Mutation models III: "exotic" models

Peter Beerli

September 15, 2009

Most of these "exotic" models are standard models in population genetics, where they dominate the field and finite-sites mutation models, such as HKY or GTR, are only recently used in the context of parameter inference using coalescence theory.

1 Infinitely many allele model

This model was developed as an extension of the k-allele model, where $k = 1, 2, , 3.....$ The k allele model allows back-mutations, so that A can mutate to a and also from a to A (see details of this model for $k = 2$ in the chapter about *Mutation models I: basic nucleotide sequence mutation models*). The infinitely many allele model, or infinite allele model (IAM) for short, sets $k = \infty$ and so de facto does not allow any back-mutation: seeing two individuals with an A means that they must have a recent common ancestor and no mutation event happened since then. New mutations will results in a new allele. The rate matrix for the transitions between alleles A_i and A_j could be expressed as

$$R = \begin{pmatrix} 1-\mu & \frac{\mu}{k} & .... \\ .... & & \\ \frac{\mu}{k} & .... & 1-\mu \end{pmatrix} \tag{1}$$

In the program MIGRATE, I used a k allele approximation to the infinite allele model, that tries to lump all unobserved states into an additional $k+1$ class.

$$\text{prob}(A_i|\mu, t, A_i) = f_{A_i} e^{-\mu t}$$
$$\text{prob}(A_i|\mu, t, A_j) = f_{A_j} e^{-\mu t}$$
$$f_{A_z} = \begin{cases} \frac{1}{k+1} & \text{if } A_z \text{ in sample,} \\ 1 - \frac{k}{k+1} & \text{if } A_z \text{ not in sample} \end{cases}$$

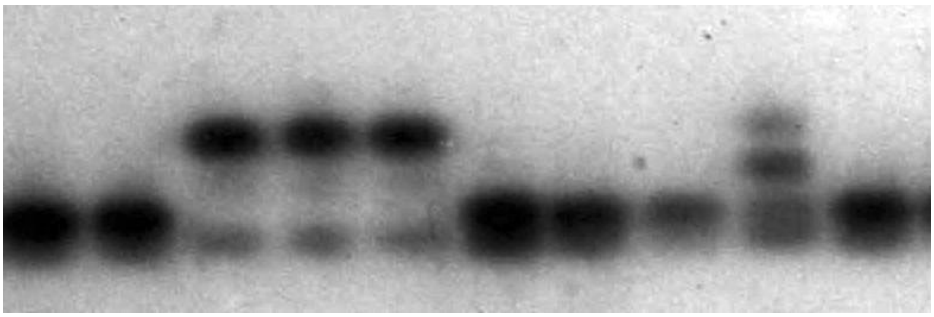

Figure 1: Example of an electrophoretic marker: malate dehydrogenase (sMDH[1]). In an electric field the enzyme moves according to the electric properties of the aminoacids. different electromorphs (alleles) show different mobility. Ten individuals were scored, from left to right: a/a, a/a, c/c, c/c, c/c, a/a, a/a, a/a, a/c, a/a. The heterozygote a/c show three bands because sMDH is a dimer and a heterozygote individuals shows functional enzyme combinations of a/a, a/c, and c/c, in proportions of 1:2:1. The bands in the position slightly lower than the a-allele belong to sMDH that is monoporphic for these 10 individuals. ^1s stands for soluble in the cytosol and m stands for located in mitochondria.

This model was developed before genetic data could be analyzed and researchers started to use it to analyze protein electrophoresis results (Figure 1). Many results in population genetics were derived using the IAM and it is probably still the most often used model.

2 Infinitely many sites mutation model (ISM)

The idea of the IAM was extended to sequences once they became available. When used with sequences IAM would not take into account multiple mutation on a sequence. We assume that a sequence of nucleotides (or aminoacids) is infinitely long and a mutation occurs at a random site, because the sequence is infinitely long every mutation happens at a different site. This mutation model is commonly used, for example haplotype networks where mutation separate haplotypes. Figure 2 shows the similarities between the two mutation model on a tree.

