Models for the evolution of microsatellites

Tristan L. Stark

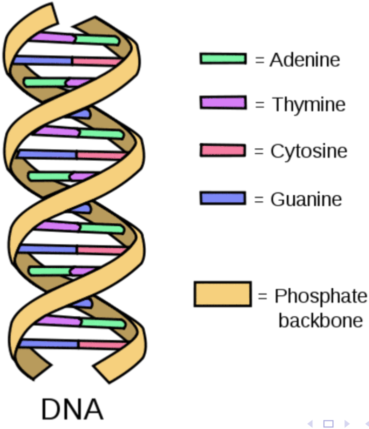
Temple University

tuk60782@temple.edu

February 14, 2019

DNA Basics

- DNA is a molecule found in all life.
- Typically appears as a double helix structure.
- Made up of nucleotides



Neucleotides

- We consider 4 distinct nucleotides:
  1. A - Adenine.
  2. T - Thymine.
  3. C - Cytosine.
  4. G - Guanine.

- The genetic code is written in the language defined by these nucleobases.
- A piece of code may be regarded as a string of nucleobases
- e.g. ATCCATATG

Base Pairing

- Nucleobases on one strand chemically bond with those on the other
- Each bonds with precisely one other:
  1. A and T bond.
  2. C and G bond.

- Base pairing leads to a natural complementary relationship between strings of code.
- E.g. ATCCATATG has TAGGTATAC as its complement

Mutation

- Two strands of double helix separate
- Each strand's complimentary sequence is generated and bonds to it.
- There exists a possibility for errors to occur.
- Most errors are corrected, some lead to a change in the code.
- We refer to uncorrected errors as mutations.

Selection

- Mutations may give rise to new alleles.
- Often, this will make an organism more or less fit.
- Many microsatellites are *not* subject to selection, which allows for demographic information to be faithfully preserved.

Microsatellites

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

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

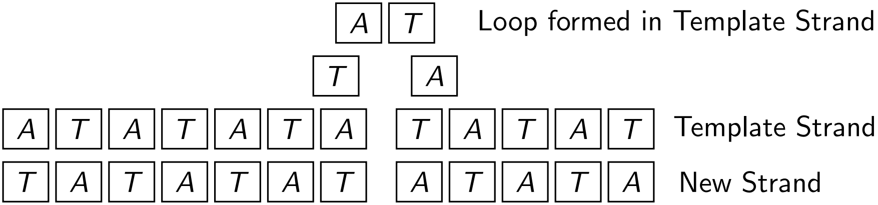
- Usually, think of microsatellites as repeat units:

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

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

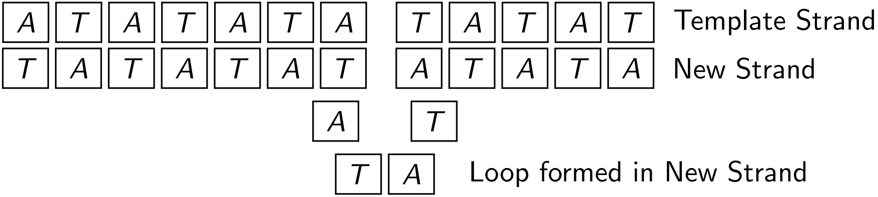
A T A T A T A T A T A T Template Strand

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

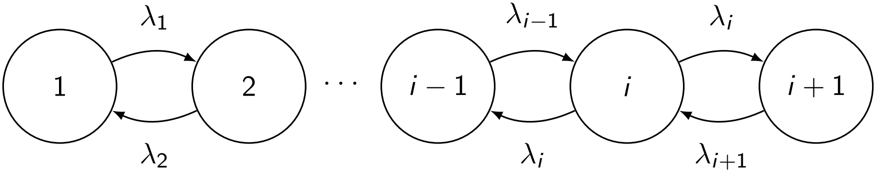
T A T A T A T   A T A T A   New Strand

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.

Point mutation

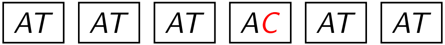
- Microsatellites also susceptible to point mutations.

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

- How to deal with this?

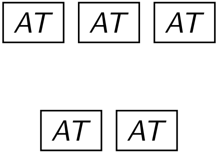
Point mutation

- Microsatellites also susceptible to point mutations.



AT AT AT AC AT AT

- How to deal with this?
- One way is to model point mutation as splitting a single microsatellite into two smaller ones.

