The SMM Model as a Boundary Value Problem using the Discrete Diffusion Equation

Joel Campbell
NASA Langley Research Center
Hampton Virginia, 23681

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.

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. 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).

A difference equation for the SMM

Let N(m,n) be the number of individuals with a particular STR marker having exactly m repeats at the nth generation. Let us suppose this marker can mutate at a rate r per generation. The probability of N[m,n] mutating to m+1 or m-1 is r/2. If N(m,n) mutates, there is an equal loss of the unmutated state. This suggests a difference equation of the following form

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

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

$$\begin{gathered} N(m,n+1) = N(m,n) + \frac{r}{2}\big(N(m+1,n) - 2N(m,n) + N(m-1,n)\big) + \\ \frac{r-r'}{2}\big(N(m,n) - N(m-1,n)\big) \end{gathered}, \tag{2}$$

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.

The SMM in the continuum

In the continuum limit this becomes

$$\frac{\partial}{\partial t}N(x,t) = \frac{r}{2}\frac{\partial^2}{\partial x^2}N(x,t). \tag{3}$$

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

$$N(x,t) = \frac{1}{\sqrt{2\pi rt}}\exp\left[-\frac{(x-x_0)^2}{2rt}\right]. \tag{4}$$

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 (ZHIVOTOVSKY 1995) using the standard deviation as the independent variable. The variance for (4) is given by,

$$\sigma^2 = rt. \tag{5}$$

This is a well known result for estimating age (GOLDSTEIN, 1995). A more general solution to (3) is (COURANT and HILBERT 1962)

$$N(x,t) = \frac{1}{\sqrt{2\pi rt}} \int_{-\infty}^{\infty} N(x',0) \exp\left[-\frac{(x-x')^2}{2rt} \right] dx'. \tag{6}$$

With an initial state of

$$N(x,0) = \delta(x - x_0), \tag{7}$$

(4) is recovered.

The interpretation of (6) is that given an initial distribution, N(x,0), which may or may not be normalized, how does that initial distribution evolve over time?

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

$$\frac{\partial}{\partial t} N(m,t) = \frac{r}{2}\left(N(m+1,t) - 2N(m,t) + N(m-1,t) \right). \tag{8}$$

It can be shown a solution to (8) is

$$N(m,t) = \exp[-rt] I_{m-m_0}(rt), \tag{9}$$

where I is the modified Bessel function of the first kind (ABRAMOWITZ and STEGUN). This is a result that has been found using statistical methods with the SMM model (WALSH 2001). A more general solution to (9) may be found using Fourier series techniques. We first multiply both sides by exp[I mω] and sum over m. The result is

$$\frac{\partial}{\partial t} \hat{N}(\omega,t) = \frac{r}{2}\left(\exp(i\omega)\hat{N}(\omega,t) - 2\hat{N}(\omega,t) + \exp(-i\omega)\hat{N}(\omega,t) \right) = -2r \sin^2\left(\frac{\omega}{2}\right) \hat{N}(\omega,t), \tag{10}$$

where

$$\hat{N}(\omega,t) = \sum_{m=-\infty}^{\infty} \exp(i\omega m) N(m,t) \tag{11}$$

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

$$N(m,t)=\frac{1}{2\pi}\int_{-\pi}^{\pi}\exp(-i\omega m)\hat{N}(\omega,t)d\omega. \tag{12}$$

Equation (10) is easily solved and has the solution

$$\hat{N}(\omega,t)=\hat{N}(\omega,0)\exp\left[-2r\sin^2\left(\frac{\omega}{2}\right)t\right]. \tag{13}$$

Using the results of (11), (12), and (13) we find

$$N(m,t)=\frac{1}{2\pi}\sum_{m'=-\infty}^{\infty}N(m',0)\int_{-\pi}^{\pi}\exp\left[-i\omega(m'-m)\right]\exp\left[-2r\sin^2\left(\frac{\omega}{2}\right)t\right]d\omega. \tag{14}$$

Using the identity $\sin^2(x)=(1-\cos(2x))/2$ and ABROMOWITZ and STEGUN we find

$$\frac{1}{2\pi}\int_{-\pi}^{\pi}\exp(-iwm)\exp\left[-2r\sin^2\left(\frac{\omega}{2}\right)t\right]d\omega=\exp(-rt)I_m(rt). \tag{15}$$

We have then

$$N(m,t)=\exp[-rt]\sum_{m'=-\infty}^{\infty}N(m',0)I_{m-m'}(rt). \tag{16}$$

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.