3 Mutation and Substitution revisited

What is the differences between mutation rate and substitution rate? We explore this using theoretical population genetics arguments. In an ideal population of diploid organisms, for example a

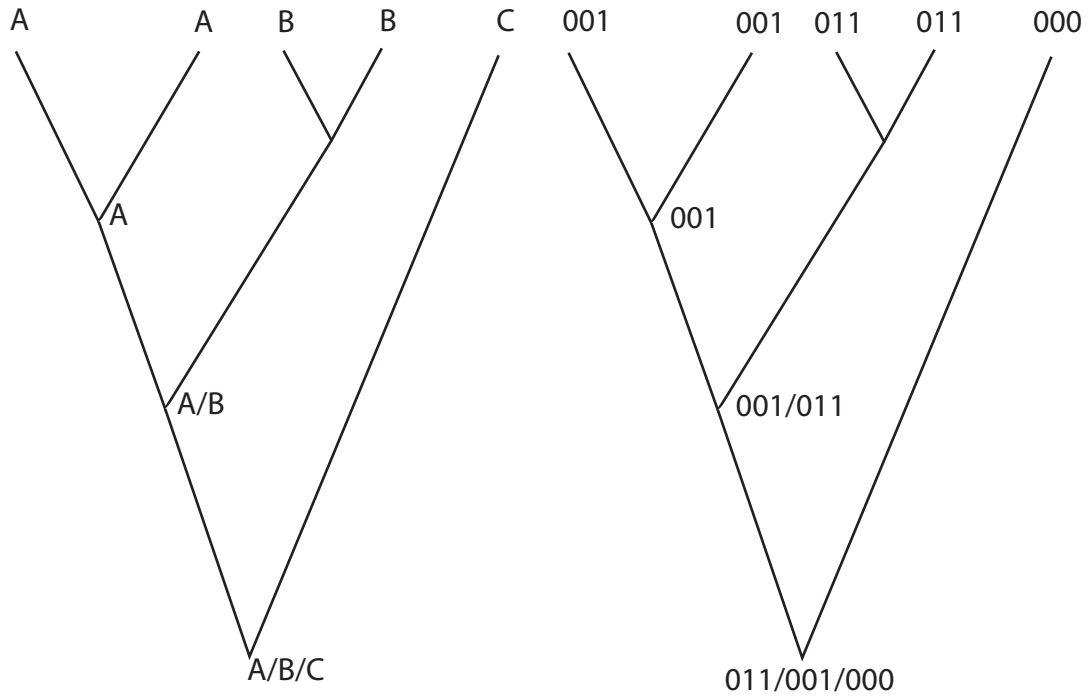

Figure 2: Comparison of the infinite allele and the infinite sites mutation model

Wright-Fisher population, there is a constant number N individuals with $2N$ gene copies. A point mutation is a change of a single nucleotide in the DNA. It happens with a rate μ (for example $\mu = 10^{-9}$ per site per generation). We expect on average $2N\mu$ mutants in our populations, if we assume that all mutations are neutral. Tracing a single mutation and ignoring all future mutation we can calculate the time to fixation of that mutation (all gene copies are descendants of that single mutation event). Each gene copy has the same chance of survival to the next generation if the system is neutral. This gives our mutant copy a chance of $1/(2N)$ to be the descendant of all gene copies in the future. The expected number mutants that arise in the current generation and substitute throughout the whole population is $2N\mu \times 1/(2N) = \mu$. So, assuming that all mutations are neither detrimental or beneficial the substitution rate and the mutation rate are the same. The effects of the population size balances the increased number of mutants with the increased chance of loosing rare alleles. Measures of rates of change based on pedigrees and based on dates of species divergence often show estimates that are 10 to 20 fold different. Most mutations are obviously not neutral and have a lower chance than $1/(2N)$ of fixation. Deleterious mutations get reduced more quickly in a large population and so we expect that the substitution rate is lower than the mutation rate. Kimura and Otha's "nearly neutral theory" might need some adjustment to accommodate the discrepancy between substitution rate and mutation rate because recent studies cannot explain the differences between empirical rates of change evaluated from pedigrees and phylogenies.

4 Microsatellite mutation models

Microsatellites are an accumulation of short repeats, typically 2 to 5 base pairs long, once could consider microsatellite as a special class of simple sequence repeats (SSR) that include repeats of 1 to several nucleotides. This marker type is often considered to be neutral because only very microsatellite loci occur in coding sequences, for example Huntington's disease repeats CAG (coding for the aminoacid glutamine). But some microsatellites can even be harmful in non-coding region, for example "fragile X" that can add under some conditions several thousand repeats, making the the X chromosome somewhat less functional.

