

**Institute of Mathematical Statistics
LECTURE NOTES-MONOGRAPH SERIES**

**Statistics in Molecular Biology
and Genetics**

Selected Proceedings of a 1997 Joint AMS-IMS-SIAM Summer Conference on Statistics
in Molecular Biology

Francoise Seillier-Moiseiwitsch, Editor

Volume 33

**Published by the Institute of Mathematical Statistics
and the American Mathematical Society**



Institute of Mathematical Statistics
LECTURE NOTES–MONOGRAPH SERIES
Volume 33

**Statistics in Molecular Biology
and Genetics**

**Selected Proceedings of a 1997 Joint AMS-IMS-SIAM
Summer Conference on Statistics in Molecular Biology**

Francoise Seillier-Moiseiwitsch, Editor

American Mathematical Society
Providence, Rhode Island

Institute of Mathematical Statistics
Hayward, California

Institute of Mathematical Statistics

Lecture Notes-Monograph Series

Editorial Board

Andrew A. Barbour, Joseph Newton, and David Ruppert (Editor)

The production of the *IMS Lecture Notes-Monograph Series* is managed by the IMS Business Office: Julia A. Norton, IMS Treasurer, and Elyse Gustafson, IMS Business Manager.

This volume was co-published with the American Mathematical Society

Library of Congress Catalog Card Number: 99-076060

International Standard Book Number 0-940600-47-1

Copyright © 1999 Institute of Mathematical Statistics

All rights reserved

Printed in the United States of America

TABLE OF CONTENTS

Preface	v
	<i>F. Seillier-Moiseiwitsch</i>

Genetic Mechanisms

On a Markov Model for Chromatid Interference	1
	<i>H. Zhao and T. Speed</i>

Population Genetics

Some Statistical Aspects of Cytonuclear Disequilibria	21
	<i>S. Datta</i>
Diffusion Process Calculations for Mutant Genes in Nonstationary Populations	38
	<i>R. Fan and K. Lange</i>
The Coalescent with Partial Selfing and Balancing Selection: An Application of Structured Coalescent Processes	56
	<i>M. Nordborg</i>

Human Genetics

Statistical Aspects of the Transmission/Disequilibrium Test (TDT)	77
	<i>W. Ewens</i>
Estimation of Conditional Multilocus Gene Identity among Relatives	95
	<i>E. Thompson and S. Heath</i>

Quantitative Genetics

A Review of Methods for Identifying QTL's in Experimental Crosses	114
	<i>K. Broman and T. Speed</i>

Evolutionary Genetics

Markov Chain Monte Carlo for the Bayesian Analysis of Evolutionary Trees from Aligned Molecular Sequences	143
	<i>M. Newton, B. Mau and B. Larget</i>
Likelihoods on Coalescents: A Monte Carlo Sampling Approach to Inferring Parameters from Population Samples of Molecular Data	163
	<i>J. Felsenstein, M. Kuhner, J. Yamato and P. Beerli</i>

Uses of Statistical Parsimony in HIV Analyses	186
<i>K. Crandall</i>	
Linear Estimators for the Evolution of Transposable Elements	207
<i>P. Joyce, L. Fox, N. Casavant and H. Wichman</i>	
A Conditional Approach to the Detection of Correlated Mutations	221
<i>M. Karnoub, F. Seillier-Moiseiwitsch and P.K. Sen</i>	
Correlated Mutations in Protein Sequences: Phylogenetic and Structural Effects	236
<i>A. Lapedes, B. Giraud, L. Liu and G. Stormo</i>	

Sequence Motifs

Compound Poisson Approximations for Occurrences of Multiple Words ...	257
<i>G. Reinert and S. Schbath</i>	

Protein Structure

Deriving Interatomic Distance Bounds from Chemical Structure	276
<i>M. Trosset and G. Phillips</i>	
Protein Fold Class Prediction is a New Field for Statistical Classification and Regression	288
<i>L. Edler and J. Grassmann</i>	

PREFACE

The papers in this volume were presented at one of the 1997 Summer Research Conferences in the Mathematical Sciences jointly sponsored by the Institute of Mathematical Statistics, the American Mathematical Society and the Society for Industrial and Applied Mathematics. The theme of the meeting was "Statistics in Molecular Biology and Genetics". That this volume is published jointly by the Institute of Mathematical Statistics and the American Mathematical Society reflects the emerging importance of Statistics in these fields.

These papers fall into broad categories: population genetics, evolutionary genetics, protein structure, genetic mechanisms, quantitative genetics, human genetics and sequence motifs. While some of these areas have a long history of statistical input (and have motivated some mainstream statistical developments), others are new statistical applications. The talks by Professors D. Botstein, M.-C. King and M. Olson underlined the great need for statistical expertise in cutting-edge biological technology. Their stimulating presentations treated us with wonderfully clear overviews of current directions in important areas of genetic research (namely, physical mapping, genetic mapping and functional genetics).

The manuscripts underwent a rigorous review process: each was scrutinized by two anonymous referees. For their critical reviews, my gratitude goes to: D. Balding, L. Edler, W. Ewens, J. Felsenstein, J. Hein, P. Joyce, A. Kong, M. Kuhner, K. Lange, A. Lapedes, M. Man, P. Marjoram, K. Mengersen, M.-S. McPeek, M. Moehle, D. Nelson, M. Nordborg, I. Painter, A. Pluzhnikov, A. Ramaswami, G. Reinert, S. Sawyer, M. Stephens, E. Thompson and M. Trosset.

I particularly want to thank Professors M. Waterman and P. Donnelly for their help in organizing the conference. Financial support for the meeting was provided by the National Science Foundation and the National Institutes of Health. Professor P. Donnelly deserves a double thank-you for dealing with the papers with which I had a conflict of interest. I extend my thanks to the anonymous referees he selected.

Finally, I am enormously indebted to R. Budrevich for his tireless help in the preparation of this volume.

F. Seillier-Moiseiwitsch
Chapel Hill

ON A MARKOV MODEL FOR CHROMATID INTERFERENCE

BY HONGYU ZHAO AND TERENCE P. SPEED

Yale University School of Medicine and University of California, Berkeley

Meiotic exchange between homologous chromosomes takes place after the formation of a bundle of four chromatids. Crossovers are precise breakage-and-reunion events. Random strand involvements (no chromatid interference) and random distribution of crossovers (no chiasma interference) are usually assumed in analyzing genetic data. In this paper, we discuss a Markov model for chromatid interference. Closed form expression for the probability of any multilocus recombination/tetrad pattern is derived. Both chromatid interference and chiasma interference can be studied together using this model. In particular, we discuss chi-square models for chiasma interference.

1. Introduction. In diploid cells, each chromosome is paired with its homologue during meiosis. Each member of a given homologous pair has two identical sister chromatids, so that each synapsed paired structure consists of four chromatids. Usually one or more crossovers occur among the four chromatids. A crossover is a precise breakage-and-reunion event occurring between two nonsister chromatids.

The types of genetic data considered here are single spore data, in which the products of a single meiosis are recovered separately, and tetrad data, in which all four meiotic products are recovered together. A tetrad consists of four spores, each of which is haploid, encased in a structure called an ascus. In some organisms, such as *Neurospora crassa* (red bread mold), tetrads consist of four spores in a linear order corresponding to the meiotic divisions; these are called ordered tetrads. In other organisms, such as *Saccharomyces cerevisiae* (baker's yeast), the four spores are produced as a group without order and are called unordered tetrads. Griffiths et al. (1996) covers relevant genetic background.

In this paper, genes (markers, loci) are denoted by script letters. For example, we use \mathcal{A} and \mathcal{B} to denote two genes. Alleles of \mathcal{A} are denoted by A and a , while alleles of \mathcal{B} are denoted by B and b . Suppose that \mathcal{A} and \mathcal{B} are on the same chromosome arm, and consider a diploid cell having AB and ab on homologous chromosomes. There are four possible products at these loci resulting from meiosis of this cell, namely, AB , ab , Ab , and aB . The first two are called

Supported in part by NIH grant HG-01093.

AMS 1991 subject classifications. Primary 60J20; secondary 60P10.

Key words and phrases. Chromatid interference, chiasma interference, Markov model, chi-square model, single spore data, tetrad data.

parental types or nonrecombinants, the other two types, Ab and aB , are called recombinants. If two markers are recombined by crossovers in a meiotic product, then during meiosis an odd number of crossovers must have occurred between the two markers on the strand carrying them. The proportion of recombinants, r_{AB} , is called the recombination fraction. Because recombination fractions are not additive, genetic (or map) distance is used as an additive measure of distance between loci. Genetic distance between two markers is defined as the average number of crossovers per strand per meiosis between these markers. The unit of genetic distance is Morgan (M). Two markers are 1M apart if on average there is one crossover occurring on a single strand per meiosis between these two markers. In practice, centiMorgan (cM = 0.01M) is more commonly used in genetic mapping.

The occurrence of crossovers cannot be observed directly and has to be inferred from observed recombination events. In the case of single spore data, a given meiotic product may be scored as recombinant or non-recombinant for each pair of markers. The map distance between two markers can be estimated from the observed recombination fraction. In the case of unordered tetrad data, there are three possible observed outcomes for each pair of markers: parental ditype when all four strands are non-recombinants, tetratype when exactly two of the four strands show recombination between the two markers, and nonparental ditype when all four strands are recombinants. The map distance between two markers can be estimated from the observed proportions of these three tetrad types. In the case of ordered tetrads bearing one marker A with alleles A and a, there are six distinguishable configurations:

1	2	3	4	5	6
A	a	a	A	A	a
A	a	A	a	a	A
a	A	a	A	a	A
a	A	A	a	A	a

Configurations 1 and 2 are called first division segregation (FDS) patterns and configurations 3 to 6 are called second division segregation (SDS) patterns. Because of random spindle to centromere attachments during meiosis, configurations 1 and 2 have the same probability, and the four configurations showing SDS pattern also have the same probability (Griffiths et al. 1996). The map distance between a marker and the centromere can then be estimated from the observed SDS proportion.