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 generating function.

The discrete method

To solve (2) directly we multiply both sides by exp[I mω] and sum over m as in the previous example. The result is

$$\hat{N}(\omega,n+1)=\left[1-2r\sin^2\left(\frac{\omega}{2}\right)\right]\hat{N}(\omega,n). \tag{17}$$

By recursion we find

$$\hat{N}(\omega,n) = \left[1 - 2r\sin^2\left(\frac{\omega}{2}\right)\right]^n \hat{N}(\omega,0). \tag{18}$$

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

$$\left[1 - 2r\sin^2\left(\frac{\omega}{2}\right)\right]^n \approx Exp\left(-2rn\sin^2\left(\frac{\omega}{2}\right)\right). \tag{19}$$

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

$$N(m,n) = \exp[-rn] \sum_{m'=-\infty}^{\infty} N(m',0) I_{m-m'}(rn). \tag{20}$$

Reflecting boundary

As stated earlier the microsatellite 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. What this does is create a special reflecting boundary for our PDEs. Let us first consider the continuous case. Our boundary condition is then

$$N(x,t) = N(-x,t). \tag{21}$$

The reasoning being that if we force the solution to be symmetric there can be no net flow across the x=0 boundary. This suggests a solution of the form,

$$N(x,t) = \frac{2}{\pi}\int_0^\infty a(\omega,t)\cos(\omega x)d\omega, \tag{22}$$

where

$$a(\omega,t) = \int_0^\infty N(x,t)\cos(\omega x)dx. \tag{23}$$

We first use (23) and substitute it in to (4). The result is,

$$\frac{\partial}{\partial t} a(\omega,t) = -\frac{r}{2}\omega^2. \tag{24}$$

The solution is,

$$a(\omega,t) = a(\omega,0)\exp\left(-\frac{r}{2}\omega^2 t\right), \tag{25}$$

so that

$$N(x,t) = \frac{2}{\pi}\int_0^\infty a(\omega,0)\exp\left(-\frac{r}{2}\omega^2 t\right)\cos(\omega x)d\omega. \tag{26}$$

This can be cast into the form,

$$N(x,t) = \int_0^\infty N(x',0)G(x,x',t)dx' \tag{27}$$

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

$$\begin{gathered} G(x,y,t) = \frac{2}{\pi}\int_0^\infty \exp\left(-\frac{r}{2}\omega^2 t\right)\cos(\omega y)\cos(\omega x)d\omega = \\ \frac{1}{\sqrt{2\pi rt}}\left[\exp\left(-\frac{(x-y)^2}{2rt}\right) + \exp\left(-\frac{(x+y)^2}{2rt}\right)\right] \end{gathered} . \tag{28}$$

We now have

$$N(x,t) = \frac{1}{\sqrt{2\pi rt}}\int_0^\infty N(x',0)\left[\exp\left(-\frac{(x-x')^2}{2rt}\right) + \exp\left(-\frac{(x+x')^2}{2rt}\right)\right]dx'. \tag{29}$$

As an example we choose an initial state N(x,0)=$\delta(x-c)$ so that

$$N(x,t) = \frac{1}{\sqrt{2\pi rt}}\left[\exp\left(-\frac{(x-c)^2}{2rt}\right) + \exp\left(-\frac{(x+c)^2}{2rt}\right)\right] \tag{30}$$

The discrete version can be solved in a similar manner. The result is,

$$\begin{gathered} N(m,t) = \exp(-rt)\sum_{m'=0}^{\infty} N(m',0)\left(I_{m-m'}(rt) + I_{m+m'}(rt)\right), m \geq 1 \\ N(m,t) = \exp(-rt)\sum_{m'=0}^{\infty} N(m',0)I_{m'}(rt), m = 0 \end{gathered} . \tag{31}$$

The general principal is that the kernal must satisfy the original PDE, must be symmetric about m=0, and N must satisfy the initial state at n=0. It is interesting to note that for m>0 the kernal in both the continuum and discrete case with a reflecting boundary is just twice the even part of the kernal of the non-reflecting boundary. With an absorptive boundary it would be twice the odd part.

