MODELLING DYNAMIC DNA IN MYOTONIC DYSTROPHY

Catherine F. Higham[1,2], Douglas E. Wilcox[3], Daniel T. Haydon[1], Christina Cobbold[2] and Darren G. Monckton[1]

[1] Faculty of Biomedical and Life Sciences, University of Glasgow,
Anderson College Building, 56 Dumbarton Road, Glasgow G11 6NU, UK,
[2] Department of Mathematics, University of Glasgow, University Gardens, Glasgow G12 8QW, UK,
[3] Scottish Muscle Network, Ferguson-Smith Center for Clinical Genetics,
Yorkhill Hospital, Dalnair Street, Glasgow G3 8SJ, UK
c.higham.1@research.gla.ac.uk

ABSTRACT

Myotonic dystrophy type 1 is the most common of about 20 human diseases associated with inheriting an abnormally large unstable DNA simple sequence repeat. New quantitative data, collected by single molecule analysis of repeat length in blood cells from 145 patients reveals the extent and nature of the genetic cell to cell variation within (somatic mosaicism) and between patients.

We are developing discrete-state continuous-time mathematical models and stochastic simulation techniques that capture key features of the mutation mechanism underlying repeat length evolution. Modern Bayesian techniques involving Markov chain Monte Carlo are employed to calibrate our models against the biological data and test model hypotheses. This work has the potential to improve prognostic information for patients, as well as providing deeper understanding of the underlying biological process.

We report here an initial finding that the distribution of repeat length is better described by a threshold birth and death process model than by a traditional pure birth process. This suggests that the underlying biological mechanism consists of both expansions and contractions and that the observed tendency towards repeat expansion is the net result of many more expansion and contraction mutations than previously thought. Our estimates for expansion and contraction give only a slight bias (2%) in favour of repeat expansion and they predict that mutation events occur at a much greater rate than a pure birth model would allow.

1. INTRODUCTION

Simple sequence tandem repeats in DNA, for example the motif CTG repeated multiple times, are known as microsatellites and are present in both coding and non-coding regions of the genome. One of the ways these sequences mutate is by a change in the number of repeats and these length changes occur more frequently than other types of mutations. Because of their high mutation rates, these microsatellites come in many forms making them popular genetic markers. However, changes in repetitive sequences, even when they occur in a non-coding region, may lead to disease. In particular, genomic amplification of simple trinucleotide repeats is the underlying genetic defect in a number of human diseases including myotonic dystrophy type 1 (DM1).

DM1 patients have an expansion of unstable CTG repeats located in the untranslated region of a gene encoding a protein kinase, reviewed in [1]. The CTG repeat is polymorphic in the general population ranging from 5 to 37 repeats in healthy individuals and from upwards of 50 to several thousand in affected DM1 patients. The different variants of a specific gene are known as alleles and expanded disease-associated alleles of greater than 50 CTG repeats are unstable in both the germline and soma. Germline expansion accounts for the phenomenon of anticipation whereby the disorder becomes apparent at an earlier age, often with an increase in the severity of the symptoms, as it is passed on to the next generation. Expansion of the unstable alleles over time and variation in the level of mutation between the somatic tissues of an individual are thought to account at least partially for the tissue specificity and progressive nature of the symptoms. Studies in the past ten years have linked instability to the DNA replication, repair and recombination machineries [2], but expansion activity does not obviously correlate with cell turnover [3] and so there is increasing focus on repair [4].

We are interested in developing a mathematical model that captures the key features of the mutation mechanism underlying repeat length evolution. Currently patients concerned about their own prognosis and their reproductive choices have limited information available to them about how their disease will progress. This is because variance in mean length only accounts for 25% of the variance in age of onset. Low correlation between age of onset of symptoms and mean repeat length is in part due to the anticipation associated with DM1 and sampling bias caused by the tendency for people to be tested only when they or a member of their family presents with symptoms. Thus, there is great potential for more sophisticated modelling and inference techniques to improve the prognostic value of genetic information. There is, however, a clear link between the degree of somatic mosaicism of the repeat lengths and the progression of the disease [3]. A more reliable measure for patients would be an indication of pro-

genitor allele length, that is, the length of the allele inherited. A key objective of our work is, therefore, to infer progenitor allele length.

In this paper we discuss the mathematical model we have developed and apply our methods to a published data set [3]. Data collected using small pool analysis of length in blood cells from a male DM1 patient aged 56 and comprises 325 alleles of different lengths so this dataset is sufficient for us to fit the model at the individual patient level (Figure 1). This means that we can examine the evolution and infer the progenitor allele length.