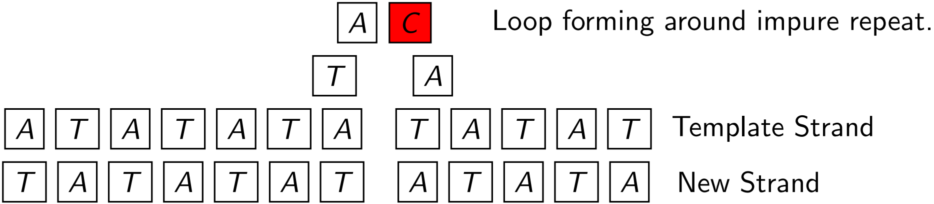


AT AT AT

AT AT

Some problems

- 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

Our model

We introduce a level-dependent QBD to model the evolution of an individual microsatellite.

Our model

We introduce a level-dependent QBD to model the evolution of an individual microsatellite.

- State space — $\mathcal{S} = \{(i,j) \in \mathbb{N}^2 : i_{\text{min}} \leq i \leq i_{\text{max}}, j \leq j^i_{\text{max}}\}$

Our model

We introduce a level-dependent QBD to model the evolution of an individual microsatellite.

- State space — $\mathcal{S} = \{(i,j) \in \mathbb{N}^2 : i_{\text{min}} \leq i \leq i_{\text{max}}, j \leq j^{i}_{\text{max}}\}$
- i tracks the repeat number, j tracks the number of mismatches at the level of the nucleotide.

Our model

We introduce a level-dependent QBD to model the evolution of an individual microsatellite.

- State space — $\mathcal{S} = \{(i,j) \in \mathbb{N}^2 : i_{\text{min}} \leq i \leq i_{\text{max}}, j \leq j^i_{\text{max}}\}$
- i tracks the repeat number, j tracks the number of mismatches at the level of the nucleotide.
- $i_{\text{min}}, i_{\text{max}}$ and j^i_{max} are all absorbing states, although taking $i_{\text{max}} = \infty$ is natural.

Our model

We introduce a level-dependent QBD to model the evolution of an individual microsatellite.

- State space — $\mathcal{S} = \{(i,j) \in \mathbb{N}^2 : i_{\text{min}} \leq i \leq i_{\text{max}}, j \leq j^i_{\text{max}}\}$
- i tracks the repeat number, j tracks the number of mismatches at the level of the nucleotide.
- $i_{\text{min}}, i_{\text{max}}$ and j^i_{max} are all absorbing states, although taking $i_{\text{max}} = \infty$ is natural.
- Generator $\mathbf{Q} = [q_{(i,j).(k,l)}]$ with

$$q_{(i,j)(k,l)} = \begin{cases} r_s(i,j)\beta(i) & \text{for } k = i+1, l = j \\ r_s(i,j)(1-\beta(i))H(j-l, iL, j, L) & \text{for } k = i-1, j-L \leq l \\ r_m(i,j) & \text{for } k = i, l = j+1 \\ r_p(i,j) & \text{for } k = i, l = j-1. \end{cases} \tag{1}$$

Our model

We assume the following forms for the functions in equation (1),

$$r_s(i,j) = (u_0 + u_1(i-1))c^j, \tag{2}$$

Our model

We assume the following forms for the functions in equation (1),

$$r_s(i,j) = (u_0 + u_1(i-1))c^j, \tag{2}$$

$$\beta(i) = \frac{1}{1+e^{-(b_0+(i-1)b_1)}}, \tag{3}$$

Our model

We assume the following forms for the functions in equation (1),

$$r_s(i,j) = (u_0 + u_1(i-1))c^j, \tag{2}$$

$$\beta(i) = \frac{1}{1 + e^{-(b_0 + (i-1)b_1)}}, \tag{3}$$

$$r_m(i,j) = d(iL - j) \tag{4}$$

Our model

We assume the following forms for the functions in equation (1),

$$r_s(i,j) = (u_0 + u_1(i-1))c^j, \tag{2}$$

$$\beta(i) = \frac{1}{1 + e^{-(b_0 + (i-1)b_1)}}, \tag{3}$$

$$r_m(i,j) = d(iL - j) \tag{4}$$

$$r_p(i,j) = \frac{1}{3}dj \tag{5}$$

A short aside on whole genome derived sequence data...

Ultimately, we want to fit our model to some data from real microsatellite sequences. In order to obtain data with information about the number of interruptions in the repeat sequence, we use whole-genome data.

