The SMM model as a boundary value problem using the discrete diffusion equation

Joel Campbell*

NASA Langley Research Center, MS 488, Hampton, VA 23681, USA

Received 28 March 2007
Available online 19 August 2007

Abstract

A generalized single-step stepwise mutation model (SMM) is developed that takes into account an arbitrary initial state to a certain partial difference equation. This is solved in both the approximate continuum limit and the more exact discrete form. A time evolution model is developed for Y DNA or mtDNA that takes into account the reflective boundary modeling minimum microsatellite length and the original difference equation. A comparison is made between the more widely known continuum Gaussian model and a discrete model, which is based on modified Bessel functions of the first kind. A correction is made to the SMM model for the probability that two individuals are related that takes into account a reflecting boundary modeling minimum microsatellite length. This method is generalized to take into account the general n-step model and exact solutions are found. A new model is proposed for the step distribution.

Keywords: Discrete diffusion; SMM; Stepwise mutation; Infinite alleles; Microsatellite; n-step; Bessel

1. Introduction

Microsatellites are a special class of tandem repeats of strings in the genome consisting of one to six bases (1–6 bp). These short strings can be repeated up to about 100 times. These simple tandem repeats (STRs) are often used to determine relatedness between individuals and even between species. In each generation there is a small probability there will be a copy mistake (mutation) in the resulting microsatellite length. Those STRs with the highest mutation rates are most useful for predicting time to a common ancestor.

There are many different approaches to solving this problem mathematically. The typical approach used in modern times is to use direct statistical methods (Walsh, 2001). The approach we take is the older one used by Ohta and Kimura (1973) and others which is to treat random genetic drift as a diffusion process and solve the resulting differential equations directly. The main disadvantage to this approach in the past has been that one must make a continuum approximation, which makes it less accurate in certain situations. This means that method is not always very useful for determining the relatedness of very closely related individuals. This particular issue is dealt with here by deriving discrete partial difference equations in the form of discrete diffusion equations and solving them instead, thereby avoiding the issues associated with continuum approximations.

The discrete diffusion equation is widely used in many other contexts. For instance, it has been applied to the area of population growth (Lu and Takeuchi, 1993) where one wishes to model geographic spread in addition to growth in number. In the area of physics it is used to model ionic diffusion on a lattice (Fath, 1998). It has also been used in digital filtering in the form of diffusion filtering (Lindeberg, 1990).

2. A difference equation for the stepwise mutation model (SMM)

Let \( N(m,n) \) be the number of individuals with a particular STR marker having exactly \( m \) repeats at the \( n \)th generation. Let us suppose this marker can mutate at a rate \( r \) per generation. The probability of \( m \) mutating to

---

*Fax: + 1 757 864 7944.
E-mail address: joel.f.campbell@nasa.gov

0040-5609/$ - see front matter Published by Elsevier Inc.
doi:10.1016/j.tpb.2007.08.001
m + 1 or m - 1 is r/2. If m mutates, there is an equal loss of the unmutated state. This suggests a difference equation of the following form:

$$N(m, n + 1) = N(m, n) + \frac{r}{2} (N(m + 1, n) - 2N(m, n) + N(m - 1, n)).$$

(1)

This is a form of the discrete diffusion equation. If we instead consider the asymmetric case where the probability of mutating to \(m + 1\) is different from mutating to \(m - 1\) we have

$$N(m, n + 1) = N(m, n) + \frac{r}{2} (N(m + 1, n) - 2N(m, n) + N(m - 1, n)) + \frac{r - r'}{2} (N(m, n) - N(m - 1, n)).$$

(2)

which is the more general case of a discrete form of the Fokker–Planck equation (Reif, 1965) of statistical mechanics. As demonstrated here, the discrete diffusion equation comes about in a very natural way.

In addition to (1) and (2) this there is a boundary condition which restricts \(m \geq 0\). This is discussed in great detail later on. We do the case without the boundary first since that is the case one finds most often in the literature for those who have found explicit analytic results using statistical or other methods. As we shall see later on there is no loss in generality since one can construct either a reflective or an absorptive boundary solution from the solution without the \(m = 0\) boundary using a well-known trick from mathematical physics.