In the non-disease case there exist models for microsatellite evolution, which are summarised in [5]. However mutation at these sites occurs at lower rates and typically involves shorter lengths than in the pathological disease case. Also these models tend to assume that an equilibrium in the distribution of lengths has been reached in the population. In the pathological disease case the data suggests that the distribution of length is time-dependent throughout the life of a patient. This makes the analysis different as we cannot assume equilibrium status. However these models form a useful basis for our work. The oldest model for microsatellite evolution is the stepwise mutation model originally proposed by Ohta and Kimura for electrophoretic alleles [6]. Kruglyak proposed a proportional slippage model where the mutation rate increases linearly with microsatellite length [7]. Although most observed microsatellite mutations are by one repeat unit not all are so, Di Rienzo proposed a model which allows for larger mutations [8]. There have been several further developments, we refer to [5] for details. Kaplan et al. [9] use a simple birth process to simulate allele evolution and derive expressions to fit clinical data for a range of diseases associated with expanded repeats. These models assume that the expansion bias observed in patients is solely due to expanding lengths. We would like to investigate the possibility that the expansion bias is due to the difference between expansion and contraction mutations. We use the same stochastic modelling framework, but introduce a threshold below which expansion and contraction is disallowed.

2. METHODS

Kaplan et al. fit a discrete-state continuous-time stochastic model (based on a simple birth process) to clinical data relating to mean allele length and age of onset using derived analytical expressions for the mean and standard software (Matlab Nelder-Mead simplex direct search least square error). We are fortunate that our data set allows us to go beyond the mean by providing us with a large set of samples from the distribution of allele lengths at a known point in time. Microsatellites are known to expand and contract and there is evidence from distributions collected for patients at two time points that some alleles may have contracted [10]. This justifies our extension of the model to include contraction as well as expansion.

We propose to consider expansion and contraction rates linearly proportional to allele length and with a threshold

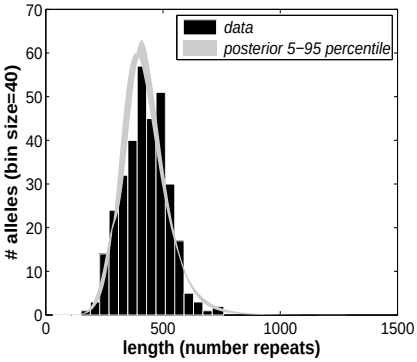

Figure 1. Allele length distribution in blood cells for a male DM1 patient aged 56. The predicted distribution is calculated for each of the MCMC samples and the shaded area shows the predicted 5-95 percentile range for this patient.

of 50, this value being based on observations from family studies. We derive an analytical expression for the derivative of the probability distribution of allele length n given expansion and contraction rates λ and μ respectively (1). We solve this equation numerically by truncating at a suitably large value of n and treating it numerically as a system of ordinary differential equations (ODEs). Several other methods were investigated [11] including manipulation of the exact solution using the exponential of the matrix, eigenvalues and Taylor series expansion. However the treatment using the stiff ODE solver ode15s from Matlab proved to be the quickest solution, in some cases 50-fold, which is an important consideration for parameter estimation where exploration of the parameter space requires the equations to be solved many times.

2.1. Mathematical Model

The approach we take to modelling is a birth and death process involving a threshold. We refer to [12] for further motivation and explanation of the basic methology. Suppose that the length, l, defined as the number of CTG repeats, is n at time t and λ is the rate of expansion above a threshold length called a and μ is the rate of contraction above a. Then at time $t + \delta t$:

- the probability that l is $n + 1 \approx \lambda (n - a) \delta t$
- the probability that l is $n - 1 \approx \mu (n - a) \delta t$
- the probability that l is $n \approx 1 - (\lambda + \mu) (n - a) \delta t$

Now let $P_n(t)$ denote the probability that an allele has length n at time t. Then the rate of change of $P_n(t)$ with respect to time is

$$\begin{aligned} \dot{P}_n(t) &= -(\lambda + \mu)(n - a) P_n(t) \\ &\quad +\lambda (n - a - 1) P_{n-1}(t) \\ &\quad +\mu (n - a + 1) P_{n+1}(t), \end{aligned} \tag{1}$$

where for convenience we set $P_{a-1}(t) \equiv 0$.

In order to derive expressions for mean and variance, we note that the first and second moments of $P_n(t)$, $M(t)$ and $M_2(t)$, respectively, are:

$$M(t) = \sum_{n \geq a} nP_n(t), \quad (2)$$

$$M_2(t) = \sum_{n \geq a} n^2 P_n(t). \quad (3)$$

Differentiating both (2) and (3) with respect to t and substituting (1) into the result gives expressions for $\dot{M}(t)$ and $\dot{M}_2(t)$ which after algebraic manipulation become:

$$\dot{M}(t) = (\lambda - \mu)(M(t) - a), \quad (4)$$

$$\dot{M}_2(t) = 2(\lambda - \mu) M_2(t) + (\lambda + \mu) M(t). \quad (5)$$

Solving (4) and (5) with $M(t=0) = n_0$ where n_0 is the length of the progenitor allele and setting $V(t) = M_2(t) - (M(t))^2$ for the variance at time t gives the following analytical expressions:

$$M(t) = (n_0 - a) e^{(\lambda-\mu)t} + a, \quad (6)$$

$$V(t) = (n_0 - a)\left(\frac{\lambda+\mu}{\lambda-\mu}\right)\left(e^{2(\lambda-\mu)t} - e^{(\lambda-\mu)t}\right),$$

which generalise the standard $a = 0$ case [12].

