and Schug, M. and Aquadro, C. (1998)
Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations
Molecular Biology and Evolution

Durrett, R. T and Kruglyak, S. (1999)
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

Ohta, T. and Kimura, M. (1973)
A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population
Genetical research

Sainudiin, R. and Durrett, R. and Aquadro, C. and Nielsen, R. (2004)
Microsatellite mutation models insights from a comparison of humans and chimpanzees
Genetics Soc America