J. Math. Biol. (2006) 53:821–841
DOI 10.1007/s00285-006-0031-0

Mathematical Biology

Consistency of estimators of population scaled parameters using composite likelihood

Carsten Wiuf

Received: 4 October 2005 / Revised: 17 July 2006 /
Published online: 8 September 2006

Abstract Composite likelihood methods have become very popular for the analysis of large-scale genomic data sets because of the computational intractability of the basic coalescent process and its generalizations: It is virtually impossible to calculate the likelihood of an observed data set spanning a large chromosomal region without using approximate or heuristic methods. Composite likelihood methods are approximate methods and, in the present article, assume the likelihood is written as a product of likelihoods, one for each of a number of smaller regions that together make up the whole region from which data is collected. A very general framework for neutral coalescent models is presented and discussed. The framework comprises many of the most popular coalescent models that are currently used for analysis of genetic data. Assume data is collected from a series of consecutive regions of equal size. Then it is shown that the observed data forms a stationary, ergodic process. General conditions are given under which the maximum composite estimator of the parameters describing the model (e.g. mutation rates, demographic parameters and the recombination rate) is a consistent estimator as the number of regions tends to infinity.

Keywords Coalescent theory · Composite likelihood · Consistency · Estimator · Genomic data

C. Wiuf (✉)
Bioinformatics Research Center,
University of Aarhus,
Høegh-Guldbergsgade 10, Building 1090,
8000 Aarhus C, Denmark
e-mail: wiuf@birc.au.dk

1 Introduction

Many large scale genomic efforts concentrate on providing comprehensive genetic data points from many regions in the genome, rather than from few regions in many individuals. To human geneticists and population biologists, the availability of large genomic data sets are exciting because such data can be used to answer many important scientific questions regarding recombination and mutation in the human genome, and regarding the demographics and ancestry of human populations. Unfortunately, there are few appropriate statistical tools available for analysing large genomic data sets.

The basic coalescent process [19] and its modifications and generalizations [14] are appropriate mathematical models to describe the evolution of chromosomes and chromosomal regions, and their genealogical history. However, despite its mathematical and biological attractiveness, it has been shown to be computationally intractable to calculate the likelihood of a sample of chromosomes with more than a few dozen variable DNA positions [4,9–12,20,23,29]. These methods are based on approximating the likelihood using sequential Importance Sampling or Markov Chain Monte Carlo methods. For smallish data sets the likelihood and the maximum likelihood estimates can easily be evaluated using simulation, however for large data sets the methods become time consuming, computationally demanding and inaccurate.

This has sparked an interest in alternative approximate methods, such as *composite likelihood methods* (see e.g. [2] for some general perspectives on composite likelihood methods in statistics and their statistical properties). In the context of genomic data sets, a composite likelihood method treats different regions of the chromosome as being evolutionary independent regions, i.e. the composite likelihood function (CLF) is obtained by multiplying the likelihood of the individual regions. Dependencies between regions must die out sufficiently fast for the *maximum composite estimate* (MCE) of the parameters in the model to be consistent. Parameters here are of two kinds, namely those describing the shape of the genealogical relationship between small homologous chromosomal regions (e.g. demographic parameters and mutation rates) and those describing the correlation between genealogies in different regions (e.g. recombination and gene conversion rates). Composite likelihood methods have been suggested and used by a number of authors, among these Hudson [17], Fearnhead and Donnelly [5], Kim and Stephan [18], McVean et al. [22], Adams and Hudson [1], and Marth et al. [21].

Let $f_i(x_i;\alpha)$ be the likelihood of data x_i in region i, $i \leq l$ (x_i is the outcome of the stochastic variable X_i), where α is some parameter describing the possible models. Then, the logarithm of CLF is given by

$$\frac{1}{l}\sum_{i=1}^{l}\log(f_i(x_i;\alpha)), \tag{1}$$

and it is natural to consider conditions for which Eq. (1) converges to the expectation of $\log(f_i(x_i;\alpha))$ under the true model (with parameter α_0) as l tends to

infinity. This is an instance of the law of large numbers. If the variables X_i are independent this has become a stardard condition in relation to asymptotics (e.g. Hoffmann-Jørgensen 1994 for a full probabilistic treatment of the consequences of this condition); if the data are not independent convergence of Eq. (1) is still an important pre-requisite for convergence of the MCE [2,27,28].

Peskir [27] discusses the case where the X_is form a stationary ergodic process and provides similar results to those of Hoffmann-Jørgensen (1994). One important point of Peskir [27] is that his results allow for model misspecification, i.e. the true model of the data is not in the class of models parameterized by α. He shows that if Eq. (1) is converging under the true model, then so is the MCE. This is a useful addition, in particular in relation to genomic data analysis, because it is very unlikely that the true model is included in the class of models parameterized by α. However, if the true model is not included, then convergence of Eq. (1) under the true model cannot be proven, but must be postulated. As a further consequence, the interpretation of the MCE in relation to the biological reality is less straightforward. In the case of independent data points, convergence of the maximum likelihood estimator under model misspecification has been treated by White [30], and given in full generality by Hoffmann-Jøregnsen (1994). The results in this paper are based on Peskir [27].

Theoretical considerations of convergence properties for coalescent-based estimators have been published previously. Fearnhead [3] discusses the basic coalescent model with recombination and proves consistency of the MCE of the recombination rate as the number of genomic regions becomes large. (He also considers estimators based on pairs of sites but these fall outside the framework of the present article.) Fearnhead's proof is also based on convergence of Eq. (1). His result is here extended to more general models, in particular his result is extended to cover convergence of other parameters, such as demographic parameters. Fundamentally, data from a series of regions of the same length (L nucleotides) are considered and a coalescent-like model for the evolution of the sequences is assumed. It is further assumed that the state of all L nucleotides are observed. Nielsen and Wiuf [25] have provided some considerations about consistency of estimators in this and similar settings; however they have not undertaking detailed theoretical investigations.

Some familiarity with the basic coalescent and its generalizations is assumed. All proofs are in the Appendix.

2 The model

A continuous time coalescent model that allows for coalescence, migration, recombination, and gene conversion is considered. The setting is somewhat more general than in Griffiths and Tavaré [10,11], Griffiths and Tavaré [13], and Griffiths and Marjoram [8], but the model and notation is straightforwardly extrapolated from their model(s) and notation; see also Hein et al. [14], Hudson [15,16], and Wiuf and Hein [32] for further background on the notation and models. The model has the following characteristics:

(M1) A DNA sequence consists of L consequetive nucleotides

(M2) $t \in \mathbb{R}$ denotes time and $\alpha \in A \subseteq \mathbb{R}^d$ is an d-dimensional vector describing possible demographic and genetic scenarios

(M3) There are K time points, $T_1(\alpha), T_2(\alpha), \ldots, T_K(\alpha)$, where scaled rates for the four types of events can change discontinuously; corresponding to $K+1$ time epochs, $k = 0, 1, \ldots, K$ beginning at time $T_0(\alpha), T_1(\alpha), \ldots, T_K(\alpha)$, respectively, with $T_0(\alpha) = 0$ always

(M4) In time epoch k there are D_k demes. A sequence in deme i of epoch $k-1$ jumps at time $T_k(\alpha)$ to a new deme of epoch k as determined by a transition probability matrix $\{q^k_{ii'}(\alpha)\}_{i,i'}$: With probability $q^k_{ii'}(\alpha)$ a sequence in deme i, $i = 1, \ldots, D_{k-1}$, moves to deme i', $i' = 1, \ldots, D_k$, of epoch k

(M5) $\lambda_{ik}(t;\alpha)$ is the reciprocal relative deme size of deme $i = 1, \ldots, D_k$ at time t, and $\lambda_{ik}(0,\alpha) = 1$