2.2. Parameter Estimation

We used a Bayesian framework [13] and Markov chain Monte Carlo (MCMC) Metropolis Hastings [14], with a Gaussian proposal function, for inferring the model parameters from the dataset described in Figure 1. Bayes' theorem links the quantities that we are interested in, the probability of the parameters given the data, to quantities we can estimate numerically. The latter involves (a) the likelihood of the data given the parameters and (b) our prior knowledge about the parameters before we see the data. The likelihood $L(\theta)$ of our parameters

$\theta = \{n_0, \lambda, \mu\}$ will be the solution $P_n(D|\theta)$ where

$D = \{d_1, d_2, \ldots d_{325}\}$ is the data comprising of 325 allele lengths. Assuming that the evolution of individual alleles is independent $P_n(D|\theta) = \prod_{k=1}^{325} P_n(d_k|\theta)$. We propose almost uniform priors for θ by using a gamma distribution $\Gamma(1,1)$. Bayes' theorem says

$$p(\theta|D) \propto P_n(D|\theta)\, p(\theta). \quad (7)$$

The relation in (7) is expressed using proportionality because the term $p(D)$ has been omitted from the denominator in the right hand side. This is appropriate for parameter estimation since the missing denominator is simply a normalisation constant not depending explicitly on the parameters.

Initial investigation of the likelihood surface suggests that that the data holds more information about $\lambda - \mu$ than about either λ or μ. This is consistent with equation (6), where the mean is seen to depend only on λ and μ through their difference, $\lambda - \mu$. We mention, as an aside, $M(t)$ in (6) would arise from the deterministic analogue of this birth and death process, in which case only $\lambda - \mu$, and not λ and μ individually, could be inferred. With only three parameters to infer, a brute force grid search is perhaps a feasible alternative to MCMC. However the MCMC approach allows us to collect samples from the posterior distribution, $p(\theta|D)$ which can be used to show the uncertainity in the fit. Use of an MCMC method also paves the way for future refinements of the model which may increase dimensionality.

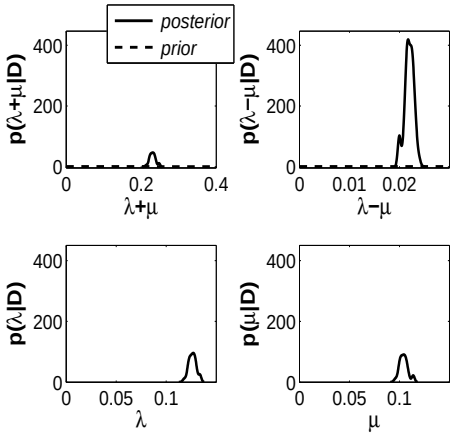

Figure 2. Prior and Posterior probability distributions for $\lambda + \mu$, $\lambda - \mu$ and the same information transformed into expansion and contraction rates λ and μ. Threshold a = 50 and t = 56 (patient age).

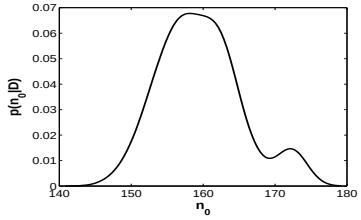

Figure 3. Posterior probability distribution for the progenitor allele length, n_0.

There are three parameters in this model λ, μ and n_0. However, we can simplify the problem by using (6) to express n_0 in terms of λ, μ and $M(t)$:

$$n_0 = (M(t) - a) e^{-(\lambda-\mu)t} + a. \quad (8)$$

3. RESULTS

In order to obtain posterior probability distribution functions (pdfs) for the model parameters, λ, μ and n_0, we used a Bayesian framework to fit the model to the data. The results for $\lambda + \mu$, $\lambda - \mu$ and the same results transformed for λ and μ are shown in Figure 2. The posterior pdfs for $\lambda + \mu$ and $\lambda - \mu$ clearly modify the uninformative prior and provide strong support for their values of around 0.23 and 0.022 respectively. The transformation of these pdfs into their counterpart pdfs for λ and μ show strong evidence for μ not equal to zero, which supports

the hypothesis that contractions as well as expansions explain the observed variance of the allele lengths. Figure 3 shows the estimated posterior pdf for n_0. The distribution has a second mode and wide variance suggesting that alternative values of n_0 would provide almost as good a fit as the main peak. Results like this show the importance of considering the full distribution and not relying on a point estimate. The MCMC samples can be used to predict the distribution of alleles at the age of the patient and the shaded area in Figure 1 shows the predicted 5-95 percentile range for this patient. The peak of the prediction lies over the data mode and closely follows the data variation.