3. The SMM in the continuum

In the continuum limit this becomes

$$\frac{\partial}{\partial t} N(x, t) = \frac{r}{2} \frac{\partial^2}{\partial x^2} N(x, t).$$

(3)

This is the well-known diffusion/heat equation (Reif, 1965) and is easily solved using standard techniques. A solution localized at \(t = 0\) is

$$N(x, t) = \frac{1}{\sqrt{2\pi rt}} \exp \left[ -\frac{(x - x_0)^2}{2rt} \right].$$

(4)

This represents the solution where each individual is descended from a single source. A similar approach has been applied using the SMM model in the continuous limit when comparing the distance between two distinct populations (Zhitkovsky and Feldman, 1995) using the standard deviation as the independent variable. The variance for (4) is given by

$$\sigma^2 = rt.$$

(5)

This is a well-known result for estimating age (Goldstein, 1995). A more general solution to (3) is (Courant and Hilbert, 1962)

$$N(x, t) = \frac{1}{\sqrt{2\pi rt}} \int_{-\infty}^{\infty} N(x', 0) \exp \left[ -\frac{(x - x')^2}{2rt} \right] dx'.$$

(6)

With an initial state of

$$N(x, 0) = \delta(x - x_0),$$

(7)

(4) is recovered.

The interpretation of (6) is that given an initial distribution, \(N(x, 0)\), \(N(x, t)\) shows how that initial distribution evolves over time.

4. The semi-discrete method

An alternative to (3) is where we allow the \(x\) variable to be discrete and the time variable to be continuous. In this case we have

$$\frac{\partial}{\partial t} N(m, t) = \frac{r}{2} (N(m + 1, t) - 2N(m, t) + N(m - 1, t)).$$

(8)

It can be shown a solution to (8) is

$$N(m, t) = \exp[-rtJ_{m-m_0}(rt)],$$

(9)

where \(J\) is the modified Bessel function of the first kind (Abramowitz and Stegun, 1972). This is a result that has been found using statistical methods with the SMM model (Walsh, 2001). He also shows how this may be applied to Y DNA and mtDNA time estimates. A more general solution to (8) may be found using Fourier series techniques. We first multiply both sides of (8) by \(\exp(i\omega t)\) and sum over \(m\). The result is

$$\frac{\partial}{\partial t} \hat{N}(\omega, t) = \frac{r}{2} \left[ \exp(i\omega t)\hat{N}(\omega, t) - 2\hat{N}(\omega, t) + \exp(-i\omega t)\hat{N}(\omega, t) \right]$$

$$= -2r \sin^2 \left( \frac{\omega}{2} \right) \hat{N}(\omega, t),$$

(10)

where

$$\hat{N}(\omega, t) = \sum_{m=-\infty}^{\infty} \exp(i\omega m) N(m, t)$$

(11)

is the characteristic function for the solution to (8). To recover the solution from the characteristic function we use

$$N(m, t) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \exp(-i\omega m) \hat{N}(\omega, t) d\omega.$$

(12)

Eq. (10) is easily solved and has the solution

$$\hat{N}(\omega, t) = \hat{N}(\omega, 0) \exp \left[ -2r \sin^2 \left( \frac{\omega}{2} \right) t \right].$$

(13)

Using the results of (11)–(13) we find

$$N(m, t) = \frac{1}{2\pi} \sum_{m'=-\infty}^{\infty} N(m', 0) \int_{-\pi}^{\pi} \exp[-i\omega(m' - m)]$$

$$\times \exp \left[ -2r \sin^2 \left( \frac{\omega}{2} \right) t \right] d\omega.$$
Using the identity $\sin^2(x) = (1 - \cos(2x))/2$ and (Abramowitz and Stegun, 1972) we find

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} \exp(-i\omega m) \exp\left[-2r \sin^2\left(\frac{\omega}{2}\right) t\right] d\omega = \exp(-rt) I_m(rt).$$

We have then

$$N(m, t) = \exp(-rt) \sum_{m' = -\infty}^{\infty} N(m', 0) I_{m-m'}(rt).$$

These same results can also be obtained by assuming discrete $t$ but small $r$ (A.1). The extension and uniqueness of this type of solution into the discrete domain has been demonstrated by others in a different context (Lindeberg, 1990). In reality $m$ can never be negative so the initial state will only include non-zero values for $m' > 0$. As long as the initial state is sufficiently localized and far from $m' = 0$ (as is typically the case) there are no boundary effects. This is particularly true when comparing modern human beings within the same haplogroup. Boundary effects are treated in the next section in some detail.