To estimate map distance from the observed data, a model is needed which connects the process of crossover to the observed outcomes. Any model has to

consider two aspects of crossover that are relevant to the observed recombination outcome: the distribution of crossover events along the bundle of four chromatids, and the pairs of nonsister chromatids involved in crossovers. To distinguish crossover events occurring on the four strand bundle and crossover events on single strands, we describe crossover events on the four strand bundle as chiasmata (singular: chiasma), and crossover events on single strands as crossovers. Chiasma interference refers to non-random distribution of chiasmata on the four strand bundle, whereas crossover interference refers to non-random distribution of crossovers along single strands. In this paper, we use the word *random* in the sense that all outcomes are equally likely. There is no chromatid interference (NCI) if any pair of non-sister chromatids are equally likely to be involved in any chiasma, independent of which pairs were involved in other chiasmata. It is possible that there is chiasma interference at the four strand bundle, but because of the presence of chromatid interference, there is no crossover interference on single strands, see Zhao and Speed (1996) for a model exhibiting this phenomenon.

With virtually no restrictions on the chiasma process, Speed, McPeek and Evans (1992) derived the constraints on multilocus recombination probabilities for single spore data under the assumption of NCI. It was shown by Zhao, McPeek and Speed (1995) that the assumption of NCI also imposes constraints on multilocus tetrad probabilities. Based on these constraints, a statistical testing procedure for NCI was proposed in the last mentioned paper and applied to data from several organisms. Though there was an excess of two-strand double recombinations in some organisms, no strong evidence was found for chromatid interference.

Crossover interference has been observed in almost all organisms. The presence of one crossover usually inhibits the formation of crossovers in a nearby region. Fisher, Lyon and Owen (1947) modeled crossovers as a renewal process; that is, crossovers along a single strand were assumed to be formed as a regular sequence starting from the centromere, with the length between two adjacent crossovers always following the same distribution. Centromeres are constricted regions of nuclear chromosomes, to which the spindle fibers attach during division. Mather (1938) appeared to be the first one to propose the above sequential model for the chiasma process. Other mathematical models have also been proposed to model crossover interference (McPeek and Speed 1995). However, the biological nature of crossover interference is still not well understood.

In this paper, we discuss a Markov model for chromatid interference. There exist closed form expressions for joint recombination and tetrad probabilities under this model. Closed form expressions still exist when this model is combined with a class of chiasma interference models, the chi-square model (Zhao, Speed

and McPeek 1995), thus allowing joint modeling of both types of interference.

2. A Markov model for chromatid interference. The Markov chromatid interference model discussed here was first introduced and studied by Weinstein (1938). Later studies on chromatid interference (Carter and Robertson 1952, Sturt and Smith 1976 and Stam 1979) essentially used Weinstein's model.

Weinstein's model assumes that chiasmata occur according to a point process originating from the centromere, and that the choice of nonsister chromatids involved in one chiasma only depends on the pair involved in the previous chiasma. Thus it is a Markov model. The two pairs of sister chromatids are labeled as w^1, w^2 and m^1, m^2 . If w^1 and m^1 are involved in the k th chiasma, the chances of the four nonsister pairs (w^1, m^1) , (w^1, m^2) , (w^2, m^1) , and (w^2, m^2) being involved in the $(k+1)$ th chiasma are denoted by α , β , β , and γ , respectively. That is, the conditional probabilities of two, three, and four-strand double chiasmata are α , 2β , and γ for any two consecutive chiasmata. Under this model, if a strand is involved in one chiasma, the chance that it will be involved in the next chiasma is $\eta = \alpha + \beta$, and the chance that it will not be involved is $\beta + \gamma = 1 - \eta$. So for a single strand, the degree of chromatid interference is determined by η . Different values of $(\alpha, 2\beta, \gamma)$ can correspond to the same η . When there is no chromatid interference, $(\alpha, 2\beta, \gamma) = (\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$ and $\eta = \frac{1}{2}$, but any chromatid interference model with parameters $(\alpha, 2\beta, \gamma)$ satisfying $\alpha + \beta = \frac{1}{2}$ will be indistinguishable from a no chromatid interference model if only single spore data are available. On the other hand, different sets of $(\alpha, 2\beta, \gamma)$ that correspond to the same η value can be distinguished using tetrad data.

Previous work on this chromatid interference model has been confined to the two-locus case. Our aim is to derive general expressions for the probabilities of recombination or not, across a set of loci, and the analogous multilocus tetrad probabilities. We consider single spore data and tetrad data separately.

2.1 Single spore data. We assign a state “ y ” or “ n ” to any locus on any one of the four strands in a bundle as follows. If there has been at least one chiasma between the centromere and that locus, then a (locus, strand) pair is assigned state “ y ” if the strand was involved in the last chiasma before the locus; otherwise the (locus, strand) pair is assigned “ n ”. If no chiasmata have occurred between a locus and the centromere, it can be shown that assigning the state “ y ” or “ n ” with probability $\frac{1}{2}$ fits in well with our development.

Suppose that \mathcal{A} and \mathcal{B} are two loci on the same chromosome in the order $CEN - \mathcal{A} - \mathcal{B}$. For any given strand, there are four possible joint states, (y, y) , (y, n) , (n, y) and (n, n) . Consider the case in which \mathcal{A} is in state y and that $k \geq 1$ chiasmata have occurred between \mathcal{A} and \mathcal{B} . A given strand could have

been involved in an odd (o) or an even (e) number of the chiasmata occurring between \mathcal{A} and \mathcal{B} , i.e., \mathcal{A} and \mathcal{B} may have recombined or not, on that strand. Given \mathcal{A} is in state y on a strand, let $p_{y,y}^o(k)$ be the conditional probability that \mathcal{B} is in the state y and that an odd number of the $k \geq 1$ chiasmata between \mathcal{A} and \mathcal{B} involve this strand. Define $p_{y,n}^o(k)$, $p_{y,y}^e(k)$, and $p_{y,n}^e(k)$ in similar fashion, where e denotes an even number. If \mathcal{A} is in state n , $p_{n,y}^o(k)$, $p_{n,n}^o(k)$, $p_{n,y}^e(k)$, and $p_{n,n}^e(k)$ can be similarly defined. Group $p_{y,y}^o(k)$, $p_{y,n}^o(k)$, $p_{n,y}^o(k)$, and $p_{n,n}^o(k)$ into a 2×2 matrix, define

$$\mathbf{T}_k^1 = \begin{pmatrix} p_{y,y}^o(k) & p_{y,n}^o(k) \\ p_{n,y}^o(k) & p_{n,n}^o(k) \end{pmatrix},$$

and similarly,

$$\mathbf{T}_k^0 = \begin{pmatrix} p_{y,y}^e(k) & p_{y,n}^e(k) \\ p_{n,y}^e(k) & p_{n,n}^e(k) \end{pmatrix}.$$

Recursive relationships among these quantities can be established as follows. Suppose that \mathcal{B} is in state y on one strand, and that \mathcal{A} and \mathcal{B} are recombined on that strand after k chiasmata have taken place on the bundle between \mathcal{A} and \mathcal{B} . If a $(k+1)$ th chiasma on the bundle occurs between \mathcal{A} and \mathcal{B} , then the strand has a chance η of being involved in it, and $1 - \eta$ of not being involved. In the first case, \mathcal{B} will remain in state y and \mathcal{A} and \mathcal{B} will not be recombined on that strand; in the second case, \mathcal{B} will change to state n and \mathcal{A} and \mathcal{B} will still be recombined on that strand. Other cases can be considered similarly, and we can thus derive the following relationship:

$$\begin{pmatrix} p_{y,y}^o(k+1) \\ p_{y,n}^o(k+1) \\ p_{y,y}^e(k+1) \\ p_{y,n}^e(k+1) \end{pmatrix} = \begin{pmatrix} 0 & 0 & \eta & 1 - \eta \\ 1 - \eta & \eta & 0 & 0 \\ \eta & 1 - \eta & 0 & 0 \\ 0 & 0 & 1 - \eta & \eta \end{pmatrix} \begin{pmatrix} p_{y,y}^o(k) \\ p_{y,n}^o(k) \\ p_{y,y}^e(k) \\ p_{y,n}^e(k) \end{pmatrix}.$$

Similarly, if \mathcal{A} is in state n , with $p_{n,y}^o(k)$, $p_{n,n}^o(k)$, $p_{n,y}^e(k)$, and $p_{n,n}^e(k)$ defined as above, similar recursive relationship can be established.

Let $p_{\mathcal{A}}^y$ (or $p_{\mathcal{A}}^n$) be the probability that \mathcal{A} is in state y (or n) on a given strand and $s_{\mathcal{A}}$ and $s_{\mathcal{B}}$ be the states at \mathcal{A} and \mathcal{B} . From the definitions of the $p_{s_{\mathcal{A}}, s_{\mathcal{B}}}^r(k)$, after $k \geq 1$ chiasmata have occurred along the four-strand bundle between \mathcal{A} and \mathcal{B} , the chance that \mathcal{A} and \mathcal{B} are recombined on that strand is

$$(p_{\mathcal{A}}^y, p_{\mathcal{A}}^n) \mathbf{T}_k^1 \begin{pmatrix} 1 \\ 1 \end{pmatrix},$$

and the chance that \mathcal{A} and \mathcal{B} are not recombined on that strand is:

$$(p_{\mathcal{A}}^y, p_{\mathcal{A}}^n) \mathbf{T}_k^0 \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

Let $\delta = 2\eta - 1$, define

$$\mathbf{M} = \frac{1}{2} \begin{pmatrix} 1 + \delta^k & 1 - \delta^k \\ 1 - \delta^k & 1 + \delta^k \end{pmatrix},$$

$$\mathbf{N}_0 = \begin{pmatrix} -\delta^{\frac{k}{2}} & 0 \\ 0 & -\delta^{\frac{k}{2}} \end{pmatrix} \quad \text{and} \quad \mathbf{N}_1 = \begin{pmatrix} \eta\delta^{\frac{k-1}{2}} & -(1-\eta)\delta^{\frac{k-1}{2}} \\ (1-\eta)\delta^{\frac{k-1}{2}} & -\eta\delta^{\frac{k-1}{2}} \end{pmatrix}.$$

General expressions of \mathbf{T}_k^1 and \mathbf{T}_k^0 can be obtained through the above recursive relationships among the $p_{s_{\mathcal{A}}, s_{\mathcal{B}}}^r(k)$ as summarized in the following theorem. The proof is given in section 6.