The main advantage (31) has over the continuum diffusive model is better accuracy. For large times (r t>>1) they are very similar. As an example the discrete counterpart to (30) is

$$N(m,t) = \exp(-rt)\left(I_{m-c}(rt) + I_{m+c}(rt)\right). \tag{32}$$

A plot of the continuous and discrete solution is shown in Figure 1. This result matches favorably with experimental results, which show a positive skewed curve (CALABRESE et. al., 2001). Some have also 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.

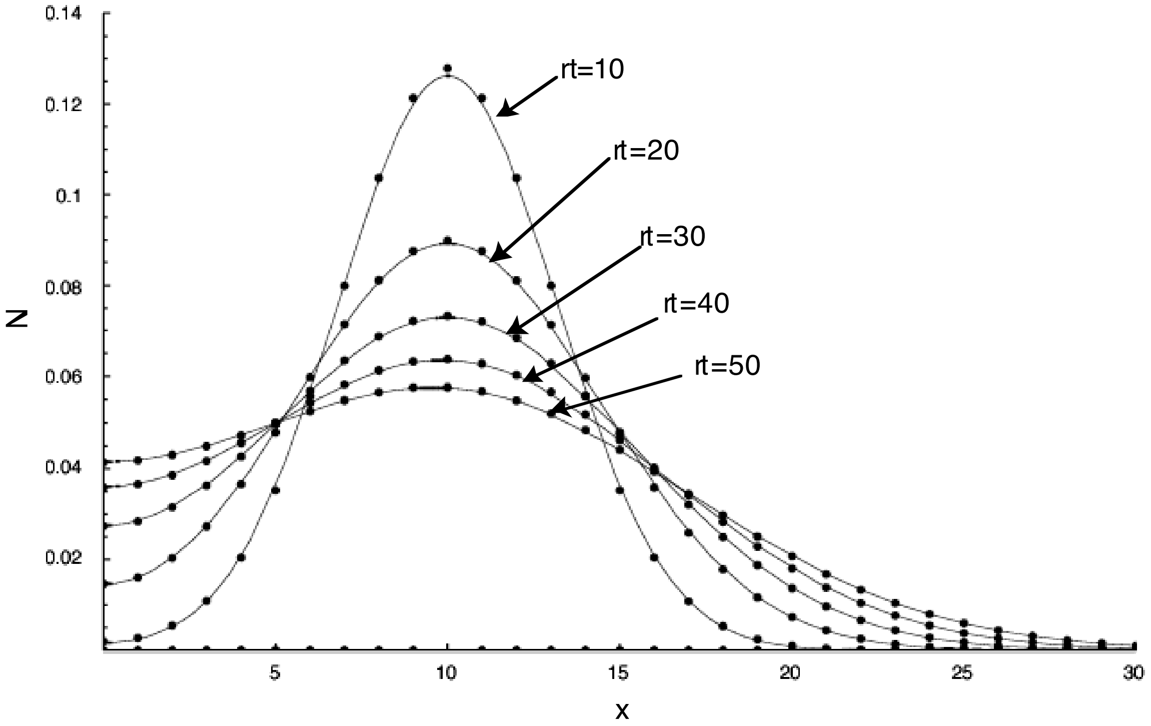

Figure 1. A comparison of continuous and discrete solution for various large times and c=10 with a reflective boundary shows that the continuous and discrete solutions are very similar. The dots represent the discrete solution.