A short aside on whole genome derived sequence data...

Ultimately, we want to fit our model to some data from real microsatellite sequences. In order to obtain data with information about the number of interruptions in the repeat sequence, we use whole-genome data. To filter microsatellites from whole genome sequences, we use a program called Tandem Repeats Finder (TRF). TRF is essentially a two component algorithm for finding microsatellite sequences.

A short aside on whole genome derived sequence data...

Ultimately, we want to fit our model to some data from real microsatellite sequences. In order to obtain data with information about the number of interruptions in the repeat sequence, we use whole-genome data. To filter microsatellites from whole genome sequences, we use a program called Tandem Repeats Finder (TRF). TRF is essentially a two component algorithm for finding microsatellite sequences.

- Detection component does some statistics to find (a sample of) candidate microsatellites in the genome.

A short aside on whole genome derived sequence data...

Ultimately, we want to fit our model to some data from real microsatellite sequences. In order to obtain data with information about the number of interruptions in the repeat sequence, we use whole-genome data. To filter microsatellites from whole genome sequences, we use a program called Tandem Repeats Finder (TRF). TRF is essentially a two component algorithm for finding microsatellite sequences.

- Detection component does some statistics to find (a sample of) candidate microsatellites in the genome.
- Analysis component determines (among other things) the repeat motif and measures how well the observed sequence matches a theoretical sequence of the same length consisting of perfect copies of the repeat sequence.

Observable sequences

It is easy to derive a criteria in terms of repeat number and number of interruptions that a sequence can have to make it through the analysis component of TRF, based on chosen parameters.

Observable sequences

It is easy to derive a criteria in terms of repeat number and number of interruptions that a sequence can have to make it through the analysis component of TRF, based on chosen parameters.

- We set the ‘absorbing boundary‘ of the model to match the boundary of observability under TRF, so that i_{min} and j^{i}_{max} are determined by the aforementioned criteria.

Observable sequences

It is easy to derive a criteria in terms of repeat number and number of interruptions that a sequence can have to make it through the analysis component of TRF, based on chosen parameters.

- We set the 'absorbing boundary' of the model to match the boundary of observability under TRF, so that i_{min} and j^{i}_{max} are determined by the aforementioned criteria.
- We chose i_{max} to be the maximum observed sequence length in each subset of our dataset (which we partitioned by motif-length).

Fitting the model to whole-genome derived sequence data

Usually in the microsatellite literature, models are fit by assuming that observed data is at equilibrium and fitting the stationary distribution to the empirical distribution. Clearly not appropriate here...

Fitting the model to whole-genome derived sequence data

Usually in the microsatellite literature, models are fit by assuming that observed data is at equilibrium and fitting the stationary distribution to the empirical distribution. Clearly not appropriate here...

In fact, not having to assume that observed data is at equilibrium is a nice feature of this model (the assumption of equilibrium has some philosophical issues in the context of evolution).

Fitting the model to whole-genome derived sequence data

Usually in the microsatellite literature, models are fit by assuming that observed data is at equilibrium and fitting the stationary distribution to the empirical distribution. Clearly not appropriate here...

In fact, not having to assume that observed data is at equilibrium is a nice feature of this model (the assumption of equilibrium has some philosophical issues in the context of evolution).

The equilibrium assumption provides a natural way to extend a model for the evolution of an individual microsatellite to a model for a population of microsatellites.

Fitting the model to whole-genome derived sequence data

Usually in the microsatellite literature, models are fit by assuming that observed data is at equilibrium and fitting the stationary distribution to the empirical distribution. Clearly not appropriate here...

In fact, not having to assume that observed data is at equilibrium is a nice feature of this model (the assumption of equilibrium has some philosophical issues in the context of evolution).

The equilibrium assumption provides a natural way to extend a model for the evolution of an individual microsatellite to a model for a population of microsatellites.

We extend the model to the population-level by assuming a Poisson birth process for microsatellites, born with some initial distribution $\underline{\alpha}$.

Fitting the model to whole-genome derived sequence data

We derive the distribution (in terms of the individual-level model) of a microsatellite observed at a time t.

- Ignoring any imperfection in the process of observation, the event that a microsatellite is observed at time t^* is equivalent to the event that it was born before time t^*, and is absorbed after time t^*.
- It follows that the density associated with the event that a microsatellite is of age t given that it is observed at time t^* is

Fitting the model to whole-genome derived sequence data

We derive the distribution (in terms of the individual-level model) of a microsatellite observed at a time t.