Theorem 1. *When k is even,*

$$\mathbf{T}_k^1 = \frac{1}{2}(\mathbf{M} + \mathbf{N}_0) \quad \text{and} \quad \mathbf{T}_k^0 = \frac{1}{2}(\mathbf{M} - \mathbf{N}_0),$$

and when k is odd,

$$\mathbf{T}_k^1 = \frac{1}{2}(\mathbf{M} + \mathbf{N}_1) \quad \text{and} \quad \mathbf{T}_k^0 = \frac{1}{2}(\mathbf{M} - \mathbf{N}_1).$$

For a strand chosen at random, \mathcal{A} has an equal chance to be in state y or n . The probability that \mathcal{A} and \mathcal{B} are recombined is $\frac{1}{2}(1 - \delta^{\frac{k}{2}})$ for k even and $\frac{1}{2}$ for k odd.

This result was proved by Weinstein (1938) and Sturt and Smith (1976). Our approach, however, is different, and easily generalizes to the multilocus case, as illustrated later. We now consider three special cases: (a) $\eta = \frac{1}{2}$, (b) $\eta = 0$, and (c) $\eta = 1$. Case (a) includes no chromatid interference. Case (b) implies that a strand is never involved in two consecutive chiasmata, i.e., that only four-strand double chiasmata occur. In case (c), all chiasmata involve the same two strands, i.e., only two-strand double chiasmata occur. We calculate \mathbf{T}_k^1 for each case. Case (a):

$$\mathbf{T}_k^1 = \begin{pmatrix} \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} \end{pmatrix}.$$

Case (b): when k is even,

$$\mathbf{T}_k^1 = \begin{pmatrix} \frac{1}{2}(1 - (-1)^{\frac{k}{2}}) & 0 \\ 0 & \frac{1}{2}(1 - (-1)^{\frac{k}{2}}) \end{pmatrix},$$

and when k is odd,

$$\mathbf{T}_k^1 = \begin{pmatrix} 0 & \frac{1}{2}(1 - (-1)^{\frac{k-1}{2}}) \\ \frac{1}{2}(1 + (-1)^{\frac{k-1}{2}}) & 0 \end{pmatrix}.$$

Case (c): when k is even,

$$\mathbf{T}_k^1 = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix},$$

and when k is odd,

$$\mathbf{T}_k^1 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}.$$

The recombination fraction is: case (a) $r = \frac{1}{2}$ for $k \geq 1$; case (b) $r = \frac{1}{2}(1 - (-1)^{\frac{k}{2}})$ for k even and $\frac{1}{2}(1 - (-1)^{\frac{k-1}{2}})$ for k odd; and case (c) $r = 0$ for k even and 1 for k odd.

Suppose q_k is the probability that there are k chiasmata between \mathcal{A} and \mathcal{B} , and let

$$\mathbf{T}_{AB}^1 = \sum q_k \mathbf{T}_k^1 = \begin{pmatrix} p_{y,y}^o & p_{y,n}^o \\ p_{n,y}^o & p_{n,n}^o \end{pmatrix}.$$

Then given that \mathcal{A} is in state $s_{\mathcal{A}}$, the value of $p_{s_{\mathcal{A}}, s_{\mathcal{B}}}^o$ is the conditional probability that \mathcal{B} is in state $s_{\mathcal{B}}$ and that \mathcal{A} and \mathcal{B} are recombined. We can similarly define

$$\mathbf{T}_{AB}^0 = \sum q_k \mathbf{T}_k^0 = \begin{pmatrix} p_{y,y}^e & p_{y,n}^e \\ p_{n,y}^e & p_{n,n}^e \end{pmatrix}.$$

Define $p_{\mathcal{A}}^y$ and $p_{\mathcal{A}}^n$ as above, the chance that \mathcal{A} and \mathcal{B} are recombined is

$$(p_{\mathcal{A}}^y, p_{\mathcal{A}}^n) \mathbf{T}_{AB}^1 \begin{pmatrix} 1 \\ 1 \end{pmatrix},$$

and the chance that they are not recombined is

$$(p_{\mathcal{A}}^y, p_{\mathcal{A}}^n) \mathbf{T}_{\mathcal{AB}}^0 \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

The recombination probabilities can be computed explicitly for some simple chiasma processes. Consider the Poisson chiasma process model. Under this model, the number of chiasmata between \mathcal{A} and \mathcal{B} is Poisson distributed, $q_k = e^{-2d}(2d)^k/k!$, where d is the genetic distance between \mathcal{A} and \mathcal{B} . There is a factor 2 because each chiasma involves two of the four chromatids, thus the average number of crossovers on a single strand is half the number of chiasmata on the four strand bundle. The matrix $\mathbf{T}_{\mathcal{AB}}^1$ is

$$\frac{1}{4} \begin{pmatrix} x & y - z \\ y + z & x \end{pmatrix},$$

where

$$x = 1 + e^{-4d(1-\eta)} - \frac{(\eta + \phi)e^{-2d(1+\phi)} - (\eta - \phi)e^{-2d(1-\phi)}}{\phi},$$

$$y = 1 - e^{-4d(1-\eta)},$$

$$z = \frac{(1 - \eta)(e^{-2d(1-\phi)} - e^{-2d(1+\phi)})}{\phi},$$

and $\phi = \sqrt{\delta}$.

Therefore, for any ϕ , the recombination probability r is $\frac{1}{4}\{2 - (e^{-2d(1+\phi)} + e^{-2d(1-\phi)})\}$. Consider the above-mentioned three special cases. For case (a), r is $\frac{1}{2}(1 - e^{-2d})$; for case (b), r is $\frac{1}{2}\{1 - \cos(2d)e^{-2d}\}$; and for case (c), r is $\frac{1}{4}(1 - e^{-4d})$.

Now consider three markers, \mathcal{A}_1 , \mathcal{A}_2 , and \mathcal{A}_3 . Given \mathcal{A}_1 is in state s_1 , and there are k_1 and k_2 chiasmata in the two intervals, denote the probability that \mathcal{A}_3 is in state s_3 and that the strand is involved in an odd number of chiasmata in both intervals (both intervals show recombination) as $p_{s_1, s_3}^{o,o}(k_1, k_2)$. Define the matrix $\mathbf{T}_{\mathcal{A}_1 \mathcal{A}_3}^{1,1}(k_1, k_2)$ as

$$\mathbf{T}_{\mathcal{A}_1 \mathcal{A}_3}^{1,1}(k_1, k_2) = \begin{pmatrix} p_{y,y}^{o,o}(k_1, k_2) & p_{y,n}^{o,o}(k_1, k_2) \\ p_{n,y}^{o,o}(k_1, k_2) & p_{n,n}^{o,o}(k_1, k_2) \end{pmatrix}.$$

Considering the state of \mathcal{A}_2 , denoted by s_2 , and using the Markovian property of the model (that the choice of strands involved in the next chiasma depends only on the pair involved in the previous one), we obtain

$$p_{y,y}^{o,o}(k_1, k_2) = p_{y,s_2=y}^o(k_1)p_{s_2=y, y}^o(k_2) + p_{y,s_2=n}^o(k_1)p_{s_2=n, y}^o(k_2).$$

Similar relationships hold for the other $p_{s_1, s_3}^{o, o}(k_1, k_2)$. Writing this in matrix form, we have

$$\mathbf{T}_{\mathcal{A}_1 \mathcal{A}_3}^{1,1}(k_1, k_2) = \mathbf{T}_{\mathcal{A}_1 \mathcal{A}_2}^1(k_1) \mathbf{T}_{\mathcal{A}_2 \mathcal{A}_3}^1(k_2).$$

Other $\mathbf{T}_{\mathcal{A}_1 \mathcal{A}_3}^{i_1, i_2}(k_1, k_2)$ can be defined similarly, where i_1, i_2 is either 1 or 0 corresponding to whether two genes are recombined or not over that interval. We have

$$\mathbf{T}_{\mathcal{A}_1 \mathcal{A}_3}^{i_1, i_2}(k_1, k_2) = \mathbf{T}_{\mathcal{A}_1 \mathcal{A}_2}^{i_1}(k_1) \mathbf{T}_{\mathcal{A}_2 \mathcal{A}_3}^{i_2}(k_2).$$

Let $p_{\mathcal{A}_1 \mathcal{A}_3}^{i_1, i_2}(k_1, k_2)$ be the conditional probability that the recombination pattern for $\mathcal{A}_1 \mathcal{A}_2 \mathcal{A}_3$ is (i_1, i_2) given that there are k_1 and k_2 chiasmata in the two intervals. Then $p_{\mathcal{A}_1 \mathcal{A}_3}^{i_1, i_2}(k_1, k_2)$ can be calculated as:

$$p_{\mathcal{A}_1 \mathcal{A}_3}^{i_1, i_2}(k_1, k_2) = (p_{\mathcal{A}_1}^y, p_{\mathcal{A}_1}^n) \mathbf{T}_{\mathcal{A}_1 \mathcal{A}_2}^{i_1}(k_1) \mathbf{T}_{\mathcal{A}_2 \mathcal{A}_3}^{i_2}(k_2) \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