It should be noted that Fourier series techniques have been applied in the past (Fath, 1998) to solve the discrete reaction–diffusion equation. The main difference here is that we are using the difference equation to solve for a characteristic function.

### 5. Reflecting and absorbing boundaries

As stated earlier the STR length, $m$, must be greater than zero in real life. Once $m$ reaches some lower limit (let us pick 0 for now) it can get no smaller. The most general boundary condition is

$$N(x, t) = 0, \quad x < 0.$$  

However, it is a bit more complicated than this because this is not specific enough in that there are different possibilities that can lead to this condition. The two possibilities considered here are reflecting and absorbing boundaries. In the absorbing case

$$N(0, t) = 0.$$  

The reflecting case is where

$$\left(\frac{\partial N(x, t)}{\partial x}\right)_{x = 0} = 0.$$  

To give a physical analogy we refer to the heat equation, which is mathematically identical. The absorbing case would be equivalent to a semi-infinite bar held at zero temperature at the origin with an arbitrary initial state. There is a net flow of heat toward the $x = 0$ boundary. In our case we would have exactly the same thing—the normalization of the probability density is not conserved over time and this leads to the eventual death of the STR. The reflecting case is just the opposite. The heat analogy is where there is either a mirror or insulator at one end. In that case there is no net flow into the origin—everything reflects back. In this case the normalization of the probability density is conserved. Although one does occasionally find the absorbing case in the literature in the context of STRs (Farrall and Weeks, 1998), the reflecting case is more commonly discussed (Nauta and Weissing, 1996). It is the opinion of the author the reflecting boundary is what most closely models nature. The reason being, that if the absorbing boundary were correct, there would not be any STRs. As soon as one formed it would soon die out in future generations. There has to be some mechanism to preserve them—whether that be natural selection or some other process.

We model this with a simple reflecting barrier with the understanding the actual process is most likely much more complicated.

Let us first consider the continuous reflecting case. Using the boundary condition in (19), suggests a solution of the form

$$N(x, t) = \frac{2}{\pi} \int_{0}^{\infty} a(\omega, t) \cos(\omega x) d\omega,$$

where

$$a(\omega, t) = \int_{0}^{\infty} N(x, t) \cos(\omega x) dx.$$  

We first use (21) and substitute it into (3). The result is

$$\frac{\partial}{\partial t} a(\omega, t) = -\frac{r}{2} \omega^2.$$  

The solution is

$$a(\omega, t) = a(\omega, 0) \exp\left(-\frac{r}{2} \omega^2 t\right),$$  

so that

$$N(x, t) = \frac{2}{\pi} \int_{0}^{\infty} a(\omega, 0) \exp\left(-\frac{r}{2} \omega^2 t\right) \cos(\omega x) d\omega.$$  

This can be cast into the form

$$N(x, t) = \int_{0}^{\infty} N(x', 0) G(x, x', t) dx'.$$

To find $G(x, y, t)$ we use $N(x, 0) = \delta(x - y)$ so that $a(\omega, 0) = \cos(\omega y)$. From this we find

$$G(x, y, t) = \frac{2}{\pi} \int_{0}^{\infty} \exp\left(-\frac{r}{2} \omega^2 t\right) \cos(\omega y) \cos(\omega x) d\omega$$

$$= \frac{1}{\sqrt{2\pi rt}} \left[ \exp\left(-\frac{(x - y)^2}{2rt}\right) + \exp\left(-\frac{(x + y)^2}{2rt}\right) \right].$$  