(M6) $\nu_{ijk}(t;\alpha)$ is the scaled migration rate from deme i to deme j at time t; $i, j = 1, \ldots, D_k$

(M7) $\rho_{ik}(t;\alpha)$ is the per sequence scaled recombination rate in deme $i = 1, \ldots, D_k$ at time t; the break point is between position x and $x+1$, $x = 1, \ldots, L-1$, with equal probability

(M8) $\gamma_{ik}(t;\alpha)$ is the per sequence scaled gene conversion rate in deme $i = 1, \ldots, D_k$ at time t; one end point of the gene conversion tract is chosen uniformly, i.e. the break point is between position x and $x+1$, $x = 1, \ldots, L-1$ with equal probability. The other break point is chosen according to a symmetric distribution $g(y;\alpha)$, $y \in \mathbb{R}$, such that the break point is y nucleotides away from x and extends in either direction with equal probability

(M9) The mutation process is Markovian and the L positions evolve independently of each other along a given genealogy. The mutation process is parameterized by α

(M10) $\mathbf{n}(t) = (n_1(t), \ldots, n_{D_k}(t))$ is the sample configuration and counts the number of ancestral sequences at time t in each deme. (Note that the sample configuration does not contain any information about the allelic state of the sequences, only their numbers in different demes.) Total sample size at time t is $n(t) = \sum_{i=1}^{D_k} n_i(t)$

The functions λ_{ik}, ν_{ijk}, ρ_{ik}, and γ_{ik} are referred to as the rate functions. Typically, the parameters describing λ_{ik}, ν_{ijk}, ρ_{ik}, and γ_{ik} are variation independent. For completeness and notational convenience these parameters are collectively referred to as α. The number of demes is not allowed to depend on α. Times and rates are all scaled in N, the effective population size at time $t = 0$.

Condition (M3) refers to mergings and splittings of populations. Rates might depend on deme, e.g. reflecting different effective deme sizes, or that demes represent different species (e.g. human and chimps) with different genetic mechanisms. Rates might also depend on time epoch, e.g. reflecting that the number of demes might change from one epoch to the next, or that effective population sizes are modelled to change abruptly. Finally, rates might depend on time

locally, either because of fluctuations in effective population size, changes in migration patterns over time, or because the genetic mechanisms change as species' evolve.

One model of gene conversion is Wiuf and Hein [32], see also Wiuf [31]. It is parameterized by $G = 4NLg$ and $Q = qL$, where N is the effective population size, g the probability of a gene conversion tract initiating in a given position per generation, and q the probability that the tract extends beyond the neighbour nucleotide, i.e. the tract length has a geometric distribution. The tract extends to the right or to the left with equal probability. It follows that $\gamma_{ik}(t;\alpha) = G[1 + (1 - \mathrm{e}^{-Q})/Q] \equiv \gamma$ for large L, and an alternative parameterization, and perhaps more natural in this context, is thus given by (γ, q).

The dependence on α is often suppressed in the rate functions, the transition probability matrices and the times of epochs, as is the dependency of t in $\mathbf{n}(t)$. Further define (again with α suppressed)

(R1) The total rate at time t,

$$R_k(t;\mathbf{n}) = \sum_{i=1}^{D_k} \binom{n_i}{2} \lambda_{ik}(t) + \sum_{i=1}^{D_k} \frac{n_i}{2} \left[\sum_{i \neq j} \nu_{ijk}(t) + \rho_{ik}(t) + \gamma_{ik}(t) \right]$$

(R2) The relative rates of coalescence (c), migration (m), recombination (r), and gene conversion (g), respectively, for sequences in deme i,

$$c_{ik}(t;\mathbf{n}) = \frac{n_i(n_i - 1)\lambda_{ik}(t)}{2R_k(t;\mathbf{n})}, \quad m_{ijk}(t;\mathbf{n}) = \frac{n_i \nu_{ijk}(t)}{2R_k(t;\mathbf{n})}$$

$$r_{ik}(t;\mathbf{n}) = \frac{n_i \rho_{ik}(t)}{2R_k(t;\mathbf{n})}, \quad g_{ik}(t;\mathbf{n}) = \frac{n_i \gamma_{ik}(t)}{2R_k(t;\mathbf{n})}.$$

For convenience, e will be short for the rate of an arbitrary event, e.g. $e(t;\mathbf{n}) = r_{ik}(t;\mathbf{n})$

With the above notation and definitions the model can be described as a birth-death process with migration between demes and time dependent rates, i.e. a time inhomogeneous birth–death process with migration between demes. The time, T_{next}, until the next event depends on the rate $R_k(t;\mathbf{n})$ and the present time, s, and has density

$$P(T_{\text{next}} > t | \mathbf{n}(s) = \mathbf{n}) = \exp\left\{ -\int_s^{s+t} R_k(u;\mathbf{n}) \mathrm{d}u \right\}, \tag{2}$$

i.e., T_{next} is a stretched exponential variable. If $T_{\text{next}} > T_k(\alpha)$ then the next event did not happen in time epoch k, and a new variable is drawn with rate $R_{k+1}(t;\mathbf{n})$. The type of the event is determined by the relative rates; if a coalescent event then the number of ancestral sequences $n(t)$ goes down by one;

if a migration event, $n(t)$ remains unchanged; and if a recombination event or a gene conversion event, then $n(t)$ goes up by one. Mutations and break points are superimposed afterwards. This formulation of the model is very similar to the birth-death process described in Griffiths and Marjoram [8] for the coalescent with recombination only.

Whenever $n(t) = 1$, a *common ancestor* of the sample has been found. The first time, T_{MRCA}, for which $n(t) = 1$, is called the time of the *most recent common ancestor* (MRCA). It is not guaranteed that the process will reach the state of a MRCA, and hence it must be assumed that this is the case,

$$P(T_{\mathrm{MRCA}} < +\infty | \mathbf{n}(0) = \mathbf{n}) \; = \; 1. \tag{3}$$

A necessary condition for condition (3) to hold is

$$\int_{T_K(\alpha)}^{\infty} \lambda_{iK}(t;\alpha)\mathrm{d}t \; = \; \infty, \tag{4}$$

for all $i = 1, \ldots, D_K$; however it is not a sufficient condition as Eq. (3) also depends on the rates of recombination and migration. However, because the birth rate is linear in the number of sequences and the death rate is quadratic, Eq. (3) is likely to hold in all reasonable models.

3 Sample histories

The aim of this section is to introduce and discuss the concept of a (time-dated) sample history and to prove that certain probabilities are continuous functions of α.

An outcome $\mathcal{T}$ of the birth-death process is called a time-dated (sample) history and consists of a series of sample configurations. Figure 1 illustrates the concept of a history; a formal definition is given below. The sequences in a sample configuration are called ancestral sequences, though these need not be ancestors of the sample in a genetic sense. Other formulations of the model avoid the genetically 'empty' or non-ancestral sequences. However, formulating the model as a birth–death process has the advantage that the probability of the data takes the form

$$P(\mathrm{Data}) \; = \; \int_{\mathcal{T}} \sum_{\text{break pts}} P(\mathrm{Data}|\mathcal{T}, \text{break pts}) P(\text{break pts}|\mathcal{T}) P(\mathcal{T}) \mathrm{d}\mathcal{T}, \tag{5}$$

(the number of ways break points can be chosen are finite for a given history) in contrast to other formulations of the process where the break points are part of the history of the data.