4. CONCLUSION

Under the assumption that a discrete stochastic birth and death process is an appropriate model to explain the evolution of allele length in the pathelogical disease case, we have shown that there is support for the expansion bias to be the net result of expansion and contraction. For the particular patient investigated, the bias is about 0.022 per unit length which equates to 102 steps forward and 100 steps back. Fitting this model also provides an estimate for the progenitor allele length n_0, which is of direct clinical interest. At a mechanistic level, this stepping forward and backwards is aggravating the level of variance within the alleles. The uncovering of this possible mechanism provides insights in to allele evolution that could be validated independently.

We are currently applying the methods discussed here to a larger data set [15] containing about 30,000 *de novo* mutations from 145 DM1 patients from families in Scotland, USA and Chile. Future work includes inference over multiple patients and the evaluation of more sophisticated birth and death models.

5. ACKNOWLEDGMENTS

CFH is funded by a Lord Kelvin PhD scholarship from the University of Glasgow. We would like to thank members of our lab, Dr Claudia Braida and Dr Jill Couto, for useful discussions about the nature of the data. We would also like to thank Dr Fernando Morales for access to unpublished data.

6. REFERENCES

[1] M. Gomes-Pereira and D. G. Monckton, "Chemical modifiers of unstable expanded simple sequence repeats: what goes up, could come down," *Mutation Research*, vol. 598, no. 1-2, June 2006.

[2] S. M. Mirkin, "Expandable DNA repeats and human disease," *Nature*, vol. 447, no. 7147, June 2007.

[3] D. G. Monckton, L. J. Wong, T. Ashizawa, and C. T. Caskey, "Somatic mosaicism, germline expansions, germline reversions and intergenerational reductions in myotonic dystrophy males: small pool PCR analyses," *Human Molecular Genetics*, vol. 4, no. 1, 1995.

[4] I. V. Kovtun and C. T. McMurray, "Features of trinucleotide repeat instability in vivo," *Cell Research*, vol. 18, no. 1, pp. 198–213.

[5] P. Calabrese and R. Sainudiin, "Models of microsatellite evolution," *Statistical Methods in Molecular Evolution*, pp. 289–305, 2005.

[6] T. Ohta and M. Kimura, "A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population," *Population Genetics, Molecular Evolution, and the Neutral Theory: Selected Papers*, 1994.

[7] S. Kruglyak, R. T. Durrett, M. D. Schug, and C. F. Aquadro, "Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, no. 18, pp. 10774–10778, 1998.

[8] A. D. Rienzo, A. C. Peterson, J. C. Garza, A. M. Valdes, M. Slatkin, and N. B. Freimer, "Mutational processes of simple-sequence repeat loci in human populations," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 91, no. 8, pp. 3166–3170, Apr. 1994.

[9] S. Kaplan, S. Itzkovitz, and E. Shapiro, "A universal mechanism ties genotype to phenotype in trinucleotide diseases," *PLoS Computational Biology*, vol. 3, no. 11, Nov. 2007.

[10] L. Martorell, D. G. Monckton, J. Gamez, K. J. Johnson, I. Gich, A. L. de Munain, and M. Baiget, "Progression of somatic CTG repeat length heterogeneity in the blood cells of myotonic dystrophy patients," *Hum. Mol. Genet.*, vol. 7, no. 2, pp. 307–312, Feb. 1998.

[11] C. Moler and C. V. Loan, "Nineteen dubious ways to compute the exponential of a matrix, Twenty-Five years later," *SIAM Review*, vol. 45, no. 1, pp. 3–49, 2003.

[12] E. Renshaw, *Modelling Biological Populations in Space and Time*, Cambridge University Press, 1991.

[13] M. A. Beaumont and B. Rannala, "The Bayesian revolution in genetics," *Nature Reviews Genetics*, vol. 5, no. 4, pp. 251–261, 2004.

[14] C. Andrieu, N. de Freitas, A. Doucet, and M. I. Jordan, "An introduction to MCMC for machine learning," *Machine Learning*, vol. 50, no. 1-2, pp. 5–43, 2003.

[15] F. A. Morales, *Somatic mosaicism and genotype-phenotype correlations in myotonic dystrophy type 1*, Ph.D. thesis, University of Glasgow, 2006.