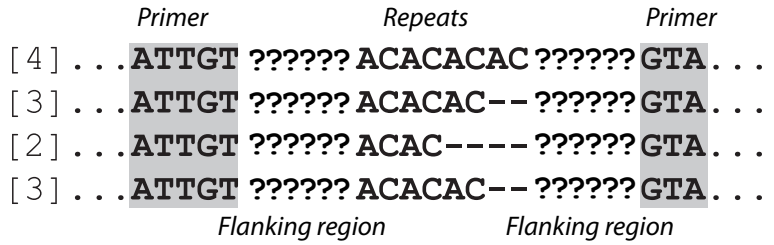

Figure 3: Example of a microsatellite locus

It is believed that the copying of a microsatellite is more error-prone than copying random nucleotide sequences because of the similarity along the repeated sequence. A most simple mutation model would allow for slipping and so adding or subtracting repeat units.

4.1 Stepwise mutation model

Kimura and Otha developed in the days of enzyme electrophoresis a model that would take into account that we cannot really know whether a band on the gel (figure 1) is the same allele or not, they also assumed that a single aminoacid change will move the band one unit up or down. Their *ladder model* is now synonym with the stepwise mutation model. It is assumed that a single mutation event is adding or subtracting a single repeat. Of course, on a long branch more than on repeat difference can arise but the model allows only single steps, in stark contrast to the infinite allele model. The transition probability under the exact one-step model can be computed as an approximation of a randomized random walk (Feller 1966). Consider a single gene that is descended from an earlier one t generations ago. The number of mutations that have occurred in this gene will be a Poisson variable with expectation $u = \mu t$. The net increase of i copies can occur by having $i+k$ mutations that increased the copy number, and k mutations that decreased it. Thus the transition probability is the sum over all values of k of the probability of having $i + 2k$ mutations, times the

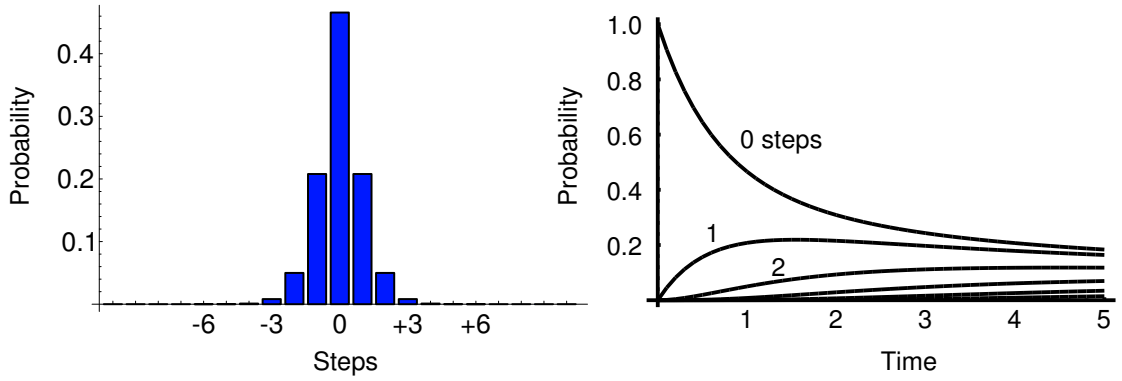

Figure 4: Probability distribution of the stepwise mutation model, on the left side: probability of seeing i steps for a fixed branch length. Right side: probability of a given branch length *time* given fixed step sizes.

binomial probability that, out of $i+2k$ mutations, $i+k$ of them increased the copy number. Then, $P_i(t,\mu)$ is the probability of a net change of i copies after t generations with mutation rate μ,

$$\text{Prob}(i \text{ steps}|t,\mu) = \sum_{k=0}^{\infty} e^{-\mu t} \frac{(\mu t)^{i+2k}}{(i+2k)!} \binom{i+2k}{i+k} \left(\frac{1}{2}\right)^{i+2k}.$$