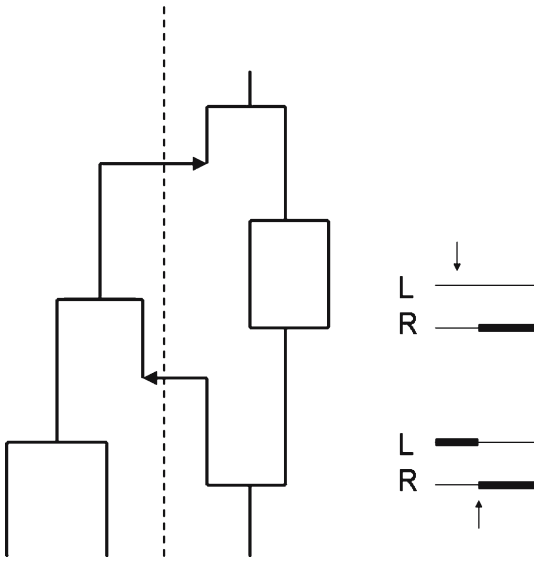

Fig. 1 The figure shows an example of a history of a sample of three sequences, two sampled in one deme and one sampled in another deme (the two demes are separated by the *dashed line*). The first event is a recombination event, splitting the sequence into two sequences with ancestral material as shown to the right of the event (*thick lines* ancestral material *thin lines* non-ancestral). L and R indicate the left and the right sequence after the event. At the second recombination event a genetically 'empty' sequence is created that does not carry any genetic information ancestral to the sample. Migration events are indicated by arrows with the head showing the direction of the migration

A time-dated history $\mathcal{T}$ of a sample $\mathbf{n}(0)$ taken at time $t = 0$ (in epoch 0) is a series of sample configurations with time of occurence attached,

$$\mathbf{n}(T_k), \mathbf{n}(t_{k1}), \ldots, \mathbf{n}(t_{kj_k}) \tag{6}$$

for each epoch $k = 0, 1, \ldots, K$, and $T_k < t_{kj} < t_{k,j+1} < T_{k+1}$ (with $T_{K+1} = \infty$), such that

(H1) $\mathbf{n}(t_{k1})$ is obtainable from $\mathbf{n}(T_k)$ by a single event
(H2) $\mathbf{n}(t_{k,j+1})$ is obtainable from $\mathbf{n}(t_{kj})$ by a single event
(H3) $n(t) > 1$ for $0 \leq t < t_{Kj_K}$
(H4) $n(t_{Kj_K}) = 1$, and
(H5) $\mathbf{n}(T_{k+1})$ is a possible transition from $\mathbf{n}(t_{kj_k})$.

The type of event transforming one configuration into another is taken as being part of the definition of a time-dated history, but it is generally suppressed in the notation.

A history $\mathcal{H}$ differs from a time-dated history in that times of sample configurations are ignored, only the order in which the configurations occur is registred and whether a configuration is the first configuration in an epoch (at times T_k, $k = 0, 1, \ldots, K$).

Informally, a history is a series of events describing the evolution of the sample. It evolves (backwards in time) through mergings (coalescent events), splittings (genetic exchange events), and migrations such that finally a MRCA of the sample is found. Mutation events and break points are not part of the (time-dated) history.

The probability $P_\alpha(\mathcal{T})$ of a time-dated history $\mathcal{T}$ can be computed from the total rate, the relative rates and the transition probability matrices defined above. It can be written in the form

$$P_\alpha(\mathcal{T}) \;=\; \prod_{k=0}^{K} U_{\alpha k}(\mathcal{T}) \prod_{k=0}^{K-1} V_{\alpha k}(\mathcal{T}), \tag{7}$$

where

$$\begin{aligned} U_{\alpha k}(\mathcal{T}) &= P_\alpha\{\mathbf{n}(t_{kj_k}), \mathbf{n}(t_{k,j_k-1}), \ldots, \mathbf{n}(t_{k1})|\mathbf{n}(T_k)\} \\ &= P_\alpha\big(\mathbf{n}(t_{k1})|\mathbf{n}(T_k)\big) \prod_{j=1}^{j_k-1} P_\alpha\left(\mathbf{n}(t_{k,j+1})|\mathbf{n}(t_{kj})\right) \end{aligned} \tag{8}$$

and

$$V_{\alpha k}(\mathcal{T}) \;=\; P_\alpha(\mathbf{n}(T_{k+1})|\mathbf{n}(t_{kj_k})). \tag{9}$$

Strictly speaking, $P_\alpha(\mathcal{T})$ is a mixture of a density with respect to a (multi-dimensional) Lebesgue measure and a Markov Chain. The probabilities in Eq. (8) depend on the total rate and the relative rates, whereas the probability in Eq. (9) also depends on the transition matrix $\{q^k_{ii'}(\alpha)\}$. To decompose the probability in Eq. (8) into a product the Markov Property of the coalescent process is used. The Markov Property also guarantees that the individual probability terms have the form

$$P_\alpha(\mathbf{n}(t_{k,j+1})|\mathbf{n}) \;=\; e(t_{k,j+1};\mathbf{n}) R_k(t_{k,j+1};\mathbf{n}) \exp\left\{ -\int_{t_{kj}}^{t_{k,j+1}} R_k(u;\mathbf{n}) \mathrm{d}u \right\}, \tag{10}$$

where $\mathbf{n} = \mathbf{n}(t_{kj})$. Also

$$P_\alpha(\mathbf{n}(t_{k1})|\mathbf{n}) \;=\; e(t_{k1};\mathbf{n}) R_k(t_{k1};\mathbf{n}) \exp\left\{ -\int_{T_k(\alpha)}^{t_{k1}} R_k(u;\mathbf{n}) \mathrm{d}u \right\}, \tag{11}$$

where $\mathbf{n} = \mathbf{n}(T_k)$. Equation (9) becomes

$$P_\alpha(\mathbf{n}(T_{k+1})|\mathbf{n}) \;=\; Q_{\alpha k}(\mathbf{n}(T_{k+1})|\mathbf{n}) \left[1 - \exp\left\{ -\int_{t_{kj}}^{T_{k+1}(\alpha)} R_k(u;\mathbf{n}) \mathrm{d}u \right\} \right], \tag{12}$$

where $\mathbf{n} = \mathbf{n}(t_{kj_k})$, and $Q_{\alpha k}(\mathbf{n}(T_{k+1})|\mathbf{n})$ is a (finite) sum of multinomial probabilities reflecting the ways $\mathbf{n}$ can be transformed into $\mathbf{n}(T_{k+1})$. It is calculated from $\{q^k_{ii'}(\alpha)\}_{i,i'}$.

Integrating $P_\alpha(\mathcal{T})$ over time provides the probability $P_\alpha(\mathcal{H})$ of the corresponding history $\mathcal{H}$,

$$P_\alpha(\mathcal{H}) \;=\; \int_{\mathbf{T}_K} \int_{\mathbf{T}_{K-1}} \cdots \int_{\mathbf{T}_0} P_\alpha(\mathcal{T})\, \mathrm{d}\mathbf{t}_K \mathrm{d}\mathbf{t}_{K-1} \cdots \mathrm{d}\mathbf{t}_0, \tag{13}$$

where $\mathbf{t}_k = (t_{k1}, t_{k2}, \ldots, t_{kj_k})$ and $\mathbf{T}_k = \{\mathbf{t}_k | T_k < t_{kj} < t_{k,j+1} < T_{k+1}\}$, again with $T_{K+1} = \infty$.

To procede a number of regularity conditions is required.

Assumption 1 Assume the rate functions (*cf.* assumptions M5–M8) are continuous in t for each time epoch (*cf.* M3–M4) and fixed $\alpha \in A$ (left/right continuous at $T_k(\alpha)$ with finite limit). Further, assume $\mathrm{int}(\mathrm{cl}(A)) \subseteq A$ and that the rate functions are continuous in α for fixed t, $T_k(\alpha)$ (*cf.* M3) is continuous in α, $\{q^k_{ii'}(\alpha)\}_{i,i'}$, $k = 1, \ldots, K$ (*cf.* M4) are continuous in α, $g(y;\alpha)$ (*cf.* M8) is continuous in α, and the mutation process (*cf.* M9) is continuous in α, for any $\alpha \in A$.