Let q_{k_1, k_2} be the chance of k_1 and k_2 chiasmata in the two intervals. Denote $p_{\mathcal{A}_1 \mathcal{A}_3}^{i_1, i_2}$ as the probability that the recombination pattern is (i_1, i_2) , we have

$$p_{\mathcal{A}_1 \mathcal{A}_3}^{i_1, i_2} = (p_{\mathcal{A}_1}^y, p_{\mathcal{A}_1}^n) \left(\sum q_{k_1, k_2} \mathbf{T}_{\mathcal{A}_1 \mathcal{A}_2}^{i_1}(k_1) \mathbf{T}_{\mathcal{A}_2 \mathcal{A}_3}^{i_2}(k_2) \right) \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

For any $l + 1$ markers, $\mathcal{A}_1, \mathcal{A}_2, \dots, \mathcal{A}_{l+1}$, an explicit expression exists for the probability of any recombination pattern $\mathbf{i} = (i_1, i_2, \dots, i_l)$, where $i_j = 0$, or 1 according to whether there is no recombination or recombination in the j th interval. Define $\mathbf{k} = (k_1, k_2, \dots, k_l)$ and let $q_{\mathbf{k}}$ be the probability of there being k_1, k_2, \dots, k_l chiasmata in these l intervals. The probability of recombination type \mathbf{i} is denoted $p_{\mathbf{i}}$. The general result for single spore data involving $l + 1$ markers is:

Theorem 2.

$$p_{\mathbf{i}} = (p_{\mathcal{A}_1}^y, p_{\mathcal{A}_1}^n) \left(\sum q_{\mathbf{k}} \mathbf{T}_{\mathcal{A}_1 \mathcal{A}_2}^{i_1}(k_1) \mathbf{T}_{\mathcal{A}_2 \mathcal{A}_3}^{i_2}(k_2) \cdots \mathbf{T}_{\mathcal{A}_l \mathcal{A}_{l+1}}^{i_l}(k_l) \right) \begin{pmatrix} 1 \\ 1 \end{pmatrix},$$

where the \mathbf{T} matrices are defined as for the two marker case, and the sum is over all $\mathbf{k} = (k_1, k_2, \dots, k_l)$.

Given the joint chiasma probabilities on the four strand bundle across studied intervals, $q_{\mathbf{k}}$, Theorem 2 allows us to obtain a closed form expression for any joint recombination probability.

2.2. Ordered tetrad data. For the simplicity of our discussion, we assume for the moment that the tetrads are ordered. Unordered tetrads will be discussed in the next subsection. For a tetrad ordered from top to bottom, marker \mathcal{A} with two alleles A and a can have six distinguishable types as illustrated in the introduction section.

For the moment, assume that there is at least one chiasma between the centromere and \mathcal{A} . Then for each type, tetrads can be further divided into subclasses according to the two strands involved in the last chiasma before \mathcal{A} . Each possible pair can be represented by (l_w, l_m) , where $l_w = 0$ if the top one of the two strands bearing A was involved in the last chiasma, $l_w=1$ otherwise, and, $l_m=0$ if the top strand bearing a was involved in the last chiasma, $l_m=1$ otherwise. Thus, we can classify each locus into 6×4 states according to the order type of the strands, and according to the nonsister pair involved in the previous crossover. Each state is written as $[h, (l_w, l_m)]$, where h is the order type of the four strands 1, 2, 3, 4, 5, or 6, and (l_w, l_m) defines the strands involved in the last chiasma before \mathcal{A} .

Consider two markers \mathcal{A} and \mathcal{B} . Suppose \mathcal{A} is in some state $s_{\mathcal{A}}$, say $[1, (0, 0)]$, and let $t^{s_{\mathcal{A}}, s_{\mathcal{B}}}(k)$ be the conditional probability that \mathcal{B} is in state $s_{\mathcal{B}}$ after k chiasmata between these two markers. For each $s_{\mathcal{B}}=[h, (l_w, l_m)]$, $t^{s_{\mathcal{A}}, s_{\mathcal{B}}}(k+1)$ is a linear function of the $t^{s_{\mathcal{A}}, s_{\mathcal{B}}}(k)$. For example, in order for \mathcal{B} to be in state $s_{\mathcal{B}}=[1, (0, 0)]$ after $k+1$ chiasmata, \mathcal{B} must be in one of the four states $[6, (0, 0)]$, $[6, (0, 1)]$, $[6, (1, 0)]$, or $[6, (1, 1)]$ after k chiasmata; these states have chance β , γ , α , and β , respectively, to change to $[1, (0, 0)]$ after the $(k+1)$ th chiasma. So $t^{s_{\mathcal{A}}, [1, (0, 0)]}(k+1)$ is equal to

$$\beta t^{s_{\mathcal{A}}, [6, (0, 0)]}(k) + \gamma t^{s_{\mathcal{A}}, [6, (0, 1)]}(k) + \alpha t^{s_{\mathcal{A}}, [6, (1, 0)]}(k) + \beta t^{s_{\mathcal{A}}, [6, (1, 1)]}(k).$$

Number the 24 possible states in the following way,

$$(1, 2, \dots, 24) = ([1, (0, 0)], [1, (0, 1)], \dots, [6, (1, 0)], [6, (1, 1)])$$

and let $\mathbf{T}(k)=(t^{i,j}(k))$, where $t^{i,j}(k), i, j = 1, \dots, 24$, is as defined above. We have $\mathbf{T}(k+1) = \mathbf{U}\mathbf{T}(k)$, where \mathbf{U} is the following 24×24 matrix.

Therefore, $\mathbf{T}(k) = \mathbf{U}^k \mathbf{T}(0) = \mathbf{U}^k$ because $\mathbf{T}(0) = I_{24 \times 24}$. Matrix $\mathbf{T}(k)$ can be divided into 36 4×4 submatrices as follows:

$$\left(\begin{array}{|c|c|c|c|c|c|} \hline \mathbf{T}^{1,1}(k) & \mathbf{T}^{1,2}(k) & \mathbf{T}^{1,3}(k) & \mathbf{T}^{1,4}(k) & \mathbf{T}^{1,5}(k) & \mathbf{T}^{1,6}(k) \\ \hline \mathbf{T}^{2,1}(k) & \mathbf{T}^{2,2}(k) & \mathbf{T}^{2,3}(k) & \mathbf{T}^{2,4}(k) & \mathbf{T}^{2,5}(k) & \mathbf{T}^{2,6}(k) \\ \hline \mathbf{T}^{3,1}(k) & \mathbf{T}^{3,2}(k) & \mathbf{T}^{3,3}(k) & \mathbf{T}^{3,4}(k) & \mathbf{T}^{3,5}(k) & \mathbf{T}^{3,6}(k) \\ \hline \mathbf{T}^{4,1}(k) & \mathbf{T}^{4,2}(k) & \mathbf{T}^{4,3}(k) & \mathbf{T}^{4,4}(k) & \mathbf{T}^{4,5}(k) & \mathbf{T}^{4,6}(k) \\ \hline \mathbf{T}^{5,1}(k) & \mathbf{T}^{5,2}(k) & \mathbf{T}^{5,3}(k) & \mathbf{T}^{5,4}(k) & \mathbf{T}^{5,5}(k) & \mathbf{T}^{5,6}(k) \\ \hline \mathbf{T}^{6,1}(k) & \mathbf{T}^{6,2}(k) & \mathbf{T}^{6,3}(k) & \mathbf{T}^{6,4}(k) & \mathbf{T}^{6,5}(k) & \mathbf{T}^{6,6}(k) \\ \hline \end{array} \right).$$

The submatrix $\mathbf{T}^{h_A, h_B}(k)$, $h_A, h_B = 1, \dots, 6$, is the transition matrix from type h_A at \mathcal{A} to type h_B at \mathcal{B} given k chiasmata between them. For example, $\mathbf{T}^{1,1}(k)$ is a 4×4 matrix with each entry being the conditional probability that, given \mathcal{A} is in state $[1, (l_w^1, l_m^1)]$, \mathcal{B} is in state $[1, (l_w^2, l_m^2)]$ after k chiasmata.

Let $\mathbf{p}_{\mathcal{A}} = (p_{\mathcal{A}}^1, \dots, p_{\mathcal{A}}^{24})'$ be the initial distribution of states at \mathcal{A} , and let $\mathbf{S} = (\mathbf{p}_{\mathcal{A}}, \dots, \mathbf{p}_{\mathcal{A}})$ and $\mathbf{P}(k) = \mathbf{S}\mathbf{T}(k) = (p^{i,j}(k))$. Then $p^{i,j}(k)$ is the joint probability that \mathcal{A} is in state i and \mathcal{B} in state j given the occurrence of k chiasmata between them. The matrix $\mathbf{P}(k)$ can be also divided into 36 4×4 submatrices, and labeled as $P^{h_{\mathcal{A}}, h_{\mathcal{B}}}(k)$. It is straightforward to obtain from $\mathbf{P}(k)$ the probability that \mathcal{A} and \mathcal{B} show parental ditype, tetratype and nonparental ditype with k chiasmata between them. For example, the chance of parental ditype with k chiasmata between the markers is the sum of all entries in the following matrices:

$$\mathbf{P}^{1,1}(k), \mathbf{P}^{2,2}(k), \mathbf{P}^{3,3}(k), \mathbf{P}^{4,4}(k), \mathbf{P}^{5,5}(k), \mathbf{P}^{6,6}(k).$$

Suppose there is chance q_k of there being k chiasmata between \mathcal{A} and \mathcal{B} and define $\mathbf{T} = \sum q_k \mathbf{T}(k)$ and $\mathbf{P} = \sum q_k \mathbf{P}(k)$. Then $t^{i,j}$ is the conditional probability that \mathcal{B} is in state j given \mathcal{A} is in state i , whereas $p^{i,j}$ is the joint probability that \mathcal{A} is in state i and \mathcal{B} in state j .