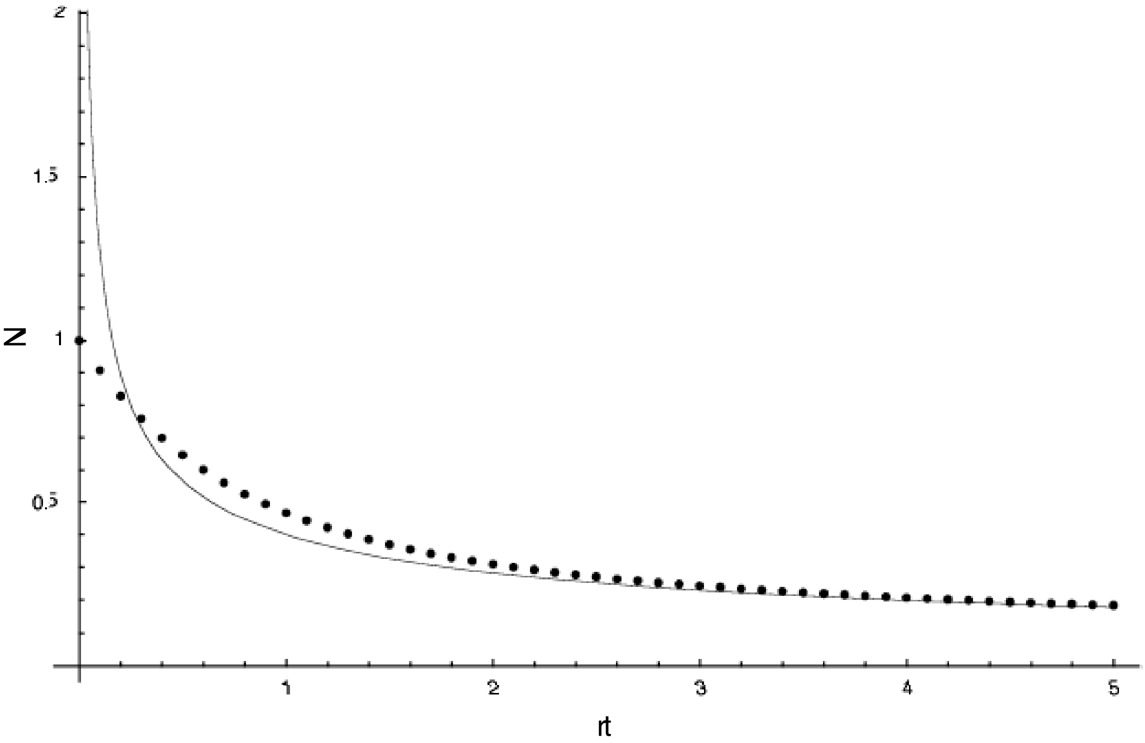

Figure 2. A comparison of the continuous and discrete solution for small times and c=x=10 with a reflective boundary shows that the continuous and discrete solutions diverge for small times.

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

$$p(x_1, x_2, t) = \exp(-2rt)\left(I_{x_1 - x_2}(2rt) + I_{x_1 + x_2}(2rt)\right). \tag{33}$$

We use 2rt because each is separated by rt from the common ancestor. In contrast, without the reflecting boundary the result is,

$$p(x_1, x_2, t) = \exp(-2rt) I_{x_1 - x_2}(2rt) \tag{34}$$

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. Figure 3 shows a comparison between two individuals with and without a reflecting boundary for $x_1=6$ and $x_2=5$.