We can calculate the probability to see i steps in time interval u as

$$\text{Prob}(i \text{ steps}|u) = \sum_{k=0}^{\infty} \frac{e^{u}(u/2)^{i+2k}}{(i+k)!k!} \tag{2}$$

with $u = \mu t$ where t is the time in generations and μ is the mutation rate per generation. This is a convolution of exponential waiting times with Poisson distributed events that we have k steps to reach i achieved steps, for example a single step during time u can be achieved by 1 step from 12 repeats to 13 repeats or by 3 steps from 12 to 13 to 14 to 13, or by stepping from 12 to 11 to 12 to 13, etc. Figure 3 shows probabilities for a given branch length for several steps and how probable some branch length is for a given step. In this model all transitions between all repeats have the same weight, the transition probability matrix P can get very large because we need to take into account all possible transitions from repeat i to repeat j and looks like this:

$$P(u) = \begin{pmatrix} \text{Prob}(0|u) & \text{Prob}(1|u) & \text{Prob}(2|u) & .... \\ \text{Prob}(1|u) & \text{Prob}(0|u) & \text{Prob}(1|u), ... & \\ \text{Prob}(2|u) & \text{Prob}(1|u) & \text{Prob}(0|u), ... & \\ ..... & & & \end{pmatrix}.$$

This P matrix is data dependent and so running tree-based programs using the stepwise mutation model can be very difficult on slow computers. Some datasets of fish populations have have repeat

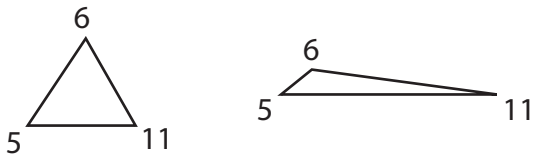

Figure 5: Comparison of the k-allele and the stepwise mutation model: On the left, the 3 alleles (labeled 5, 6, and 11) have all the same mutaitonal distance from each other, on the right the stepwise model suggest that the alleles with 5 and 6 repeats are more closely related to each other than to the allele "11 repeats".

ranges over more than 100 repeats resulting in large P matrix of 100×100 that need to be evaluated for every branch length change. Computational speedups are possible by ignoring large differences of repeats and setting the probabilities for those to zero, for example $\text{Prob}(10|u) = 0.0$ instead of a tiny value, such as 10^-20. this reduces the work essentially to the calculation of a band around the diagonal.

4.2 Brownian motion mutation model

The calculation of the transition probability matrix for the stepwise mutation model is such a burden that the program MIGRATE incorporated a model that is based on the concept of Brownian motion, where a process moves randomly one step at a time one repeat up or down. This model was formally described by Blum et al. (2004) but the outline below follows Felsenstein's description for his progrma `contml`.

The Brownian motion approximation (BMA) replaces the transition probability based on the stepwise by a normal probability density. The expectation is 0, and the variance is chosen to match the one-step model. As that model expects u changes, each of which contributes a variance of 1, the variance is taken to be u, so that the density at i is

$$f(i;u) = \frac{1}{\sqrt{2\pi u}} \exp\left(-\frac{i^2}{2u}\right) \tag{3}$$

To convert this to a probability we multiply by the width of the interval, which is 1. For the case when $i = 0$ the resulting probability can be greater than 1, so we do not allow it to rise above 1:

$$\text{Prob}\ (i|u) = \min[1, f(i;u)]. \tag{4}$$

Figure 6 shows the comparison between the transition probabilities under the SSM and the BMM.

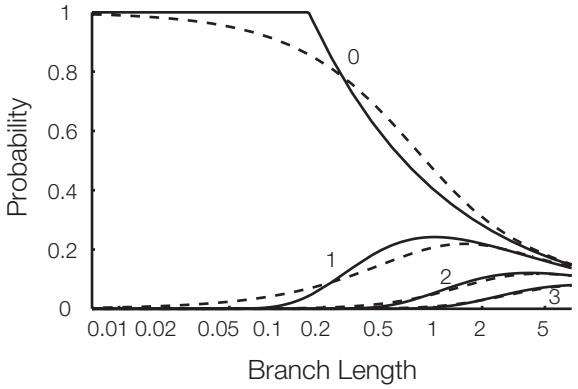

Figure 6: Comparison of the Brownian motion approximation and the stepwise mutation model. Single step mutation model (dashed curves) and Brownian motion approximation (solid curves) for various values of u and for $i = 0, 1, 2, 3$.