Using arguments similar to the single spore data case, we may obtain general results for multilocus ordered tetrads with $l+1$ markers, $\mathcal{A}_1, \mathcal{A}_2, \dots, \mathcal{A}_{l+1}$, ordered starting from the centromere. Let $\mathbf{k} = (k_1, k_2, \dots, k_l)$ and denote by $q_{\mathbf{k}}$ the joint probability of having k_i chiasmata between \mathcal{A}_i and \mathcal{A}_{i+1} , $i = 1, \dots, l$. Similarly, write $\mathbf{h} = (h_1, h_2, \dots, h_{l+1})$, where h_i is the order type of the i th marker. If $\mathbf{p}_{\mathcal{A}_1}^{h_1}$ be the initial distribution of the state at \mathcal{A}_1 , then we have

Theorem 3. *The multilocus probability of tetrad type \mathbf{h} is*

$$\mathbf{p}_{\mathcal{A}_1}^{h_1} \left(\sum_{\mathbf{k}} q_{\mathbf{k}} \mathbf{T}^{h_1, h_2}(k_1) \cdots \mathbf{T}^{h_l, h_{l+1}}(k_l) \right) \mathbf{1}'.$$

If there is no chiasma interference, this probability can be factored as

$$\mathbf{p}_{\mathcal{A}_1}^{h_1} (\mathbf{T}_{\mathcal{A}_1 \mathcal{A}_2}^{h_1, h_2} \cdots \mathbf{T}_{\mathcal{A}_l \mathcal{A}_{l+1}}^{h_l, h_{l+1}}) \mathbf{1}'.$$

In the above discussion it is assumed that there is at least one chiasma before \mathcal{A}_1 , so the state of \mathcal{A}_1 can be defined by the order type of the strands and the pair involved in the previous chiasma. The above results still hold if each pair is assigned the same chance of being involved in the previous chiasma when no chiasmata have occurred before \mathcal{A}_1 .

2.3. Unordered tetrad data. To analyze unordered tetrads, the most common type of tetrad data obtained from genetic experiments, we also begin with two markers \mathcal{A} and \mathcal{B} . Recall that there are three possible types of unordered tetrad data with two markers. If the unordered tetrads are thought to be generated from ordered tetrads but with the order lost, the parental ditype would result

from ordered tetrads with order types at two markers being one of $(1, 1)$, $(2, 2)$, \dots , $(6, 6)$; nonparental ditype would result from ordered tetrads with order types being one of $(1, 2)$, $(2, 1)$, \dots , $(6, 5)$; and tetratype tetrads would result from all other order pairs. Let $\mathbf{p}_{\mathcal{A}}^h$ be the initial distribution at \mathcal{A} , and write $p^p(k)$, $p^{np}(k)$, and $p^t(k)$ as the probability of parental ditype, nonparental ditype, and tetratype with k chiasmata. Then

$$p^p(k) = (\mathbf{p}_{\mathcal{A}}^1 \mathbf{T}^{1,1}(k) + \mathbf{p}_{\mathcal{A}}^2 \mathbf{T}^{2,2}(k) + \dots + \mathbf{p}_{\mathcal{A}}^6 \mathbf{T}^{6,6}(k)) \mathbf{1}',$$

$$p^{np}(k) = (\mathbf{p}_{\mathcal{A}}^1 \mathbf{T}^{1,2}(k) + \mathbf{p}_{\mathcal{A}}^2 \mathbf{T}^{2,1}(k) + \dots + \mathbf{p}_{\mathcal{A}}^6 \mathbf{T}^{6,5}(k)) \mathbf{1}',$$

$$p^t(k) = (\mathbf{p}_{\mathcal{A}}^1 (\mathbf{T}^{1,3}(k) + \dots + \mathbf{T}^{1,6}(k)) + \dots + \mathbf{p}_{\mathcal{A}}^6 (\mathbf{T}^{6,1}(k) + \dots + \mathbf{T}^{6,4}(k))) \mathbf{1}'.$$

Define $\mathbf{u} = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$, it can be shown that:

$$\begin{aligned} p^p(k) &= \mathbf{u} \mathbf{T}^{1,1}(k) \mathbf{1}', \\ p^{np}(k) &= \mathbf{u} \mathbf{T}^{1,2}(k) \mathbf{1}', \\ p^t(k) &= 4 \mathbf{u} \mathbf{T}^{1,3}(k) \mathbf{1}'. \end{aligned}$$

Suppose now that we “order” unordered tetrads by always assigning order type 1 to \mathcal{A} and order type 3 to \mathcal{B} for tetratype data. There will only be three possible ordered tetrad types, $(1, 1)$, $(1, 2)$ and $(1, 3)$ for parental ditype, non-parental ditype and tetratype. Then $p^p(k)$, $p^{np}(k)$, and $p^t(k)$ can be shown to be: $p^p = \mathbf{u} \mathbf{T}^{1,1} \mathbf{1}'$, $p^{np} = \mathbf{u} \mathbf{T}^{1,2} \mathbf{1}'$, and $p^t = 4 \mathbf{u} \mathbf{T}^{1,3} \mathbf{1}'$, where $\mathbf{T}^{h_{\mathcal{A}}, h_{\mathcal{B}}} = \sum_k q_k \mathbf{T}^{h_{\mathcal{A}}, h_{\mathcal{B}}}(k)$ is as defined in the previous subsection.

For unordered tetrads with $l+1$ markers, we may assign the order type to each marker by assigning order type 1 to \mathcal{A}_1 . If there are no tetratypes among the l intervals, the order types of other markers are uniquely determined, either 1 or 2. If \mathcal{A}_j and \mathcal{A}_{j+1} is the first pair starting from \mathcal{A}_1 having tetratype, we assign order type 3 to \mathcal{A}_{j+1} . There will be no ambiguity of assigning order types afterwards since the order type of the four strands is fixed after this step. A one to one correspondence can be established between unordered tetrad types and ordered tetrad types from this procedure. Let $\mathbf{h} = (1, h_2, \dots, h_{l+1})$ be the ordered tetrad type corresponding to the unordered tetrad type \mathbf{g} . Then the probability of type \mathbf{g} is $p^{\mathbf{g}} = p^{\mathbf{h}}$ when there is no tetratype among all l intervals, and $p^{\mathbf{g}} = 4p^{\mathbf{h}}$ otherwise.

The matrix \mathbf{U} and its powers play an important role in the above discussion, but no simple expression for \mathbf{U}^k has been obtained. In the case of unordered tetrads with two markers, explicit expressions for $p^p(k)$, $p^{np}(k)$, and $p^t(k)$ do exist.

Theorem 4. Let $a = \alpha + \gamma$, $c_1 = 3 + a$, $c_2 = -1 + a$, and $c_3 = \sqrt{-3 + 6a + a^2}$ (c_3 could be a complex number). Then

$$p^t(k) = \frac{2}{3} - \left(\frac{1}{2}\right)^k \frac{1}{12c_3} ((c_1 - c_3)(c_2 + c_3)^{k+1} - (c_1 + c_3)(c_2 - c_3)^{k+1}).$$

When $k > 0$ is even,

$$p^p(k) = \frac{1}{2}(1 - p^t(k) + (\alpha - \gamma)^{\frac{k}{2}}),$$

$$p^{np}(k) = \frac{1}{2}(1 - p^t(k) - (\alpha - \gamma)^{\frac{k}{2}}).$$

When k is odd,

$$p^p(k) = p^{np}(k) = \frac{1}{2}(1 - p^t(k)).$$

The proof is given in section 6. The number c_3 could be complex, but $p^t(k)$ is always a real number. Note that when $(\alpha, 2\beta, \gamma) = (\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$, i.e., there is no chromatid interference, $a = \frac{1}{2}$ and $p^t(k) = \frac{2}{3}(1 - (-\frac{1}{2})^k)$. This well-known result was first proved by Mather (1935). When $\alpha + \gamma = 1$, i.e., consecutive chiasmata always involve the same two strands or four strands, $p^t(k) = 1$ when k is odd and 0 when k is even. The probability that two markers are recombined on a single strand can also be derived from Theorem 4. Because one half of the strands on tetratype tetrads are recombined, and all strands on nonparental ditype tetrads are recombined, the recombination fraction r is $\frac{1}{2}p^t(k) + p^{np}(k) = \frac{1}{2}\{1 - (p^p(k) - p^{np}(k))\}$. Note that r is $\frac{1}{2}$ when k is odd, and $\frac{1}{2}(1 - (\alpha - \gamma)^{\frac{k}{2}})$ when k is even. This result was proved earlier for single spore data by another approach.