Let $1_S(x)$ be the indicator function for a set S. Assume the functions in u, indexed by n,

$$1_{[T_K(\alpha_n),t]}(u)\, R_k(u;\alpha_n,\mathbf{n}), \tag{14}$$

for $t > 0$, and

$$1_{[T_k(\alpha_n),T_{k+1}(\alpha_n)]}(u)\, R_k(u;\alpha_n,\mathbf{n}), \tag{15}$$

for $k = 0, \ldots, K-1$, are uniformly integrable for any series $\alpha_n \to \alpha$ and any $\mathbf{n}$ and t (fixed).

Further assume the functions in t, indexed by n,

$$1_{[T_K(\alpha_n),\infty)}(t)\, R_k(t;\alpha_n,\mathbf{n}) \exp\left\{ -\int_{T_K(\alpha_n)}^{t} R_k(u;\alpha_n,\mathbf{n})\mathrm{d}u \right\} \tag{16}$$

are uniformly integrable for any series $\alpha_n \to \alpha$ and any $\mathbf{n}$ (fixed).

The uniform integrability conditions are typically fulfilled, e.g. Eqs. (14) and (15) are fulfilled if $R_k(u;\alpha_n,\mathbf{n}) \le C(\mathbf{n})$ for all u and some constant depending on $\mathbf{n}$, and Eq. (16) is fulfilled if $R_k(u;\alpha_n,\mathbf{n}) \ge \varepsilon(\mathbf{n}) > 0$ for some constant depending on $\mathbf{n}$.

Note that Condition (3) and Assumption 1 ensure that there are countable many histories $\mathcal{H}$ for a given sample configuration $\mathbf{n}$, such that

$$1 \;=\; \sum_{\mathcal{H}} P_\alpha(\mathcal{H}), \tag{17}$$

and

$$P_\alpha(\text{Data}) = \sum_{\mathcal{H}} P_\alpha(\text{Data}|\mathcal{H})\mathcal{P}_\alpha(\mathcal{H}). \tag{18}$$

Asumption 1 guarantees the following result.

Lemma 1 *$P_\alpha(\mathcal{T})$, $P_\alpha(\mathcal{H})$ and $P_\alpha(\text{Data})$ are continuous in α.*

4 Two-locus sample histories

The aim of this section is to prove that two co-evolving loci, separated by a large genetic distance almost evolve independently of each other. In this section a fixed α is considered, in contrast to the previous section where continuity properties in α were investigated.

Let two loci each L nucleotides long and separated by M nucleotides be given. Only the DNA sequences of the two loci are observed ($2L$ nucleotides). A (time-dated) history for the two loci is embedded in the (time-dated) history of the $2L + M$ nucleotides. However, the history might be described by a modified birth-death process with migration that has fewer events than the full history of the $2L + M$ nucleotides. It can be done in the following way.

There are three types of sequences: (1) Those that are ancestral to locus 1 only, (2) those that are ancestral to locus 2 only, and (3) those that are ancestral to both loci. In the beginning all sequences are of type 3. If a recombination event happens in a type 3 sequence between the two loci (in the M nucleotides) then it is replaced by one sequence of type 1 and one of type 2. If a recombination event happens in a type 3 sequence in locus 1 (2), a type 3 sequence and a type 1 (2) sequence are created. If a recombination event happens in a type 1 (2) sequence in locus 1 (2), then two type 1 (2) sequences are created. After a MRCA is found for locus 1 (2) that locus is subsequently ignored when tracing the history of the other locus further back in time. See Fig. 2 for an illustration.

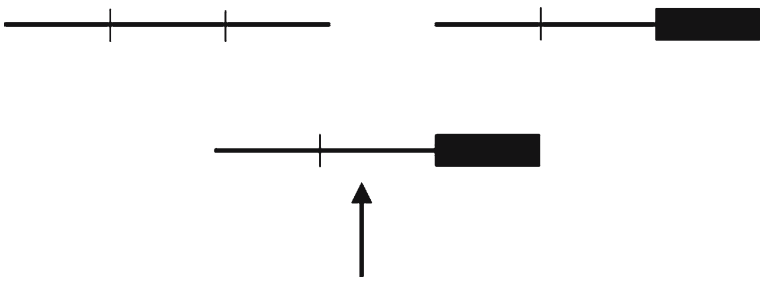

Fig. 2 An illustration of the difference between the full coalescent model and the modified model. Shown is a sequence of $L + M + L$ nucleotides separated by small vertical bars. Only locus 2 is ancestral to the sample in this example. If a recombination event happens in the middle M nucleotides it counts in the (full) history of the $2L + M$ nucleotides, whereas it does not count in the modified history

The recombination and gene conversion rates of type 1 and 2 sequences do not depend on M, only L, whereas the rates of type 3 sequences depend on L and M. To procede the following assumption is required.

Assumption 2 The recombination rate $\rho_{ik}(t)$ is linear in sequence length, such that the rate for type 1 and 2 sequences is $\rho_{ik}(t) = L\rho_{0ik}(t)$, and the rate for type 3 sequences is $(2L + M)\rho_{0ik}(t)$. Further, it is assumed that $\rho_{0ik}(t)$ is bounded uniformly away from 0, i.e. $\rho_{0ik}(t) > \rho_0 > 0$ for all $t \geq 0$, i, and k.

This modification of the process reduces the total number of recombination events in a sample history substantially. For the standard coalescent process the number of recombination events is of order $e^{(2L+M)\rho}$ for the full birth-death process, whereas it is of order $e^{2L\rho}$ for the modified process [9]. Here ρ is the per site recombination rate.

Assumption 2 implies that the rate of recombination between two ancestral loci is $M\rho_{0ik}(t)$. Informally, this has the consequence that for large M, a type 3 sequence is likely to break up into a type 1 and a type 2 sequence before being involved in other events. The statement will be made more rigourous at the end of the section.

The rate of gene conversion events for sequences of types 1 and 2, respectively, is just the rate $\gamma_{ik}(t)$. Only break points within an ancestral locus affect the history of the sample. The rate for sequences of type 3 can be divided into the rate of events with one or two break points in locus 1 and none in locus 2 (and *vice versa* for locus 2), and the rate of events with a break point in each locus. Thus, the rate, $\gamma^*_{ik}(t)$, of gene conversion events affecting one or both of the loci can be decomposed into

$$\gamma^*_{ik}(t) \;=\; 2\gamma^1_{ik}(t) + \gamma^2_{ik}(t), \tag{19}$$

where $\gamma^1_{ik}(t)$ is the rate of gene conversion in locus 1 (2) with the second break point not in locus 2 (1) and $\gamma^2_{ik}(t)$ is the rate of gene conversion affecting both loci. The term $\gamma^1_{ik}(t)$ appears once for each locus. Events with two break points within the M nucleotides can be ignored as they do not affect the sample's history.

It follows that $\gamma^1_{ik}(t) + \gamma^2_{ik}(t) = \gamma_{ik}(t)$ and $\gamma^*_{ik}(t) = 2\gamma_{ik}(t) - \gamma^2_{ik}(t) \leq 2\gamma_{ik}(t)$. According to Wiuf [32], $\gamma^2_{ik}(t) \to 0$ as $M \to \infty$. It shows that for large M only gene conversion events in type 1 and 2 sequences are likely to occur.

The next lemma shows that the histories of two loci with large distance M are very similar to the histories of two unlinked loci (corresponding to $M = \infty$). First a definition is required.