We now have

$$N(x, t) = \frac{1}{\sqrt{2\pi rt}} \int_{0}^{\infty} N(x', 0) \left[ \exp\left(-\frac{(x - x')^2}{2rt}\right) + \exp\left(-\frac{(x + x')^2}{2rt}\right) \right] dx'.$$
As an example we choose an initial state \( N(x,0) = \delta(x-c) \) so that
\[
N(x,t) = \frac{1}{\sqrt{2\pi rt}} \left[ \exp\left(-\frac{(x-c)^2}{2rt}\right) + \exp\left(-\frac{(x+c)^2}{2rt}\right) \right].
\]
(28)

The discrete version can be solved in a similar manner. The result is
\[
N(m,t) = \exp(-rt) \sum_{m'=0}^{\infty} N(m',0)(I_{m-m'}(rt) + I_{m+m'}(rt)),
\]
m \( \geq 1,
\]
\[
N(m,t) = \exp(-rt) \sum_{m'=0}^{\infty} N(m',0)I_{m'}(rt), \quad m = 0.
\]
(29)

The general principle is that the kernel must satisfy the original PDE, must satisfy the no flow boundary at \( m = 0 \), and \( N \) must satisfy the initial state at \( t = 0 \). It is interesting to note that for \( m > 0 \) the kernel in both the continuum and discrete case with a reflecting boundary is just twice the even part of the kernel of the non-reflecting boundary. This is equivalent to the solution without the \( m = 0 \) boundary plus its mirror image reflection added on. With an absorptive boundary it would be twice the odd part, which is the same as the solution without the \( m = 0 \) boundary minus its mirror image reflection. This is a principal that is generally true and has been demonstrated elsewhere for both the diffusion equation (Schulten and Kosztin, 2000) and the discrete random walk problem with a reflective boundary (Chandrasekhar, 1943). Another way to construct the reflective boundary would be to solve the problem with an initial state and with a mirror image of that same initial state on the left-hand axis. Although we have derived our solution from the boundary conditions, one of the more common methods to solve this problem is through the use of virtual mirror images on a virtual left hand axis (Szymczak and Ladd, 2003). Fig. 1 illustrates the use of this technique. Even though the physical solution is only defined along the right axis, one may construct reflecting or absorbing boundaries by analytically continuing that same solution onto the left axis.

The main advantage (29) has over the continuum diffusive model is better accuracy. For large times \( (rt \gg 1) \) they are very similar as shown in Fig. 2. As an example the
discrete counterpart to (28) is
\[ N(m, t) = \exp(-rt)(I_{m-c}(rt) + I_{m+c}(rt)). \] (30)

A plot of the continuous and discrete solution is shown in Fig. 3. This result matches favorably with experimental results, which show a positive skewed curve (Calabrese et al., 2001). Some have also numerically treated an upper bound as a reflecting boundary (Nauta and Weissing, 1996) in order to model maximum microsatellite length. Although this limits the microsatellite length, it is not clear whether that particular model has any basis in physical reality. There is, on the other hand, a real hard limit on the lower end, which cannot be violated. As a result we feel a reflecting boundary on the lower limit is a valid boundary condition.

One could also use this analytic technique to place a second reflecting boundary on the upper bound too. There are a couple of ways this could be done. One could use the method of mirror images using multiple infinite reflections of the initial state. One could also solve the Neumann problem with the boundary condition (19) applied to each reflecting barrier.

6. Comparison of two individuals with a reflecting boundary

Up to this point we have shown how the distribution of marker length of a population evolves over time. If we instead wanted to compare the probability that two individuals are related, let us suppose two individuals have a known common ancestor in time \( t \). The probability of them having a common ancestor after \( t \) generations is
\[ p(x_1, x_2, t) = \exp(-2rt)(I_{x_1-x_2}(2rt) + I_{x_1+x_2}(2rt)). \] (31)

We use 2\( rt \) because each is separated by \( rt \) from the common ancestor. In contrast, without the reflecting boundary the result is
\[ p(x_1, x_2, t) = \exp(-2rt)I_{x_1-x_2}(2rt). \] (32)

In practice, the reflecting boundary has very little effect except in cases where the microsatellite length is small or when the time between common ancestors is very long. Fig. 4 shows a comparison between two individuals with and without a reflecting boundary for \( x_1 = 6 \) and \( x_2 = 5 \).