3. Data Analysis. We fit the Markov chromatid interference model to unordered tetrad *S. pombe*. The data were kindly provided by Peter Munz. In this organism, chiasma interference is thought to be absent, the chiasma process can be assumed to follow the Poisson process. The parameters α , 2β , γ , and genetic distances were estimated using maximum likelihood. The estimates of α , 2β and γ are summarized in Table 1. Four markers on chromosome I (*leu1*, *his7*, *mat*, *his5*) were used in cross XB1050. A total of 277 tetrads were genotyped. In Table 1, XB1050-1 used markers *leu1*, *his7*, and *mat*, whereas XB1050-2 used markers *his7*, *mat*, and *his5*. Seven markers on chromosome II (*mat*, *ura5*, *his3*, *tps13*, *leu3*, *ade1*, *lys4*) were used in cross XC2 which had 458 offspring. XC2-3

TABLE 1

*Estimates of α , 2β , and γ (and their standard errors) from experimental crosses using *S. pombe*. The data were provided by Peter Munz.*

Cross	α	se_α	2β	$se_{2\beta}$	γ	se_γ
XB1051-1	0.34	0.03	0.46	0.02	0.21	0.03
XB1051-2	0.14	0.03	0.77	0.02	0.09	0.02
XC2-3	0.25	0.02	0.45	0.02	0.30	0.02
XC2-5	0.26	0.02	0.48	0.02	0.26	0.02

used markers *tps13*, *mat* and *leu3*. XC2-5 used markers *leu3*, *ade1* and *lys4*. The standard errors were calculated from the numerical approximation of the Fisher information. Except for XB1051-2, estimates of α , 2β , γ are close to $\frac{1}{4}$, $\frac{1}{2}$, $\frac{1}{4}$, which correspond to no chromatid interference.

4. Incorporating chromatid and chiasma interference. The chiasma process has to be specified if we are to fit this chromatid interference model to single spore and tetrad data. Due to the difficulty of separating chromatid interference from chiasma interference and the fact that chiasma interference is observed in many organisms, a suitable model for the chiasma process is essential to the estimation of α , β , and γ . If the chiasma model is misspecified, the estimates of α , β , and γ might indicate the presence of chromatid interference even when it is, in fact, absent.

Among different chiasma process models, the chi-square model has been found to give good fit to data from a variety of organisms (Zhao, Speed and McPeek 1995). The chi-square model can be traced back to Fisher, Lyon and Owen (1947) and has generally been of interest due to its mathematical tractability. Recently the chi-square model was suggested as a plausible biological model by Foss et al. (1993). However there are now doubts concerning the appropriateness of this motivation (Foss and Stahl 1995). The model is represented in the form $Cx(Co)^m$, as follows: assume that chiasma intermediates (C events) are randomly distributed along the four-strand bundle, and every C event will either resolve in a chiasma (Cx) or not (Co). When a C resolves as a Cx , the next m C 's must resolve as Co events, and after m Co 's the next C must resolve as a Cx , i.e., the C 's resolve in a sequence $\cdots Cx(Co)^m Cx(Co)^m \cdots$. To make the process stationary, given a set of C events, the leftmost C has an equal chance to be one of $Cx(Co)^m$. We say a C event is in state k if it is the k th C after a Cx , i.e., the state “0” means that this C event is a Cx and “ k ” ($k \neq 0$) means this C event is the k th Co after a Cx . If we start counting C events from the leftmost marker, then the state of the n th C , s_n , forms a homogeneous Markov Chain.

Under the chi-square model, the state of a marker along the chromosome can be defined to be the same as the state of the last C event before the marker. If we also consider strands involved in each chiasma, a joint state of a marker \mathcal{A} can be defined as $(r, "y")$ if the last C event before \mathcal{A} is in state r and that this strand was involved in the last Cx event. Marker \mathcal{A} has joint state $(r, "n")$ if the last C event before \mathcal{A} is in state r and this strand was not involved in the last Cx event. Given \mathcal{A} being in state $s_{\mathcal{A}}$ and there being k chiasmata between \mathcal{A} and \mathcal{B} , define $p_{s_{\mathcal{A}}, s_{\mathcal{B}}}^o(k)$ as the conditional probability that \mathcal{B} is in state $s_{\mathcal{B}}$ and that the strand is involved in an odd number of chiasmata. Define $p_{s_{\mathcal{A}}, s_{\mathcal{B}}}^e(k)$ as the conditional probability that \mathcal{B} is in state $s_{\mathcal{B}}$ and that the strand is involved in an even number of chiasmata. Recursive relationships among the $p_{s_{\mathcal{A}}, s_{\mathcal{B}}}^o(k)$ and $p_{s_{\mathcal{A}}, s_{\mathcal{B}}}^e(k)$ can be easily established. For example, under the $CxCo$ model, the following relationships hold:

$$\begin{pmatrix} p_{s_{\mathcal{A}},(0,y)}^o(k+1) \\ p_{s_{\mathcal{A}},(1,y)}^o(k+1) \\ p_{s_{\mathcal{A}},(0,n)}^o(k+1) \\ p_{s_{\mathcal{A}},(1,n)}^o(k+1) \\ p_{s_{\mathcal{A}},(0,y)}^e(k+1) \\ p_{s_{\mathcal{A}},(1,y)}^e(k+1) \\ p_{s_{\mathcal{A}},(0,n)}^e(k+1) \\ p_{s_{\mathcal{A}},(1,n)}^e(k+1) \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & \eta & 0 & 1-\eta \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1-\eta & 0 & \eta & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & \eta & 0 & 1-\eta & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1-\eta & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} p_{s_{\mathcal{A}},(0,y)}^o(k) \\ p_{s_{\mathcal{A}},(1,y)}^o(k) \\ p_{s_{\mathcal{A}},(0,n)}^o(k) \\ p_{s_{\mathcal{A}},(1,n)}^o(k) \\ p_{s_{\mathcal{A}},(0,y)}^e(k) \\ p_{s_{\mathcal{A}},(1,y)}^e(k) \\ p_{s_{\mathcal{A}},(0,n)}^e(k) \\ p_{s_{\mathcal{A}},(1,n)}^e(k) \end{pmatrix}.$$

These relationships can be used to derive closed form expressions for multilocus recombination probabilities. Because the techniques are essentially the same as before, we omit the details here. Similarly, closed form multilocus tetrad probabilities can be derived under this Markov chromatid interference model and the chi-square chiasma interference model.

5. Discussion. In this paper, we studied a Markov model for chromatid interference. Although both chromatid and chiasma interference are commonly assumed to be absent in analyzing genetic data, crossover interference has been observed in almost all organisms studied, including humans. Thus a reasonable mathematical model, which can capture the main features of the genetic data, may help the understanding of the underlying biological mechanisms and the construction of genetic maps.

Genetic map functions, $r = M(d)$, are often used to relate the unobservable map distance (d) to the observable recombination fraction, r , from single spore data. Map functions are also used to infer the map distance between the centromere and the marker to the proportion of second division segregants (SDS)

at the marker using ordered tetrad data. Most of the map functions proposed in the literature were constructed to deal with crossover interference, and some do fit data well. But when there are more than three markers present in the data, multilocus recombination/tetrad probabilities are not, in general, uniquely determined by the map function. Map functions under different chromatid and chiasma interference models were compared in Zhao and Speed (1998). For single spore data, it was found that recombination fraction is no longer a monotone function of map distance when both types of interference are present. In general, there is no one to one correspondence between the map distance and the recombination fraction, unless the NCI assumption holds. The presence of chiasma interference diminishes the effect of chromatid interference. For ordered tetrad data, for different chromatid interference models, the SDS proportion never exceeds $\frac{2}{3}$ under the Poisson model. When consecutive chiasmata always involve the same pair or different pairs of strands, the SDS proportion never goes above $\frac{2}{3}$. When consecutive chiasmata always involve three strands, except for the Poisson model, the SDS proportions all rise above $\frac{2}{3}$. As with single spore data, in general, there is no one to one correspondence between the map distance and the SDS proportion, unless NCI holds. Therefore, in the presence of genetic interference, extra caution should be taken when gene centromere distance is estimated from the SDS proportion, especially when the observed SDS proportion is large.

Both the Markov chromatid interference model and the chi-square model discussed in this paper are based on discrete time Markov chains, which are equivalently definable as one-dimensional random fields. Therefore, there is no preferred directionality in theory for both models.

The chromatid interference model discussed in this paper relies on the simplified assumption that the interference parameters do not depend on the distance between two crossovers. When the distance increases, the degree of interference might decrease. Recall that for single spore data, chromatid interference is determined by $\alpha - \gamma$. Carter and Robertson (1952) and Stam (1979) proposed that $\alpha - \gamma$ is a function $g(t)$ of the distance between two consecutive chiasmata, and considered the special form $g(t) = g(0)e^{-ct}$. For tetrad data, α , β , and γ have to be specified as a function of the distance. A simple model is

$$\begin{aligned}\alpha(t) &= \alpha(0)e^{-ct} + \frac{1}{4}(1 - e^{-ct}), \\ \beta(t) &= \beta(0)e^{-ct} + \frac{1}{4}(1 - e^{-ct}), \\ \gamma(t) &= \gamma(0)e^{-ct} + \frac{1}{4}(1 - e^{-ct}).\end{aligned}$$

However, no closed form expressions for multilocus probabilities have been obtained for this model.

Like any mathematical model, the models discussed above have to be tested and validated using real data sets. A variety of goodness-of-fit tests can be performed to examine whether the model provides a reasonable fit to the data sets (Read and Cressie 1988).

Because the existence of crossover interference has been well established in many organisms, chromatid interference should be considered together with chiasma interference. Although two types of interference are not separable using single spore data (Zhao and Speed 1996), they can be distinguished from tetrad data as demonstrated in this paper. When chromatid interference is present, genetic mapping assuming the absence of chromatid interference can lead to incorrect genetic maps. In such cases, the model studied in this paper provides a useful approach to incorporating chromatid interference.