Definition 1 *A T3-history* (*"Type* 3 *history"*) *is a time-dated history of two loci fulfilling the following condition: If the sample configuration* $\mathbf{n}(t)$ *at time t contains sequences of type* 3*, then the first event after time t is a recombination event in a type* 3 *sequence with break point between the two loci* (*in the M nucleotides*).

Note that all histories of two unlinked loci are T3 in that any type 3 sequence break up instantaneously. Let P_M denote the joint distribution of a sample of

two loci, and P_∞ the joint distribution of a sample of two independent loci, i.e. for $M = \infty$.

Lemma 2 *Let $T_{\mathrm{MRCA}}(i)$ be the time of the MRCA of locus i, $i = 1, 2$; $E(12)$ the number of events where a type 3 sequence is created from a type 1 and a type 2 sequence; $E(i)$, $i = 1, 2$, the number of events only affecting the history of locus i; and Δ the minimum time span between two events none of which are recombination events in type 3 sequences.*

Choose, $t_\epsilon < +\infty$, $d_\epsilon > 0$, $e_\epsilon(12)$, and $e_\epsilon(i)$, $i = 1, 2$, such that $P_\infty(K_\epsilon) > 1-\epsilon$, where $K_\epsilon = \{T_{\mathrm{MRCA}}(i) < t_\epsilon, \delta_\epsilon < \Delta, E(12) < e_\epsilon(12), E(i) < e_\epsilon(i), i = 1, 2\}$ for $\epsilon > 0$. Then

$$P_M(K_\epsilon, \mathrm{T3}) \; > \; 1 - 2\epsilon$$

for $M > M_\epsilon$, where M_ϵ depends on K_ϵ.

To prove consistency of the MCE bounds (depending on M) on the probabilities of individual histories are required. However these appear of little general interest and will not be reproduced here, but derived in the Appendix in connection with the proofs of Lemmas 2 and 3.

5 Consistency

In order to prove consistency Peskir (1996) is followed closely. He considers general stationary and ergodic processes, that further fulfill a number of regularity conditions. The regularity conditions are similar to (but slightly stronger than) conditions that normally are required for the maximum likelihood estimator to be consistent under repeated (independent) sampling. Peskir's conditions are generally met in coalescent models that have been used for analyses of data.

Some conditions are now imposed to ensure ergodicity of the process. Assume

(C1) An infinite array of consecutive segments of L nucleotides each, sampled from an infinitely long chromosome is given
(C2) The array $\mathrm{Data} = (\mathrm{Data}_1, \mathrm{Data}_2, \mathrm{Data}_3, \ldots)$, where Data_j is the observed data in segment j, forms a stationary process, i.e. the distribution of the data is translational invariant, for any $\alpha \in A$
(C3) $P_\alpha(\mathrm{Data}_j)$ is positive for all $\alpha \in A$ and all possible Data_j
(C4) Assumptions 1 and 2 are true.

The first two items are natural requirements in the context of models for large genomic data sets. Then the following lemma holds.

Lemma 3 *The stationary process $\mathrm{Data} = (\mathrm{Data}_1, \mathrm{Data}_2, \mathrm{Data}_3, \ldots)$ is ergodic. In particular, the CLF*

$$h_l(\alpha; \mathrm{Data}) \; = \; \frac{1}{l} \sum_{j=1}^{l} \log(P_\alpha(\mathrm{Data}_j)) \tag{20}$$

converges almost surely for $l \to \infty$ to a limit, say, $I_{\alpha_0}(\alpha)$, for any $\alpha \in A$, where $\alpha_0 \in A$ denotes the true value.

The proof of Lemma 3 suggests that the rate of convergence of the MCE is $\log(M)/M$, where M is the length between the two most distant segments. However, the convergence rate cannot be made exact by the methods used here: The proof shows that outside a set of measure $\epsilon > 0$ (where ϵ can be chosen arbitrary small) the rate of convergence of $h_l(\alpha; \text{Data})$ is $\log(M)/M$. This cannot directly be translated into a rate of convergence of the MCE.

Assumption 3 Assume

$$\inf_{\alpha \in A} h_l(\alpha; \text{Data}) > -\infty \tag{21}$$

for any outcome of Data.

For example, Assumption 1 ensures that Assumption 3 is fulfilled if A is closed and bounded (remember there are only finitely many possible data points). The above conditions and assumptions guarantee the following:

Theorem 1 *Let $A_{\max}$ be the set of maximum points of $I_{\alpha_0}(\alpha)$, where α_0 denotes the true value. Always α_0 is in $A_{\max}$. Let $\hat{\alpha}_l$ be the MCE of α obtained by maximizing the CLF $h_l(\alpha; \text{Data})$ with respect to α. Then the set of accumulation points of the series $\{\hat{\alpha}_l\}_{l \geq 1}$ is in $A_{\max}$. In particular, if $A_{\max} = \{\alpha_0\}$, then $\hat{\alpha}_l$ converges almost surely to α_0 for $l \to \infty$.*

If the model is identifiable (not over-parameterized), then $A_{\max} = \{\alpha_0\}$. For a model to be identifiable it is required that for all α and α' there is some Data_j such that $P_\alpha(\text{Data}_j) \neq P_{\alpha'}(\text{Data}_j)$.

If the true model is not in A the ergodic property cannot be proven from the assumptions. If the ergodic property holds then Theorem 1 is still true and $\hat{\alpha}_l$ has accumulation points in $A_{\max}$. In particular, if $A_{\max} = \{\alpha_1\}$ for some α_1 in A, then $\hat{\alpha}_l$ converges almost surely to α_1 for $l \to \infty$.

6 Discussion

Consistency of the MCE has been discussed in a very general coalescent framework and conditions for which the MCE is consistent has been provided. The examples that fall under this framework are many, including the basic coalescent with recombination and a general mutation process (e.g. Jukes–Cantor, Kimura, or F84; see e.g. [8]). Coalescent models allowing for exponential growth, or logarithmic growth, and bottlenecks similarly fulfill the conditions for Theorem 1 to apply. For example the demographic models that are allowed in Hudson's program `ms` fulfill the conditions (http://www.home.uchicago.edu/~rhudson1). Similarly, Theorem 1 applies to models with a fixed number of demes of constant size and migration between the demes [26], and Theorem 1

also applies to the model by Nielsen and Wakeley [24] in which two populations mix some time in the past and migration is (not) allowed while the two populations are separated. All the parameters used to describe these models can be estimated consistently, or at least the MCE will approach a set of equally optimal points, as specified in Theorem 1. Initially one might investigate $I_{\alpha_0}(\alpha)$ computationally to ensure it is likely to have only one maximum point. Only in rare cases will this information be available analytically.

The break point distributions for recombination and gene conversion are perhaps not as general as sometimes required. The break point distributions, uniform in both cases, are both independent of time and translational invariant, i.e., variation in recombination and gene conversion rates along the chromosome is not modelled, neither is variation in tract length. These features are realistic and important biological features. One way to accomodate for variation in the rates is to adopt a prior distribution on the rates. This will introduce additional correlation between genealogies and data of linked loci and also complicate the likelihood of the data given the history, because break points are no longer drawn uniformly on the sequence but according to some other distribution. This kind of model is not covered by the theory presented here – and cannot be incoorporated without modifications.

The tract length distribution $g(y;\alpha)$ can be made more general (e.g. time and deme dependent) – however it was kept simple for convenience. None of the proofs nor techniques used in the paper require special modification to apply more generally.

The present theory does not apply to models with selection, and to models with contex dependent mutation rates, e.g if the mutation rate of a nucleotide depends on the states of its neighbours. In these cases the likelihood of the data does not separate in the form given in Eq. (5) and the techniques applied here are not sufficient to prove consistency. An exception is codon-based models where codons evolve independently of each other [7].