- Ignoring any imperfection in the process of observation, the event that a microsatellite is observed at time t^* is equivalent to the event that it was born before time t^*, and is absorbed after time t^*.
- It follows that the density associated with the event that a microsatellite is of age t given that it is observed at time t^* is

$$\begin{aligned} f_{T_0}(t^* - t \mid T_0 < t^* < T_a) &= \frac{S(t)}{\int_{t=0}^{T} S(t)dt} \\ &= \frac{\underline{\alpha}_0 e^{\mathbf{Q}^* t} \underline{1}}{\underline{\alpha}_0 (e^{\mathbf{Q}^* t} - \mathbf{I})(\mathbf{Q}^*)^{-1} \underline{1}}. \end{aligned} \tag{6}$$

where S is the survival function and $\mathbf{Q}^*$ is the subgenerator associated with the model.

Fitting the model to whole-genome derived sequence data

Now we can write the probability that a microsatellite observed at time t is in state s as

$$\begin{aligned}
&P(X(t^* - T_0) = s \mid T_0 < t^* < T_a) \\
&= \int_{t=0}^{t^*} P(X(t) = s \mid T_0 = t^* - t < t^* < T_a) f_{T_0}(t^* - t \mid T_0 < t^* < T_a) dt \\
&= \int_{t=0}^{t^*} \left(\frac{[\underline{\alpha}_0 e^{\mathbf{Q}^* t}]_s}{\underline{\alpha}_0 e^{\mathbf{Q}^* t} \underline{1}} \right) \left(\frac{\underline{\alpha}_0 e^{\mathbf{Q}^* t} \underline{1}}{\underline{\alpha}_0 (e^{\mathbf{Q}^* t^*} - \mathbf{I})(\mathbf{Q}^*)^{-1} \underline{1}} \right) dt \\
&= \frac{[\underline{\alpha}_0 (e^{\mathbf{Q}^* t^*} - \mathbf{I})(\mathbf{Q}^*)^{-1}]_s}{\underline{\alpha}_0 (e^{\mathbf{Q}^* t^*} - \mathbf{I})(\mathbf{Q}^*)^{-1} \underline{1}}. \qquad (7)
\end{aligned}$$

Fitting the model to whole-genome derived sequence data

Now we can write the probability that a microsatellite observed at time t is in state s as

$$
\begin{aligned}
&P(X(t^* - T_0) = s \mid T_0 < t^* < T_a) \\
&= \int_{t=0}^{t^*} P(X(t) = s \mid T_0 = t^* - t < t^* < T_a) f_{T_0}(t^* - t \mid T_0 < t^* < T_a) dt \\
&= \int_{t=0}^{t^*} \left(\frac{[\underline{\alpha}_0 e^{\mathbf{Q}^* t}]_s}{\underline{\alpha}_0 e^{\mathbf{Q}^* t} \underline{1}} \right) \left(\frac{\underline{\alpha}_0 e^{\mathbf{Q}^* t} \underline{1}}{\underline{\alpha}_0 (e^{\mathbf{Q}^* t^*} - \mathbf{I})(\mathbf{Q}^*)^{-1} \underline{1}} \right) dt \\
&= \frac{[\underline{\alpha}_0 (e^{\mathbf{Q}^* t^*} - \mathbf{I})(\mathbf{Q}^*)^{-1}]_s}{\underline{\alpha}_0 (e^{\mathbf{Q}^* t^*} - \mathbf{I})(\mathbf{Q}^*)^{-1} \underline{1}}. \qquad (7)
\end{aligned}
$$

which we write in vector form as

$$
\underline{\pi}^*(t^*) = \frac{\underline{\alpha}_0 (e^{\mathbf{Q}^* t^*} - \mathbf{I})(\mathbf{Q}^*)^{-1}}{\underline{\alpha}_0 (e^{\mathbf{Q}^* t^*} - \mathbf{I})(\mathbf{Q}^*)^{-1} \underline{1}}, \qquad (8)
$$

Fitting the model to whole-genome derived sequence data

Thus, we can fit $\underline{\pi}^*(1)$ to the data, which naturally defines a molecular clock via the mutation-rate parameters of the model.

Fitting the model to whole-genome derived sequence data

Thus, we can fit $\underline{\pi}^*(1)$ to the data, which naturally defines a molecular clock via the mutation-rate parameters of the model.