4.3 Co-dominant versus dominant markers

We assumed that we can see all the genetic variability in the sample. this is only possible if we are able to see the heterozygote state in diploids or polyploids. For several data types this is not possible, such as restriction length polymorphisms (RFLP) and amplified fragment length polymorphisms (AFLP) or morphological data (these will be considered in the next chapter). One of the alleles is dominant and hiding the other allele in heterzygotes. In RFLP and AFLP this is not really caused by a genetic dominance but by the scoring technique.

4.4 RFLP – Restriction fragment length polymorphism

Restriction enzymes cut DNA at particular locations, for example the restriction enzyme of the bacterium *Hemophilus aegypticus*, named HaeIII, cuts DNA wherever it encounters the sequence

```
5'GGCC3'
3'CCGG5'
```

and cuts it between the `GC` nucleotides. Once cut the DNA-fragments are run on a gel in an electric field and are separated by size (Figure 7). This technique is now rarely used to generate data for population genetics analyses, but still is the base of AFLPs and procedures to reduce the size of complete sequences of DNA to more workable chunks.

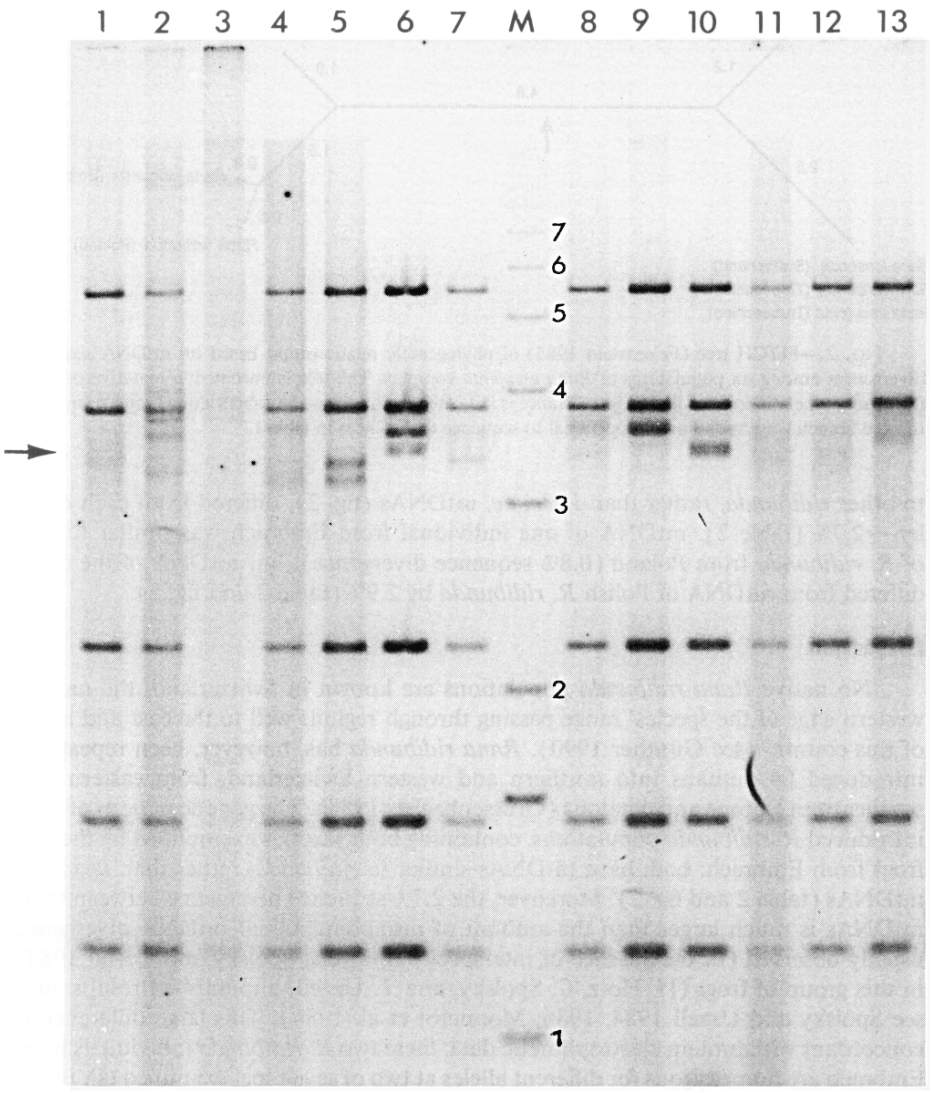

FIG. 3.—Autoradiogram of *Hin*dIII restriction-fragment patterns for mtDNAs of water frogs from northern Switzerland. Fragments were separated on a 0.7% agarose gel. Fragment lengths of the size marker (M; 1-kb ladder) are given in kilobases. The arrow indicates the region of the length-variable fragment. Lane 1, *Rana lessonae* from Frauenfeld. Lanes 2–4, *R. ridibunda* from Trubeschloo. Lanes 5–11, *R. esculenta* from Trubeschloo. Lane 12, *R. lessonae* from Frauenfeld. Lane 13, *R. lessonae* from Poznań. In this gel, much of material in lane 3 remained at the origin; the same six bands as in all other lanes were apparent, however, on the autoradiogram.

Figure 7: Example of an RFLP digest.

Smouse and Li(1987), Felsenstein (1992) recognized that this data type is different to the other ones we mentioned, because it is only used when investigators find differences in their data sets. One needs to take into account that if a point mutation happens in the recognition sequence then one will not see the band on the gel. We correct for not seeing mutations in the recognition sequence.

Such biases occur in almost any DNA data type. We use primers to fish for single copy genes, if the primer is not attaching to the DNA we miss the locus and might not use it for our investigation. More serious is the problem for the detection single nucleotide polymorphisms and microsatellites, where we often do not consider monopmorphic loci for our study – a mistake of not taking into account that we are selecting variable markers. Correcting for such a bias is not yet common in the standard sequence models and microsatellite models.

4.5 RAPD – Random amplified polymorphism data