If the regions are not equally spaced the theory still applies. This will often be the case for empirically collected data. Similarly, it is possible to prove consistency if the regions are not of equal size. However, in this case it must be assumed that all regions have size less than L (for some L), because otherwise convergence of the CLF cannot be guaranteed. Strictly speaking it is not necesaary to assume a lower bound on the size, because the theory applies equally well for regions of size $L = 1$ as for regions of size $L > 1$. However, if L is very small (e.g. $L = 1$) then the model is likely to be over-parametrized and the CLF will not have a unique maximum.

Acknowledgments Rasmus Nielsen is thanked for raising the question of whether composite likelihood estimators are consistent at a meeting at the Bannf Research Station, and for reading and commenting on the manuscript. An anonymous reviewer is thanked for providing useful criticism that improved the presentation of the proofs. The author is supported by The Danish Cancer Society and a travel grant from the Carlsberg Foundation that made the author's participation in the meeting possible.

Appendix

In this section proofs of the lemmas in the main text are given.

Proof of Lemma 1. Consider $P_\alpha(\mathcal{T})$. Continuity follows from continuity (by assumption) of $e(t;\alpha,\mathbf{n})$, $R_k(t;\alpha,\mathbf{n})$, and $Q_{\alpha k}(\mathbf{n}(T_{k+1})|\mathbf{n}(t_{kj_k}))$. The latter is a sum of finitely many continuous terms. To prove continuity of the integrals in Eq. (10)–(12) it suffices to note that the rate functions are uniformly integrable according to Assumption 1, Eqs. (14) and (15).

Consider $P_{\alpha_n}(\mathcal{H})$ written in the form of Eqs. (13):

$$P_{\alpha_n}(\mathcal{H}) = \int_{\mathbf{T}_K}\int_{\mathbf{T}_{K-1}} \dots \int_{\mathbf{T}_0} P_{\alpha_n}(\mathcal{T})\, \mathrm{d}\mathbf{t}_K \mathrm{d}\mathbf{t}_{K-1} \dots \mathrm{d}\mathbf{t}_0. \tag{22}$$

Using Eq. (7) each of the terms $U_{\alpha_n k}(\mathcal{T})$ and $V_{\alpha_n k}(\mathcal{T})$, $k = 0,\dots,K-1$, are uniformly integrable by Assumption 1, Eqs. (14) and (15), because Eqs. (10) and (11) are bounded by Eqs. (14) and (15). The last term $U_{\alpha_n K}(\mathcal{T})$ is uniformly integrable by Assumption 1, Eq. (16). Because $P_{\alpha_n}(\mathcal{T}) \to \mathcal{P}_\alpha(\mathcal{T})$ for every time-dated history consistent with $\mathcal{H}$, it follows that $P_{\alpha_n}(\mathcal{H}) \to P_\alpha(\mathcal{H})$, and that $P_\alpha(\mathcal{H})$ is continuous in α.

Finally, consider $P_\alpha(\mathrm{Data}) = \sum_{\mathcal{H}} P_\alpha(\mathrm{Data}|\mathcal{H})P_\alpha(\mathcal{H})$. For given $\mathcal{H}$, one has $P_{\alpha_n}(\mathcal{H}) \to \mathcal{P}_\alpha(\mathcal{H})$, and further $1 = \sum_{\mathcal{H}} P_{\alpha_n}(\mathcal{H})$. It follows, using Fatou's lemma, that

$$\sum_{\mathcal{H}\in\Omega} P_{\alpha_n}(\mathcal{H}) \to \sum_{\mathcal{H}\in\Omega} P_\alpha(\mathcal{H})$$

for any $\alpha_n \to \alpha$ and any set Ω of histories. Hence for given $\epsilon > 0$ one can choose E such that

$$\sum_{\mathcal{H}\in\Omega_E} P_{\alpha_n}(\mathcal{H}) < \epsilon$$

for large n, where Ω_E is the set of histories with more than E events. Also $P_{\alpha_n}(\mathrm{Data}|\mathcal{H}) \to P_\alpha(\mathrm{Data}|\mathcal{H})$, because

$$P_{\alpha_n}(\mathrm{Data}|\mathcal{H}) = \int_{\mathbf{T}_K}\int_{\mathbf{T}_{K-1}} \cdots \int_{\mathbf{T}_0} P_{\alpha_n}(\mathrm{Data}|\mathcal{T})P_{\alpha_n}(\mathcal{T})\mathrm{d}\mathbf{t}_K \mathrm{d}\mathbf{t}_{K-1} \cdots \mathrm{d}\mathbf{t}_0,$$

$P_{\alpha_n}(\mathrm{Data}|\mathcal{T})P_{\alpha_n}(\mathcal{T})$ and $P_{\alpha_n}(\mathcal{T})$ are uniformly integrable, and $P_{\alpha_n}(\mathrm{Data}|\mathcal{T})$ is continuous by assumption (a sum of finitely many continuous terms).

It follows that

$$P_{\alpha_n}(\mathrm{Data}) = \sum_{\mathcal{H}\in\Omega_E^c} P_{\alpha_n}(\mathrm{Data}|\mathcal{H})P_{\alpha_n}(\mathcal{H}) + \sum_{\mathcal{H}\in\Omega_E} P_{\alpha_n}(\mathrm{Data}|\mathcal{H})P_{\alpha_n}(\mathcal{H})$$

converges to $P_\alpha(\text{Data})$ because the first summation is over finitely many continuous terms and the second is at most ϵ for sufficiently large n. □

Proof of Lemma 2. The lemma is proven in the special case where $\mathbf{n}(0)$ is a configuration with $2n$ sequences such that there are n sequences of type 1 and n sequences of type 2. The proof in the general case can be derived in the same way as the proof presented here.

Let $\mathcal{T}_{12}$ be a T3-history. The probability of $\mathcal{T}_{12}$ is (with the terms explained below)

$$P_M(\mathcal{T}_{12}) = P_\infty(\mathcal{T}_{12}^\star) \prod_{j=1}^{E(12)} r_{ik_j}(s_j;\mathbf{n}_j) R_{k_j}(s_j;\mathbf{n}_j) \times \exp\left\{ -\int_{t_j}^{s_j} R_{k_j}(u;\mathbf{n}_j) - R_{k_j}(u;\mathbf{n}_j')\mathrm{d}u \right\}, \tag{23}$$

where $\mathcal{T}_{12}^\star$ is the corresponding history for $M = \infty$; and t_j, $j = 1,\ldots,E(12)$, are the times of the $E(12)$ events, where a type 3 sequence is created, and s_j, $j = 1,\ldots,E(12)$, are the times when the sequence again is broken up into a type 1 and a type 2 sequence. The configuration $\mathbf{n}_j$ has one type 3 sequence and $\mathbf{n}_j'$ is the same as $\mathbf{n}_j$ but with the type 3 sequence broken up.

The times of events in $\mathcal{T}_{12}^\star$ are the same as the times of events in $\mathcal{T}_{12}$ with the exception that type 3 sequences break up instantaneously, i.e. at time t_j. Consequently, there is a one-many relation between histories $\mathcal{T}_{12}^\star$ and $\mathcal{T}_{12}$. The term

$$r_{ik_j}(s_j;\mathbf{n}_j) R_{k_j}(s_j;\mathbf{n}_j) \exp\left\{ -\int_{t_j}^{s_j} R_{k_j}(u;\mathbf{n}_j)\mathrm{d}u \right\}$$

in Eq. (23) is the density of a recombination event in a type 3 sequence at time s_j. For $M = \infty$, the recombination event happens at time t_j with probability 1 and the rate becomes $R_{k_j}(u;\mathbf{n}_j')$ for $t_j \le u \le s_j$.