Notice that as $t^* \to \infty$ $\underline{\pi}^*(t)$ tends to the ratio of means distribution,

$$\lim_{t^* \to \infty} \underline{\pi}^*(t^*) = \frac{\underline{\alpha}_0 (\mathbf{Q}^*)^{-1}}{\underline{\alpha}_0 (\mathbf{Q}^*)^{-1} \underline{1}}. \tag{9}$$

Fitting the model to whole-genome derived sequence data

Thus, we can fit $\underline{\pi}^*(1)$ to the data, which naturally defines a molecular clock via the mutation-rate parameters of the model.

Notice that as $t^* \to \infty$ $\underline{\pi}^*(t)$ tends to the ratio of means distribution,

$$\lim_{t^* \to \infty} \underline{\pi}^*(t^*) = \frac{\underline{\alpha}_0(\mathbf{Q}^*)^{-1}}{\underline{\alpha}_0(\mathbf{Q}^*)^{-1}\underline{1}}. \tag{9}$$

This provides a way to test the validity of the assumption of independence — if the fitted $\underline{\pi}^*(1) \approx \lim_{t^* \to \infty} \underline{\pi}^*(t^*)$, then we can conclude that the empirical distribution is at or near equilibrium.

Fitting the model to whole-genome derived sequence data

We have a method to fit our model to empirical microsatellite distributions which can

- identify the extent of the slowdown in slipped-strand mispairing due to interruptions in the repeat sequence

Fitting the model to whole-genome derived sequence data

We have a method to fit our model to empirical microsatellite distributions which can

- identify the extent of the slowdown in slipped-strand mispairing due to interruptions in the repeat sequence
- test the support for the commonly used dynamic bias function

Fitting the model to whole-genome derived sequence data

We have a method to fit our model to empirical microsatellite distributions which can

- identify the extent of the slowdown in slipped-strand mispairing due to interruptions in the repeat sequence
- test the support for the commonly used dynamic bias function
- test the veracity of the assumption of equilibrium in empirical microsatellite distributions

Fitting the model to whole-genome derived sequence data

We have a method to fit our model to empirical microsatellite distributions which can

- identify the extent of the slowdown in slipped-strand mispairing due to interruptions in the repeat sequence
- test the support for the commonly used dynamic bias function
- test the veracity of the assumption of equilibrium in empirical microsatellite distributions

However, there is one major problem with this model — we require data from bona-fide microsatellite sequences which includes information about interruptions.

Fitting the model to whole-genome derived sequence data

We have a method to fit our model to empirical microsatellite distributions which can

- identify the extent of the slowdown in slipped-strand mispairing due to interruptions in the repeat sequence
- test the support for the commonly used dynamic bias function
- test the veracity of the assumption of equilibrium in empirical microsatellite distributions

However, there is one major problem with this model — we require data from bona-fide microsatellite sequences which includes information about interruptions.

In theory, TRF (or similar software) applied to whole-genome data should provide this. In practice, our data appears to be polluted with non-microsatellite sequences.

More on microsatellites

Fundamentally, the problem is that repetitive structure is not enough for a sequence to be considered a microsatellite. It must also exhibit ‘characteristic microsatellite behaviour‘ — i.e. it should undergo high rates of slipped-strand mispairing.

More on microsatellites

Fundamentally, the problem is that repetitive structure is not enough for a sequence to be considered a microsatellite. It must also exhibit ‘characteristic microsatellite behaviour‘ — i.e. it should undergo high rates of slipped-strand mispairing.

This is particularly problematic when considering sequences which are slowed down due to the introduction of interruptions to the repeat sequence.

More on microsatellites

Fundamentally, the problem is that repetitive structure is not enough for a sequence to be considered a microsatellite. It must also exhibit ‘characteristic microsatellite behaviour‘ — i.e. it should undergo high rates of slipped-strand mispairing.

This is particularly problematic when considering sequences which are slowed down due to the introduction of interruptions to the repeat sequence.

We not only need to separate repetitive non-microsatellite sequences from proper microsatellites, but we also need to identify ‘ex-microsatellites‘ — repetitive sequences which were evolving rapidly due to slipped-strand mispairing before becoming highly interrupted.

Acknowledgements

Supervisors

- Dr Małgorzata O'Reilly
- Dr Barbara Holland

- Dr Bennet McComish