Inference in a probabilistic model of dynamic DNA

April 2009

Catherine Higham, Lord Kelvin 1st year PhD student
Darren Monckton, Dan Haydon, Christina Cobbold and Douglas Wilcox
Probabilistic model
- Stochastic “Birth and Death” process
- Discrete-state continuous-time
- Capture key features DNA mutation mechanism

Rich data set (30,000 *de novo* mutations)
Bayesian framework to calibrate model
Dynamic DNA
Myotonic dystrophy
Initial findings
myotonic dystrophy type 1 (DM1)

- Common (1/8000)
- Variable symptoms including
  - Myotonia
  - Wasting limb and facial muscles
  - Cataracts

Need for quantitative predictive tools
CTG repeat expansion in DM1

- >1,000 congenital
- 200-500 adult onset
- <100 late onset
Anticipation in DM1

- 72
- 112
- 205
- 160
- 2100
- 730

- late onset
- adult onset
- congenital
Variation between blood cells from one patient

Small pool analysis

32 x 6 molecules

1,284
951
618
501
284
234

F Morales

~4 cells

Learning in Computational Systems Biology
Patient repeat length distributions

Number of CTG repeats

- Age 66
  - Onset asym

- Age 31
  - Onset 19

- Age 13
  - Onset 10

- Age 35
  - Onset 27
Kaplan et al, *PLoS Computational Biology* 2007 proposed probabilistic model based on expansion only

- Fitted model to mean length and age of onset for several diseases using standard fitting tools

Expansion = “Birth”
Kaplan et al, *PLoS Computational Biology* 2007 proposed probabilistic model based on expansion only

We are generalising this model to include contraction

- Evidence of contractions
- Use Bayesian framework to evaluate this hypothesis
- Calibration using the whole dataset and not key statistics

Expansion = “Birth”
Contraction = “Death”
Evolution of repeat length

Using biologically realistic parameters

CTG repeat length vs. age (years)
Evolution of repeat length

Using biologically realistic parameters

![Graph showing evolution of repeat length over age (years). The y-axis represents CTG repeat length, and the x-axis represents age (years). The graph includes multiple lines of different colors, indicating various scenarios or conditions. The left inset graph provides a close-up view of the evolution at younger ages.](image-url)
Bayesian framework

Bayes’ Theorem links what we can quantify to what we would like to know

\[
prob(\theta|data) \text{ is proportional to } prob(data|\theta) \cdot prob(\theta)
\]

- Let \( P_n(t) \) be probability that #CTG repeats is \( n \) at patient age \( t \)
- \( P_n(t) \) obtained by solving high dimensional master equation one ODE for each possible repeat length
- Assuming cell lengths evolve independently, likelihood is product of \( P_n(t) \) over all the data points
- Use Markov chain Monte Carlo (MCMC) to obtain posterior pdfs for parameters
Posterior pdfs

DM1 patient (data from blood cells) age 56
Posterior envelope
Initial findings

• Evidence for contractions
• The observed tendency towards expansion of repeat length is the net result of many more expansion and contraction mutations than previously thought

Future work

• Investigate the full dataset
• Explore other forms for expansion and contraction
• Mathematical modelling will be extended to incorporate new data being generated by our lab
Aims

Improve prognostic information
genetic counselling, age of onset, severity of disease

Clinical trials
account for variation and lower error bars for drugs

Slow/reverse repeat is therapeutic target

Biological understanding
Shed light mechanism unstable DNA
Stochastic “birth and death” process

Suppose repeat length is $n$ at time $t$, $\lambda$ is the expansion rate, $\mu$ is the contraction rate and $a$ is the threshold, then at time $t + \delta t$:

- Probability length is $n + 1 \approx \lambda (n-a) \delta t$
- Probability length is $n - 1 \approx \mu (n-a) \delta t$
- Probability length is $n \approx 1 - (\lambda + \mu) (n-a) \delta t$

Each coloured line represents the evolution of repeat length in one cell.
We derive an expression for $P_n(t | \lambda, \mu, a)$ the probability of length $n$ at time $t$ given parameter values for $\lambda$, $\mu$ and $a$:

$$P_n(t+\delta t) = P_n(t) \left[1-(n-a)(\lambda-\mu)\delta t\right] + P_{n-1}(t) (n-a-1) \lambda \delta t + P_{n+1}(t) (n-a+1) \mu \delta t$$

Dividing by $\delta t$ and letting $\delta t \to 0$ gives

$$\frac{dP_n(t)}{dt} = \lambda (n-a-1) P_{n-1}(t) - (\lambda+\mu) (n-a) P_n(t) + \mu (n-a+1) P_{n+1}(t)$$

or equivalently

$$\frac{dP_n}{dt} = AP_n$$

with solution $P_n = e^{At}P_a$

Numerical approach will allow $\lambda$ and $\mu$ to vary with $n$