Integrating out s_j provides an upper and a lower bound to $P_M(\mathcal{T}_{12})$. First note that on K_ϵ, the number of events is bounded by $e_\epsilon = 2e_\epsilon(12) + e_\epsilon(1) + e_\epsilon(2)$, and $c_1(t), c_2(t), d(t) > 0$ can be chosen such that

$$R_k(u;\mathbf{n}_j) - R_k(u;\mathbf{n}_j') > d(t)M, \tag{24}$$

$$1 + \frac{c_1(t)}{M} > \frac{R_k(u;\mathbf{n}_j)}{R_k(u;\mathbf{n}_j) - R_k(u;\mathbf{n}_j')} \ge 1, \tag{25}$$

and

$$1 \ge r_{ik}(u;\mathbf{n}_j) > 1 - \frac{c_2(t)}{M} \tag{26}$$

for all $0 \leq u \leq t$, $i = 1, \ldots, D_k$, $k = 1, \ldots, K$ and any possible configuration $\mathbf{n}_j$ with at least one type 3 sequence. It is possible to choose such numbers because of Assumption 1 and 2 and because $2 \leq \sum_i n_i \leq 2n + e_\epsilon$ is a (crude) upper bound to the total number of sequences in the sample at any time on K_ϵ. It follows that

$$P_\infty(\mathcal{T}_{12}^\star)\left(1 + \frac{c_1(t_\epsilon)}{M}\right)^{e(12)} > \int\limits_{S_{12}} P(\mathcal{T}_{12})\mathrm{d}\mathbf{s}, \tag{27}$$

where

$$S_{12} = \{t_j' \geq u \geq t_j | j = 1, \ldots, E(12)\}$$

and t_j' is the time of the event following that at time s_j. To prove the inequality, relations (25) and (26) have been used, in addition to $e(12) \geq E(12)$, and

$$1 \geq \int\limits_{t_j \leq s_j \leq t_j'} \Lambda(s_j) \exp\left\{-\int\limits_{t_j}^{s_j} \Lambda(u)\mathrm{d}u\right\} \mathrm{d}s_j,$$

where $\Lambda(s)$ is a function such that $\Lambda(s) > 0$; in particular this is true for $\Lambda(u) = R_k(u; \mathbf{n}_j) - R_k(u; \mathbf{n}_j')$.

Similarly, it follows that

$$\begin{aligned}\int\limits_{S_{12}} P(\mathcal{T}_{12})\mathrm{d}\mathbf{s} &> P_\infty(\mathcal{T}_{12}^\star)\left(1 - \frac{c_2(t_\epsilon)}{M}\right)^{e(12)} \prod_{j=1}^{e(12)} \left[1 - \mathrm{e}^{-(t_j' - t_j)d(t_\epsilon)M}\right] \\ &> P_\infty(\mathcal{T}_{12}^\star)\left(1 - \frac{c_2(t_\epsilon)}{M}\right)^{e(12)} \left[1 - \mathrm{e}^{-\delta_\epsilon d(t_\epsilon)M}\right]^{e(12)}, \end{aligned} \tag{28}$$

where it has been used that

$$\int\limits_{t_j \leq s_j \leq t_j'} \Lambda(s_j) \exp\left\{-\int\limits_{t_j}^{s_j} \Lambda(u)\mathrm{d}u\right\} \mathrm{d}s_j \;\geq\; 1 - \mathrm{e}^{-(t_j' - t_j)\delta},$$

if $\Lambda(u) > \delta$. This is in particular true for $\Lambda(u) = R_{k_j}(u; \mathbf{n}_j) - R_{k_j}(u; \mathbf{n}_j') > d(t_\epsilon)M = \delta$.

For convenience define the constants k_1 and k_2 by

$$k_1 \;=\; \left(1 + \frac{c_1(t_\epsilon)}{M}\right)^{e(12)}, \tag{29}$$

and

$$k_2 \;=\; \left(1-\frac{c_2(t_\epsilon)}{M}\right)^{e(12)}\left[1-\mathrm{e}^{-\delta_\epsilon d(t_\epsilon)M}\right]^{e(12)}. \tag{30}$$

Both of these are $1+O(1/M)$ and depend on K_ϵ. (The constants k_1 and k_2 will be used again in the proof of Lemma 3.)

Integrating over the remaining times, keeping the constraint K_ϵ, gives

$$k_1\, P_\infty(\mathcal{H}_{12}^\star, K_\epsilon) \;>\; P_M(\mathcal{H}_{12}, K_\epsilon) \;>\; k_2\, P_\infty(\mathcal{H}_{12}^\star, K_\epsilon).$$

Summing over all T3 histories compatible with marginal histories $\mathcal{H}_1$ and $\mathcal{H}_2$ yields

$$k_1 \sum_{\text{comp}} P_\infty(\mathcal{H}_{12}^\star, K_\epsilon) \;\geq\; \sum_{\text{T3, comp}} P_M(\mathcal{H}_{12}, K_\epsilon) \;\geq\; k_2 \sum_{\text{comp}} P_\infty(\mathcal{H}_{12}^\star, K_\epsilon)$$

with equality if there are no histories $\mathcal{H}_{12}$ and $\mathcal{H}_{12}^\star$ that fulfill the constraints in K_ϵ. Next, this results implies that

$$\begin{aligned} k_1\, P(\mathcal{H}_1)P(\mathcal{H}_2) \;&\geq\; P_M(\mathcal{H}_1, \mathcal{H}_2, K_\epsilon, \mathrm{T3}) \\ &\geq\; k_2\left[P(\mathcal{H}_1)P(\mathcal{H}_2) - P_\infty(\mathcal{H}_1, \mathcal{H}_2, K_\epsilon^c)\right]. \end{aligned}$$

As a consequence M_ϵ can be chosen such that

$$P_M(K_\epsilon, \mathrm{T3}) \;>\; 1-2\epsilon$$

for $M > M_\epsilon$. This completes the proof. □

Proof of Lemma 3. Note that because the number of nucleotides L is fixed, any function that does not take the value $\pm\infty$ is bounded. Consider a function $f(\mathrm{Data}_j)$ of the Data_j and the average over all l regions

$$F_l(\mathrm{Data}) \;=\; \frac{1}{l}\sum_{j=1}^{l} f(\mathrm{Data}_j).$$

It is to be proven that $F_l(\mathrm{Data})$ converges for all bounded functions. To do so it will be shown that the variance of $F_l(\mathrm{Data})$ converges to zero as $l \to \infty$. Now

$$\begin{aligned} Var[F_l(\mathrm{Data})] = \frac{1}{l}E[f(\mathrm{Data}_1)^2] \\ +\frac{2}{l^2}\sum_{i<j} E[f(\mathrm{Data}_i)f(\mathrm{Data}_j)] - E[f(\mathrm{Data}_1)]^2, \end{aligned} \tag{31}$$

using stationarity of the process. The first term converges to zero.

Let K_ϵ be chosen as in Lemma 2. Then

$$E[f(\text{Data}_i)f(\text{Data}_j)] = \sum \int_{K_\epsilon, T3} f(\text{Data}_i)f(\text{Data}_j)P_M(\text{Data}_i, \text{Data}_j|\mathcal{T}_{12}) \times P_M(\mathcal{T}_{12})\mathrm{d}\mathbf{t} \pm 2\epsilon f_{\max} = \sum \int_{K_\epsilon, T3} f(\text{Data}_i)f(\text{Data}_j) \times P(\text{Data}_i|\mathcal{T}_1)P(\text{Data}_j|\mathcal{T}_2)P_M(\mathcal{T}_{12})\mathrm{d}\mathbf{t} \pm 2\epsilon f_{\max}, \quad (32)$$

where the sum is over all possible observations, $f_{\max}$ is the maximum absolute value $f(\text{Data}_i)$ can obtain, and $\mathcal{T}_1$ and $\mathcal{T}_2$ are the marginal histories of locus 1 and 2, extracted from the joint history $\mathcal{T}_{12}$.