Using two random primers for PCR reactions. It is expected that the primers are close to each other and produce the fragment between the two primers in enough copies to sequence it. this produces often non-reproducable results because the PCR conditions are rarely identical. Often used for paternity analyses but no great mutation model available.

4.6 AFLP – Amplified fragment length polymorphisms

AFLP is a more sophisticated variants of RAPDs that allows us to get consistent results. First one cuts the DNA with two restriction enzymes, a frequent cutter (MseI, 4 bp recognition sequence) and a rare cutter (EcoRI, 6 bp recognition sequence). Second, a preselective PCR amplification using primers complementary to each of the two adaptor sequences, except for the presence of one additional base at the 3' end. Which base is chosen by the user. Amplification of only 1/16th of EcoRI-MseI fragments occurs. Third, In a second, "selective", PCR, using the products of the first as template, primers containing two further additional bases, chosen by the user, are used. The EcoRI-adaptor specific primer used bears a label (fluorescent or radioactive). (from http://opbs.okstate.edu/ melcher/MG/MGW1/MG11128.html)

Models to analyze AFLP data are sparse, typically people resort to assume hardy-Weinberg proportions and estimate the allele frequencies of the two alleles (band present or absent) using the NULL allele, assuming that that missing band is a homozygote A_2A_2, that does not contain the fragment or does not had the correct recognition sequence. The bands will be a mixture of A_1A_1

and A_1A_2. It seems that under some condition some reseachers are able to show that the band of a heterozygote is fainter than the homozygote A_1 band and so can score a co-dominant marker. without knowing the heterozygotes and without assuming Hardy-Weinberg equilibrium it is rather difficult the generate a useful mutation model, because the current present absent data is rather uninformative for on a per locus base. Despite this AFLPs are wildly successful in genetic trait mapping and paternity analyses.

4.7 Literature

Blum, M. G. B., C. Damerval, et al. (2004). Brownian models and coalescent structures. Theoretical Population Biology 65(3): 249-261. Brownian motions on coalescent structures have a biological relevance, either as an approximation of the stepwise mutation model for microsatellites, or as a model of spatial evolution considering the locations of individuals at successive generations. We discuss estimation procedures for the dispersal parameter of a Brownian motion defined on coalescent trees. First, we consider the mean square distance unbiased estimator and compute its variance. In a second approach, we introduce a phylogenetic estimator. Given the UPGMA topology, the likelihood of the parameter is computed thanks to a new dynamical programming method. By a proper correction, an unbiased estimator is derived from the pseudomaximum of the likelihood. The last approach consists of computing the likelihood by a Markov chain Monte Carlo sampling method. In the one-dimensional Brownian motion, this method seems less reliable than pseudomaximum-likelihood.