6. Proofs.

6.1. *Proof of Theorem 1.* Define $\mathbf{S}_k = \mathbf{T}_k^1 + \mathbf{T}_k^0$ and $\mathbf{D}_k = \mathbf{T}_k^1 - \mathbf{T}_k^0$. We have,

$$\mathbf{S}_{k+1} = \mathbf{S}_k \begin{pmatrix} \eta & 1-\eta \\ 1-\eta & \eta \end{pmatrix} = \mathbf{S}_k \mathbf{U},$$

and

$$\mathbf{D}_{k+1} = \mathbf{D}_k \begin{pmatrix} -\eta & 1-\eta \\ -(1-\eta) & \eta \end{pmatrix} = \mathbf{D}_k \mathbf{V}.$$

Thus,

$$\mathbf{S}_k = \mathbf{S}_0 \mathbf{U}^k \quad \text{and} \quad \mathbf{D}_k = \mathbf{D}_0 \mathbf{V}^k,$$

where

$$\mathbf{S}_0 = \mathbf{T}_0^1 + \mathbf{T}_0^0 = \begin{pmatrix} +1 & 0 \\ 0 & +1 \end{pmatrix},$$

and

$$\mathbf{D}_0 = \mathbf{T}_0^1 - \mathbf{T}_0^0 = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}.$$

It is easy to show that

$$\mathbf{U}^k = \frac{1}{2} \begin{pmatrix} 1 + (2\eta - 1)^k & 1 - (2\eta - 1)^k \\ 1 - (2\eta - 1)^k & 1 + (2\eta - 1)^k \end{pmatrix}.$$

When k is even,

$$\mathbf{V}^k = \begin{pmatrix} \delta^{\frac{k}{2}} & 0 \\ 0 & \delta^{\frac{k}{2}} \end{pmatrix}.$$

When k is odd,

$$\mathbf{V}^k = \begin{pmatrix} -\eta\delta^{\frac{k-1}{2}} & (1-\eta)\delta^{\frac{k-1}{2}} \\ -(1-\eta)\delta^{\frac{k-1}{2}} & \eta\delta^{\frac{k-1}{2}} \end{pmatrix}.$$

It is then straightforward to arrive at the expressions for \mathbf{T}_k^1 and \mathbf{T}_k^0 .

6.2 Proof of Theorem 4. Without loss of generality, assume \mathcal{A} is in state $[1, (0, 0)]$. Let $p^{s_B}(k)$ be the conditional probability that \mathcal{B} is in state s_B given k chiasmata between \mathcal{A} and \mathcal{B} . Define

$$p(k) = p^{[1,(0,0)]}(k) + p^{[1,(0,1)]}(k) + p^{[1,(1,0)]}(k) + p^{[1,(1,1)]}(k),$$

$$np(k) = p^{[2,(0,0)]}(k) + p^{[2,(0,1)]}(k) + p^{[2,(1,0)]}(k) + p^{[2,(1,1)]}(k),$$

$$t_1(k) = p^{[3,(0,0)]}(k) + p^{[4,(0,0)]}(k) + p^{[5,(0,0)]}(k) + p^{[6,(0,0)]}(k),$$

$$t_2(k) = p^{[3,(0,1)]}(k) + p^{[4,(0,1)]}(k) + p^{[5,(0,1)]}(k) + p^{[6,(0,1)]}(k),$$

$$t_3(k) = p^{[3,(1,0)]}(k) + p^{[4,(1,0)]}(k) + p^{[5,(1,0)]}(k) + p^{[6,(1,0)]}(k),$$

$$t_4(k) = p^{[3,(1,1)]}(k) + p^{[4,(1,1)]}(k) + p^{[5,(1,1)]}(k) + p^{[6,(1,1)]}(k).$$

Theorem 4 can be proved using the following recursive relationships:

$$\begin{aligned} p(k+1) &= \beta t_1(k) + \gamma t_2(k) + \alpha t_3(k) + \beta t_4(k), \\ np(k+1) &= \beta t_1(k) + \alpha t_2(k) + \gamma t_3(k) + \beta t_4(k), \\ t_1(k+1) &= \alpha t_1(k) + \beta t_2(k) + \beta t_3(k) + \gamma t_4(k), \\ t_2(k+1) &= np(k), \\ t_3(k+1) &= p(k), \\ t_4(k+1) &= \gamma t_1(k) + \beta t_2(k) + \beta t_3(k) + \alpha t_4(k). \end{aligned}$$

Acknowledgments. We thank Dr. Peter Munz for providing the S. Pombe data for our analysis and two referees for their insightful comments.

REFERENCES

- CARTER, T. C. and ROBERTSON, A. (1952). A mathematical treatment of genetical recombination using a four-strand model. *Proc. Roy. Soc., B* **139** 410–426.

- FISHER, R. A., LYON, M. F. and OWEN, A. R. G. (1947). The sex chromosome in the house mouse. *Heredity* **1** 335–365.
- FOSS, E. and STAHL, F. W. (1995). A test of a counting model for chiasma interference. *Genetics* **139** 1201–1209.
- FOSS, E., LANDE, R., STAHL, F. W. and STEINBERG, C. M. (1993). Chiasma interference as a function of genetic distance. *Genetics* **133** 681–691.
- GRIFFITHS, A. J. F., MILLER, J. H., SUZUKI, D. T., LEWONTIN, R. C. and GELBART, W. M. (1996). *An Introduction to Genetic Analysis*, 6 th edition. W. H. Freeman and Company, New York.
- MATHER, K. (1935). Reduction and equational separation of the chromosomes in bivalents and multivalents. *J. Genet.* **30** 53–78.
- MATHER, K. (1938). Crossing-over. *Biological Reviews* **13** 252–292.
- MCPEEK, M. S. and SPEED, T. P. (1995). Modeling interference in genetic recombination. *Genetics* **139** 1031–1044.
- READ, T. R. C. and CRESSIE, N. A. C. (1988) *Goodness-of-Fit Statistics for Multivariate Discrete Data*. Springer-Verlag, New York.
- SPEED, T. P. (1995). What is a genetic map function? in *Genetic Mapping and Sequencing*, edited by T. P. SPEED and M. S. WATERMAN. IMA volumes on Mathematics and its Applications. Springer-Verlag, New York.
- SPEED, T. P., MCPEEK, M. S. and EVANS, S. N. (1992). Robustness of the no-interference model for ordering genetic markers. *Proc. Natl. Acad. Sci. USA* **89** 3103–3106.
- STAM, P. (1979). Interference in genetic crossing over and chromosome mapping. *Genetics* **92** 573–594.
- STURT, E. and SMITH, C. A. B. (1976). The relationship between chromatid interference and the mapping function. *Cytogen. Cel. Genet.* **17** 212–220.
- WEINSTEIN, A. (1938). Mathematical study of multiple-strand crossing over and coincidence in the chromosomes of Drosophila. *Am. Phil. Soc. Yearbook* 1937. 227–228.
- ZHAO, H. and SPEED, T. P. (1996). On Genetic Map Functions. *Genetics* **142** 1369–1377.
- ZHAO, H. and SPEED, T. P. (1998). Stochastic modeling of the crossover process during meiosis. *Communications in Statistics, Theory and Methods*. In Press.
- ZHAO, H., MCPEEK, M. S. and SPEED, T. P. (1995). Statistical analysis of chromatid interference. *Genetics* **139** 1057–1065.
- ZHAO, H., SPEED, T. P. and MCPEEK, M. S. (1995). Statistical analysis of crossover interference using the chi-square model. *Genetics* **139** 1045–1056.

DIVISION OF BIOSTATISTICS
YALE UNIVERSITY SCHOOL OF MEDICINE
NEW HAVEN, CT 06520
HONGYU.ZHAO@YALE.EDU

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA AT BERKELEY
BERKELEY, CALIFORNIA 94720
TERRY@STAT.BERKELEY.EDU

SOME STATISTICAL ASPECTS OF CYTONUCLEAR DISEQUILIBRIA

BY SUSMITA DATTA

Georgia State University

The purpose of this paper is to review the statistical properties of cytonuclear disequilibria, which measure the association of cytoplasmic genes with nuclear genes and genotypes within a hybrid zone, under different evolutionary models. We report the exact dynamics of the expected cytonuclear genotypic disequilibria for both the homozygotes and heterozygotes in a finite population, with or without having reproductively isolated subdivisions, under random drift alone and random drift along with mutation. The dynamics for the variance is studied using Monte Carlo simulation for a subdivided population, whereas its exact formula for a single undivided population is available. The asymptotic formulas for both the expectation and variance are obtained which are compared between populations with and without reproductive barriers. Construction of a goodness of fit type statistical test using the dynamics of the cytonuclear disequilibria is discussed. An existing test in an undivided population is reviewed and a new test for a subdivided population is outlined.

1. Introduction. Scientists have noticed dramatic nonrandom associations of cytoplasmic markers with nuclear markers in a variety of hybrid populations such as mice (Ferris *et. al.*, 1983), waterfrogs (Spolsky and Uzzel, 1984), treefrogs (Lamb and Avise, 1986) etc. One needs to correctly define these associations and check whether these cytonuclear associations can be explained without invoking natural selection. These nonrandom associations of cytoplasmic genes with nuclear genes and genotypes, i.e., cytonuclear disequilibria (Lamb and Avise, 1986; Asmussen *et al.*, 1987; Arnold, 1993) can be used to infer the natural history of a particular species. For example, in hybrid zones, cytonuclear disequilibria provide important information about the directionality of the mating events between hybridizing taxa, levels of assortative mating by conspecifics and the kinds of selection on hybrids (Arnold 1993).

Rand and Harrison (1989) have suggested that some hybrid zones may be fundamentally different in character from the more clinal picture as in Mallet *et al.* (1990). For example, in the case of an extensive hybrid zone between two types of cricket along the Appalachians (*Gryllus pennsylvanicus* and *Gryllus firmus*), individuals may sort themselves according to the soil type. As a consequence, the hybrid zone becomes a mosaic of populations of one species or the

Research partially supported by NIH grant AI07442.

AMS 1991 subject classifications. Primary 92D25, 92D15; secondary 62P06, 92D10.