Using the the inequalities (27) and (28) provides the following upper and lower bound to the sum in Eq. (32). Upper bound:

$$k_1 \sum \int_{K_\epsilon} f(\text{Data}_i)f(\text{Data}_j)P(\text{Data}_i|\mathcal{T}_1)P(\text{Data}_j|\mathcal{T}_2)P_\infty(\mathcal{T}_{12}^\star)\mathrm{d}\mathbf{t},$$

and lower:

$$k_2 \sum \int_{K_\epsilon} f(\text{Data}_i)f(\text{Data}_j)P(\text{Data}_i|\mathcal{T}_1)P(\text{Data}_j|\mathcal{T}_2)P_\infty(\mathcal{T}_{12}^\star)\mathrm{d}\mathbf{t},$$

where k_1, k_2, and $\mathcal{T}_{12}^\star$ are as in the proof of Lemma 3 (k_1 and k_2 are defined in Eqs. (29) and (30), respectively), and $\mathcal{T}_i$, $i = 1,2$ are the marginal histories extracted from $\mathcal{T}_{12}^\star$. For corresponding histories $\mathcal{T}_{12}$ and $\mathcal{T}_{12}^\star$, the marginal histories are identical.

Regarding the upper bound. Integrating over all possible histories (instead of over K_ϵ) yields the bound

$$k_1 E[f(\text{Data}_i)]E[f(\text{Data}_j)]. \quad (33)$$

Regarding the lower bound. Integrating over all possible histories yields the bound

$$k_2 E[f(\text{Data}_i)]E[f(\text{Data}_j)] - k_2 f_{\max}\epsilon. \quad (34)$$

Note that k_1 and k_2 are $1 + O(1/M)$, where $O(1/M)$ depends on the chosen ϵ (see proof of Lemma 2). Inserting Eqs. (33) and (34) into Eq. (31) shows that the variance can be made arbitrary small by first choosing $\epsilon > 0$ sufficiently small and then M sufficiently large. Note that there is a term depending on $\log(M)/M$ that converges to zero for large M. The proof is completed. □

Proof of Theorem 1. With the assumptions made in this paper the following is also true: The regularity assumptions in Peskir [27], Sect. 2; the conditions in Peskir [27], Lemma 1; and Eqs. (7)–(9), p. 307 in Peskir [27] with $\Gamma = A$ (in Peskir's notation). Hence $\hat{M} \subseteq M$ (in Peskir's notation) and the theorem follows from Theorem 1, Eqs. (2) and (3), in Peskir [27]. □

References

1. Adams, M., Hudson, R.R.: Maximum-likelihood estimation of demographic parameters using the frequency spectrum of unlinked single-nucleotide polymorphisms. Genetics **168,** 1699–1712 (2004)
2. Cox, D.R., Reid, N.: A note on pseudolikelihood constructed from marginal densities. Biometrika **91,** 729–737 (2004)
3. Fearnhead, P.: Consistency of estimators of the population-scaled recombination rate. Theor. Pop. Biol. **64,** 67–79 (2003)
4. Fearnhead, P., Donnelly, P.: Estimating recombination rates from population genetic data. Genetics **159,** 1299–1318 (2001)
5. Fearnhead, P., Donnelly, P.: Approximate likelihood methods for estimating local recombination rates. P. J. Roy. Stat. Soc. B **64,** 657–680 (2002)
6. Felsenstein, J.: Inferring Phylogenies. Sinauer Associates, Sunderland, (2003)
7. Goldman, N., Yang, Z.: A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol. Biol. Evol. **11,** 725–736 (1994)
8. Griffiths, R.C., Marjoram, P.: Ancestral inference from samples of DNA sequences with recombination. J. Comput. Biol. **3,** 479–502 (1996)
9. Griffiths, R.C., Marjoram, P.: An ancestral recombination graph. In: Donnelly, P., Tavaré, S. (eds.) Progress in Population Genetics and Human Evolution, IMA, Volumes in Mathematics and its Applications, Vol. 87, pp. 257–270 Springer, Berlin Heidelberg New York (1997)
10. Griffiths, R.C., Tavaré, S.: Simulating probability distributions in the coalescent. Theor. Pop. Biol. **46,** 131–159 (1994)
11. Griffiths, R.C., Tavaré, S.: Sampling theory for neutral alleles in varying environment. Phil. Trans. R. Soc. Lond. B **344,** 403–410 (1994)
12. Griffiths, R.C., Tavaré, S.: Markov chain inference methods in population genetics. Math. Comput. Modelling **23,** (8/9), 141–158 (1996)
13. Griffiths, R.C., Tavaré, S.: Computational methods for the coalescent. In: Donnelly, P., Tavaré, S. (eds.) Progress in Population Genetics and Human Evolution, IMA Volumes in Mathematics and its Applications, Vol. 87, pp. 165–182. Springer, Berlin Heidelberg New York (1997)
14. Hein, J., Schierup, M., Wiuf, C.: Gene Genealogies, Variation, and Evolution. Oxford University Press (2005)
15. Hudson, R.R.: Properties of the neutral alelle model with intergenic recombiantion. Theor. Pop. Biol. **23,** 183–201 (1983)
16. Hudson, R.R.: Gene genealogies and the coalescent process. Oxford Surv. Evol. Biol. **7,** 1–47 (1991)
17. Hudson, R.R.: Two-locus sampling distributions and their application. Genetics **159,** 1805–1817 (2001)
18. Kim, Y., Stephan, W.: Joint effects of genetic hitchhiking and background selection on neutral variation. Genetics **155,** 1414–1427 (2002)
19. Kingman, J.F.C.: The coalescent. Stoch. Appl. Process. **13,** 235–248 (1982)
20. Kuhner, M.K., Yamato, J., Felsenstein, J.: Maximum likelihood estimation of recombination rates from population data. Genetics **156,** 1393–1401 (2000)
21. Marth, G., Czabarka, E., Murvai, J., Sherry, S.T.: The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. Genetics **166,** 351–372 (2004)
22. McVean, G., Awadalla, P., Fearnhead, P.: A coalescent-based method for detecting and estimating recombination from gene sequences. Genetics **160,** 1231–1241 (2002)
23. Nielsen, R.: Estimation of population parameters and recombination rates from single nucleotide polymorphisms. Genetics **154,** 931–942 (2000)

24. Nielsen, R., Wakeley, J.: Distinguishing migration from isolation: a Markov chain Monte Carlo approach. Genetics **158,** 885–896 (2001)
25. Nielsen, R., Wiuf, C.: Composite likelihood estimation applied to single nucleotide polymorphism (SNP) data. ISI Conference Proceedings (2005)
26. Nordborg, M.: Coalescent theory. In: Balding, D.J., Bishop, M.J., Cannings, C. (eds.) Handbook of Statistical Genetics. pp. 179–212 J Wiley, New York (2001)
27. Peskir, G.: Consistency of statistical models in the stationary case. Math. Scand. **78,** 293–319 (1996)
28. Serfling, R.J.: Approximation Theorems of Mathematical Statistics. J. Wiley, New York (1980)
29. Stephens, M., Donnelly, P.: Inference in molecular population genetics. J. Roy. Stat. Soc. B **62,** 605–655 (2000)
30. White, H.: Maximum likelihood estimation of misspecified models. Econometrica **50,** 1–26 (1982)
31. Wiuf C.: A coalescence approach to gene conversion. Theor. Pop. Biol. **57,** 357–367 (2000)
32. Wiuf, C., Hein, J.: The coalescent with gene conversion. Genetics **155,** 451–462 (2000)