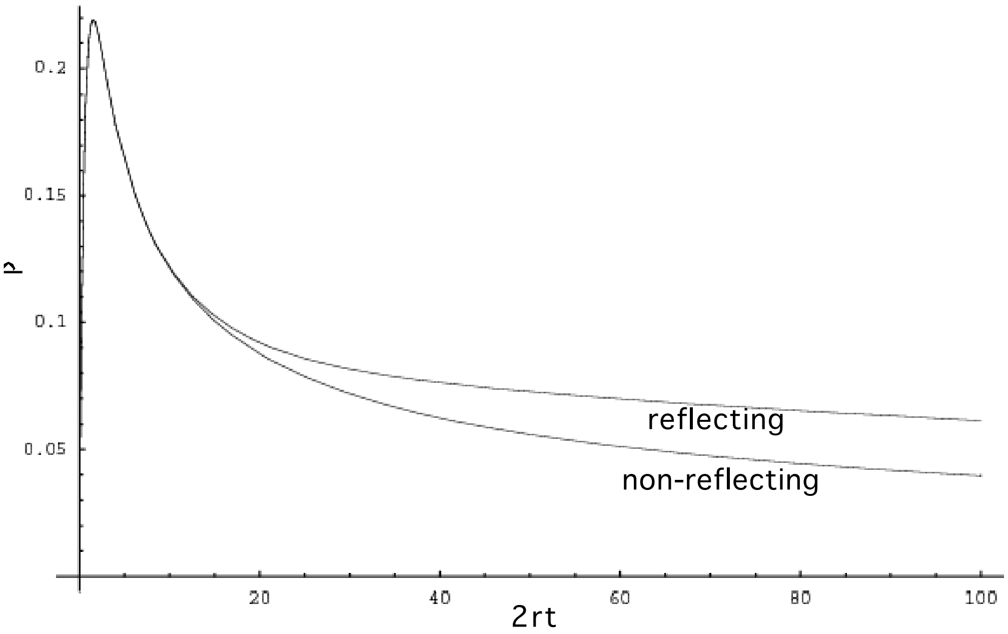

Figure 3. Comparison of the probability two individuals are related at time t for $x_1=6$ and $x_2=5$ for the reflecting and non-reflecting boundary cases.

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

$$\frac{\partial}{\partial t}N(m,t)=\frac{1}{2}\sum_{k=1}^{K}r_k\left(N(m+k,t)-2N(m,t)+N(m-k,t)\right), \tag{35}$$

where r$_k$ are the mutation rates for the kth step size. Applying the same transformation as in (9) and solving the resulting differential equation we find

$$\hat{N}(\omega,t) = \hat{N}(\omega,0)\exp\left[-2\sum_{k=1}^{K} r_k \sin^2\left(\frac{\omega}{2}k\right)t\right], \tag{36}$$

so that

$$N(m,t) = \sum_{m'=-\infty}^{\infty} N(m',0)G(m-m',t), \tag{37}$$

where

$$G(m,t) = \frac{1}{2\pi}\int_{-\pi}^{\pi}\exp(-i\omega m)\exp\left[-2\sum_{k=1}^{K} r_k \sin^2\left(\frac{\omega}{2}k\right)t\right]d\omega. \tag{38}$$

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.

Equation (38) is an amazingly simple result and can be calculated for any combination of step size. To test the validity of (38) we take the simple case of the two-step model, where

$$G(m,t) = \frac{1}{2\pi}\int_{-\pi}^{\pi}\exp(-i\omega m)\exp\left[-2r_1\sin^2\left(\frac{\omega}{2}\right)t\right]\exp\left[-2r_2\sin^2(\omega)t\right]d\omega. \tag{39}$$

By expanding (39) in a Fourier series and using the convolution theorem for Fourier series (ZEMANIAN, 1965), it can be shown this reduces to

$$G(m,t) = \exp(-r_1 t)\exp(-r_2 t)\sum_{m'=-\infty}^{\infty} I_{m'}(r_2 t) I_{m-2m'}(r_1 t). \tag{40}$$

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 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

$$r_k = r(1-q)q^{k-1}, \tag{41}$$

where q is a parameter such that $0 \le q \le 1$. The sum in (38) becomes (DWIGHT 1961)

$$-2\sum_{k=1}^{\infty} r_k \sin^2\left(\frac{\omega}{2}k\right) = -2r\frac{1-q}{q}\sum_{k=1}^{\infty} q^k \sin^2\left(\frac{\omega}{2}k\right) = r\frac{(1-q)(\cos(\omega)-q)}{1-2q\cos(\omega)+q^2} - r. \qquad (42)$$

Equation (38) becomes

$$G(m,t) = \frac{\exp(-rt)}{2\pi}\int_{-\pi}^{\pi} \exp(-i\omega m)\exp\left[\frac{(1-q)(\cos(\omega)-q)}{1-2q\cos(\omega)+q^2} rt\right] d\omega. \qquad (43)$$