7. The symmetric \( n \)-step model

It has been shown that for some loci a single-step model may no longer be sufficient (Whittaker et al., 2003) and an \( n \)-step model may be more appropriate. For this model (8) becomes
\[ \frac{\partial}{\partial t} N(m, t) = \frac{1}{2} \sum_{k=1}^{K} r_k (N(m+k, t) - 2N(m, t) + N(m-k, t)), \] (33)

where \( r_k \) are the mutation rates for the \( k \)th step size. Applying the same transformation as in (9) and solving the resulting differential equation we find
\[ \hat{N}(\omega, t) = \hat{N}(\omega, 0) \exp \left[ -2 \sum_{k=1}^{K} r_k \sin^2 \left( \frac{\omega}{2} k \right) t \right], \] (34)

so that
\[ N(m, t) = \sum_{m'=\infty}^{\infty} N(m', 0) G(m-m', t), \] (35)
where

\[
G(m, t) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \exp(-i\omega m) \exp\left[ -2 \sum_{k=1}^{K} r_k \sin^2\left(\frac{\omega}{2} k\right) t \right] d\omega.
\]  

(36)

As in the previous sections, it is understood that \( m \) is only defined for values of \( m \) greater than zero in nature. Therefore, this mathematical solution is only valid as long as the interaction with the \( m = 0 \) boundary is minimal. Should this not be the case, one may easily apply a reflective boundary as we have in the previous section, which is a very simple thing to do. All that is needed is to add the mirror image reflection.

Eq. (34) is an amazingly simple result and can be calculated for any combination of step size. To test the validity of (36) we take the simple case of the two-step model, where

\[
G(m, t) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \exp(-i\omega m) \exp\left[ -2r_1 \sin^2\left(\frac{\omega}{2} t\right) \right] \exp[-2r_2 \sin^2(\omega t)] d\omega.
\]  

(37)

By expanding (37) in a Fourier series and using the convolution theorem for Fourier series (Zemanian, 1965), it can be shown this reduces to

\[
G(m, t) = \exp(-r_1 t) \exp(-r_2 t) \sum_{m'=-\infty}^{\infty} I_{m'}(r_2 t) I_{m-2m'}(r_1 t).
\]  

(38)

This is exactly the result found by Wehrhahn (1975) using probability generating functions.

In a recent paper, Watkins (2007) found a special case of an \( n \)-step solution using matrix methods similar to Zhivotovsky et al. (1997) and taking a limit as the matrix size approaches infinity. In his solution he found the special case where the rate distribution drops off geometrically (inspired by the work of Whittaker et al., 2003) and found an explicit expression for this. We are able to easily derive that result using our methods instead. In our notation this distribution is

\[
r_k = r(1 - q)q^{k-1},
\]  

(39)

where \( q \) is a parameter such that \( 0 \leq q \leq 1 \). The sum in (36) becomes (Dwight, 1961)

\[
-2 \sum_{k=1}^{\infty} r_k \sin^2\left(\frac{\omega}{2} k\right) = -2r \frac{1-q}{q} \sum_{k=1}^{\infty} q^k \sin^2\left(\frac{\omega}{2} k\right)
\]

\[
= r \frac{(1-q)(\cos(\omega) - q)}{1 - 2q \cos(\omega) + q^2} - r.
\]  

(40)

Eq. (36) becomes

\[
G(m, t) = \frac{\exp(-rt)}{2\pi} \int_{-\pi}^{\pi} \exp(-i\omega m) \exp\left[ \frac{(1-q)(\cos(\omega) - q)}{1 - 2q \cos(\omega) + q^2} \right] d\omega.
\]  

(41)