As demonstrated by WATKINS, 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

$$r_k = \frac{2r}{\vartheta_3(0,q)-1} q^{k^2}, \qquad (44)$$

where ϑ_3 is the elliptic Theta function of the third kind (WHITTAKER and WATSON 1927), the sum in (38) becomes

$$-2\sum_{k=1}^{\infty} r_k \sin^2\left(\frac{\omega}{2}k\right) = -\frac{4r}{\vartheta_3(0,q)-1}\sum_{k=1}^{\infty} q^{k^2} \sin^2\left(\frac{\omega}{2}k\right) = r\frac{\vartheta_3\left(\frac{\omega}{2},q\right)-1}{\vartheta_3(0,q)-1} - r. \qquad (45)$$

In this case (38) becomes

$$G(m,t) = \frac{\exp(-rt)}{2\pi}\int_{-\pi}^{\pi} \exp(-i\omega m)\exp\left[\frac{\vartheta_3\left(\frac{\omega}{2},q\right)-1}{\vartheta_3(0,q)-1} rt\right] d\omega \qquad (46)$$

Just as in the previous example, this model also approaches the single step model in the limit $q \to 0$ and the infinite allele model in the limit $q \to 1$.

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 generating 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 breath new life into an older method.

Acknowledgements

Many thanks go to Bruce Walsh for his helpful comments.

References

Abramowitz, M. and I. A Stegun, 1972 Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables, Dover.

Calabrese, Peter P., Richard T. Durrett and Charles F. Aquadro, 2001 Dynamics of Microsatellite Divergence Under Stepwise Mutation and Proportional Slippage/Point Mutation Models. Genetics 159, pp. 839–852

Courant, R. and D. Hilbert, 1962 Methods of Mathematical Physics, Vol II, pp. 198

Dwight, H. 1961. Table of integrals and other mathematical data. MacMillan Publishing.

Fath, Gabor 1998 Propagation failure of traveling waves in a discrete bistable medium. Physica D 116 pp. 176-190

Goldstein, D. 1995. An Evaluation of Genetic Distances for Use With Microsatellite Loci. Genetics 139, pp. 463-471

Gradshteyn, I. and I. Ryzhik, 1980 Table of integrals, series, and products. Academic Press.

Lindeberg, Tony 1990 Scale-Space for Discrete Signals. IEEE Transactions of Pattern Analysis and Machine Intelligence, 12(3), pp. 234-254

Lu, Zhengyi and Yasuhiro Takeuchi, 1993 Global asymptotic behavior in single-species discrete diffusion systems. Journal of Mathematical Biology Vol 32, pp. 67-77

Nauta, Maarten J. and Franz J. Weissing, 1996 Constraints on Allele Size at Microsatellite Loci: Implications for Genetic Differentiation. Genetics 143 pp. 1021-1032

Ohta, T., and M. Kimura, 1973 A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a genetic population. Genet. Res. **22:** 201–204.

Reif, F., 1965. Fundementals of statistical and thermal physics. McGraw-Hill.

Wehrhahn, C. F., 1975 The evolution of selectively similar electrophoretically detectable alleles in finite natural populations. Genetics 80 pp. 375–394.

Watkins, Joseph C. 2007 Microsatellite Evolution: Markov Transition Functions for a Suite of Models. Theoretical Population Biology 71

Whittaker, John C., Roger M. Harbord, Nicola Boxall, Ian Mackay, Gary Dawson and Richard M. Sibly, 2003 Likelihood-Based Estimation of Microsatellite Mutation Rates. Genetics 164, pp. 781–78

Whittaker and Watson, 1927 (reprint 1992) A course of modern analysis. Cambridge University Press, pp. 462

Zemanian, A, 1965 Distribution theory and transform analysis. Dover Publications.

Zhivotovsky, Lev A. and Marcus W. Feldman, 1995 Microsatellite variability and genetic distances. Proc. Natl. Acad. Sci. USA, Vol. 92, pp. 11549-11552, December 1995

Zhivotovsky, Lev A., Marcus W. Feldman, and Sergei A. Grishechkint, 1997 Biased Mutations and Microsatellite Variation. Molecular Biology and Evolution, Vol 14, 926-933