As demonstrated by Watkins (2007), this approaches the single-step model in the limit as \( q \to 0 \) and the infinite alleles model as \( q \to 1 \). However, this is not the only possible model with that behavior. As an alternative to the geometric drop off in the mutation distribution with step size, we propose it could very well be a Gaussian instead, with the center of the curve representing no mutation, then dropping off as a Gaussian for 1 or more steps. That makes more sense if it is a random event. It is a simple matter to calculate this. With

\[
r_k = \frac{2r}{\partial^2_{\bar{z}}(0, q) - 1} q^k.
\]  

(42)
where \( \vartheta_3 \) is the Jacobi Theta function of the third kind (Whittaker and Watson, 1927), the sum in (36) becomes

\[
-2 \sum_{k=1}^{\infty} r_k \sin^2 \left( \frac{\omega}{2} k \right) = -\frac{4r}{\vartheta_3(0,q)} - \sum_{k=1}^{\infty} q^k \sin^2 \left( \frac{\omega}{2} k \right)
\]

\[
= r - \frac{\vartheta_3(0,q)}{\vartheta_3(0,q) - 1} - r.
\]

(43)

In this case (36) becomes

\[
G(m, t) = \frac{\exp(-rt)}{2\pi} \int_{-\pi}^{\pi} \exp(-iom) \times \exp \left[ \frac{\vartheta_3(\omega/2, q) - 1}{\vartheta_3(0, q) - 1} - rt \right] \, d\omega.
\]

(44)

Just as in the previous example, this model also approaches the single-step model in the limit \( q \to 0 \) and the infinite alleles model in the limit \( q \to 1 \).

8. Conclusion

We have demonstrated the utility in using the discrete diffusion approach in solving the SMM model. We feel this approach has many advantages to straight statistical methods. The first is the high degree of flexibility in specifying the type of solution, initial state (if any), and the natural way different types of boundaries are dealt with by the extension to continuum PDEs. Many of the techniques applied to continuum PDEs may be applied here. One of the most powerful of these is using transform methods. Translated into the discrete domain we use Fourier series techniques rather than Fourier transforms. We have demonstrated one may transform a partial difference equation to a first-order differential equation for the characteristic function for the solution using very simple methods. Kimura and others have done much work using the continuum diffusion equation. We feel that by extending his methods to the discrete diffusion equation and using our techniques one may breathe new life into an older method.

Much work remains to be done. This work was mainly intended as an introduction to the technique of using the discrete diffusion equation in relation to STRs. One possible extension of this work would be to work the case of mutation rates that vary with length. Although it is doubtful this problem could be solved analytically in general, it is quite possible specific cases could be solved. The reason for this is because of the similarity of the diffusion equation with the Schrödinger wave equation. It may be possible to cast it in a form where the mutation rate represents a potential of some sort. As a result some of the techniques used to solve problems in quantum mechanics might be applied here.

Acknowledgments

Many thanks go to Bruce Walsh for his helpful comments regarding the use of reflecting barriers in the context of STRs. I also like to thank the reviewers for their many helpful comments and suggestions.

Appendix

A.1. The discrete method

To solve (1) directly we multiply both sides by \( \exp(i\omega t) \) and sum over \( m \) as in the previous example. The result is

\[
\hat{N}(\omega, n + 1) = \left[ 1 - 2r \sin^2 \left( \frac{\omega}{2} \right) \right] \hat{N}(\omega, n).
\]

By recursion we find

\[
\hat{N}(\omega, n) = \left[ 1 - 2r \sin^2 \left( \frac{\omega}{2} \right) \right]^n \hat{N}(\omega, 0).
\]

Although we could integrate this expression in terms of a finite series, since \( r \) is typically very small, we instead use the approximation

\[
\left[ 1 - 2r \sin^2 \left( \frac{\omega}{2} \right) \right]^n \approx \exp \left( -2rn \sin^2 \left( \frac{\omega}{2} \right) \right).
\]

Using the results of (11) and (12) we find using the same techniques as in the previous sections,

\[
N(m, n) = \exp(-rn) \sum_{m' = -\infty}^{\infty} N(m', 0)I_{m-m'}(rn).
\]

(A.1)

References


Walsh, J.B., 2001. Estimating the time to the most recent common ancestor for the Y chromosome or mitochondrial DNA for a pair of individuals. Genetics 158, 897–912.