Key words and phrases. Genetic drift, neutrality hypothesis, goodness of fit test.

other reflecting the patchwork character of soil types. This kind of subdivided population structure is exactly what Wright (1968) hypothesized as necessary for the shifting balance theory. Under this kind of structure a population has the opportunity to explore repeatedly new gene combinations created by hybridization, and natural selection can retain those complexes that are adaptive.

Until recently, only single locus models of cytoplasmic diversity have been considered by Birky *et al.* (1989). Fu and Arnold (1991) have discussed the behavior of the allelic disequilibria in a subdivided population or a mosaic hybrid zone under a cytonuclear system. Earlier, Ohta (1982) considered this problem in a nuclear system.

In a recent article, Datta and Arnold (1998) studied the dynamics of two other genotypic disequilibria measures (to be defined in the next section) in a subdivided cytonuclear system. These results are contrasted with those in the case of a single (or undivided) cytonuclear population (Fu and Arnold, 1991; Datta *et al.*, 1996a).

In the next section, we define the various cytonuclear disequilibria in a subdivided population. In Section 3, we include the dynamics (in terms of number of generations) of the overall cytonuclear disequilibria under random drift alone and random drift in combination with mutation, respectively, for a subdivided population and compare the results with those in an undivided population. In Section 4, we present the asymptotic results for the same disequilibria under the random drift model. Construction of statistical tests using the expected trajectory of the cytonuclear disequilibria is discussed in Section 5. An overall discussion of the results of Sections 3 and 4 including their biological significance is given in Section 6.

2. Cytonuclear disequilibria. We consider a population consisting of n isolated subpopulations each having the same discrete, nonoverlapping generations. These subpopulations are isolated reproductively either by intrinsic or extrinsic barriers to gene exchange. Hence they evolve independently. We observe the whole population at two restriction sites, one at a nuclear site with allelic types A and a and the other at a cytoplasmic site with alleles M and m . As a result, there are six possible cytonuclear genotypes with (relative) frequencies denoted by p_k , $k = 1, \dots, 6$ (Table 1). The corresponding frequencies within the subpopulations will be denoted by double suffixes. For example, p_{1i} is the frequency of the AA/M cytonuclear genotype in the i th subpopulation. The measures of the cytonuclear disequilibria for the homozygous case AA/M and for the heterozygous case Aa/M , within subpopulation i are defined by

$$D_{1i} = p_{1i} - u_i q_i, i = 1, \dots, n, \quad D_{2i} = p_{2i} - v_i q_i, i = 1, \dots, n,$$

TABLE 1
Frequencies of cytonuclear genotypes

Cytoplasm	Nuclear genotype			Total
	AA	Aa	aa	
M	p_1	p_2	p_3	q
m	p_4	p_5	p_6	$1 - q$
Total	u	v	w	1

respectively. Here n is the total number of subpopulations, and for each i ,

$$u_i = p_{1i} + p_{4i},$$

$$q_i = p_{1i} + p_{2i} + p_{3i}, \quad v_i = p_{2i} + p_{5i}.$$

The overall frequencies of these genotypes in the entire population are denoted by P_1, \dots, P_6 , where each P_k is defined as

$$P_k = (\sum_{i=1}^n n_i p_{ki}) / (\sum_{i=1}^n n_i); \quad k = 1, \dots, 6,$$

with n_i being the size of the i th subpopulation, $1 \leq i \leq n$. Thus, the **overall cytonuclear disequilibria** for the entire subdivided population corresponding to the homozygous case AA/M and the heterozygous case Aa/M , are given by

$$D_{1,ST} = P_1 - UQ \quad \text{and} \quad D_{2,ST} = P_1 - VQ,$$

respectively, where

$$U = P_1 + P_4, \quad V = P_2 + P_5, \quad Q = P_1 + P_2 + P_3.$$

An overall gametic disequilibrium can be defined along the same line as

$$D_{ST} = e_{1,ST} - PQ,$$

where P is given by

$$P = e_{1,ST} + e_{2,ST}.$$

$e_{1,ST}$ is the frequency of the gametic type A/M and $e_{2,ST}$ is the frequency of the gametic type A/m in the entire population. Results concerning the behavior of $E(D_{ST})$ and $Var(D_{ST})$ over time can be found in Fu and Arnold (1992) for the case $n = 1$, and Fu and Arnold (1991) for $n \geq 1$, respectively. In the next

section, we will discuss the dynamics of the expected disequilibria $E(D_{1,ST})$ and $E(D_{2,ST})$ over time (generation). Datta and Arnold (1998) show that for the special case when all the subpopulations have the same size, $E(D_{1,ST})$ can be written as

$$E(D_{1,ST}) = \bar{D}_1 + \frac{n-1}{n} \overline{\text{cov}}(u, q).$$

Here the quantity $\bar{D}_1 = \sum_{i=1}^n E(p_{1i} - u_i q_i)/n$ measures the average homozygous cytonuclear disequilibria within subpopulations and $\overline{\text{cov}}(u, q) = \frac{1}{n} \sum_{i=1}^n [E(u_i q_i) - E(u_i)E(q_i)]$, is the covariance (cov) in site frequencies u and q averaged across subpopulations. Similarly, $E(D_{2,ST})$ can be represented as

$$E(D_{2,ST}) = \bar{D}_2 + \frac{n-1}{n} \overline{\text{cov}}(v, q).$$

Since under the scenario of a random drift model, all the individual subpopulations will reach equilibrium eventually (Datta *et al.*, 1996a), the average cytonuclear disequilibria converges to zero with time. Note that this is in contrast with the eventual behavior for an undivided population, i.e., $n = 1$, where there is no covariance term. Hence any nonzero value of the overall cytonuclear disequilibria after a long period of time will have to come from the between population covariance term. However the behavior of the cytonuclear disequilibria for a relatively small generation number is much more complicated as the results in the next section show.

3. Dynamics of the cytonuclear disequilibria. The behavior of the expectation and variance curves over time (generation number) for the cytonuclear disequilibria $D_{1,ST}$ and $D_{2,ST}$ for $n = 1$ were studied by Datta *et al.* (1996a) and were later generalized to the case $n \geq 1$ in Datta and Arnold (1998). Here, we present the more general results for an arbitrary n first and then note its various ramifications for an undivided population by specializing to the case $n = 1$. Unlike Fu and Arnold (1991), Datta and Arnold (1998) considered the most general case, where the subpopulation sizes are allowed to be different, and they are subject to change over generations. We denote the size of the i th subpopulation at time t by $n_i(t)$, ($i = 1, \dots, n$, $t \geq 1$) where there are n subpopulations in the entire population.

The formulas reported in this paper are somewhat more complicated than the ones in Fu and Arnold (1991). This is mostly due to the fact that we are dealing with genotypic disequilibria rather than allelic disequilibria and also partly due to the fact that we made no special assumption on the subpopulation sizes. However, they are not too difficult to calculate in a computer program.

Sometimes to save space and to keep the notation simple we may not show this dependence on t , but it is to be understood.

3.1. *Random drift alone.* Recall that the overall cytonuclear disequilibria for the entire population due to the homozygote AA at the nuclear locus and M at the mtDNA locus is

$$D_{1,ST} = P_1 - UQ = P_1 - (P_1^2 + P_1P_2 + P_1P_3 + P_1P_4 + P_2P_4 + P_3P_4),$$

where $P_k(t) = \sum_{i=1}^n n_i(t)p_{ki}(t)/N(t)$; $k = 1, \dots, 6$, and $N(t) = \sum_{i=1}^n n_i(t)$ is the size of the entire population at time t . The six cytonuclear genotypic frequencies in the i th subpopulations are random variables with joint probability mass function (p.m.f) g_i . These random variables have the Markov property, i.e., this joint density of the current generation can be calculated by conditioning on the previous generation. See, e.g. Datta *et al.* (1996a) for the exact description of this conditional distribution. We assumed that the subpopulations are independent and hence the cytonuclear genotypic frequencies of different subpopulations are independent. Therefore all the expectations are calculated with respect to the product p.m.f g_1, \dots, g_n . Thus the expected value of $D_{1,ST}$ is given by

$$\begin{aligned} E(D_{1,ST}) &= \frac{1}{N} E\left(\sum_{i=1}^n n_i p_{1i}\right) - \frac{1}{N^2} E\left\{\sum_{i=1}^n \sum_{j=1}^n n_i n_j (p_{1i}p_{1j} + p_{1i}p_{2j} \right. \\ &\quad \left. + p_{1i}p_{3j} + p_{1i}p_{4j} + p_{2i}p_{4j} + p_{3i}p_{4j})\right\}, \\ &= \frac{1}{N} E\left(\sum_{i=1}^n n_i p_{1i}\right) - \frac{1}{N^2} \left[\left(\sum_{i=1}^n n_i E p_{1i}\right)^2 + \left(\sum_{i=1}^n n_i E p_{1i}\right) \left(\sum_{j=1}^n n_j E p_{2j}\right) \right. \\ &\quad \left. + \left(\sum_{i=1}^n n_i E p_{1i}\right) \left(\sum_{j=1}^n n_j E p_{3j}\right) + \left(\sum_{i=1}^n n_i E p_{1i}\right) \left(\sum_{j=1}^n n_j E p_{4j}\right) \right. \\ &\quad \left. + \left(\sum_{i=1}^n n_i E p_{2i}\right) \left(\sum_{j=1}^n n_j E p_{4j}\right) + \left(\sum_{i=1}^n n_i E p_{3i}\right) \left(\sum_{j=1}^n n_j E p_{4j}\right) \right] \\ &\quad + \frac{1}{N^2} \sum_{i=1}^n n_i^2 E D_{1i} - \frac{1}{N^2} \sum_{i=1}^n n_i^2 E p_{1i} \\ &\quad + \frac{1}{N^2} \left[\sum_{i=1}^n n_i^2 \{E p_{1i}^2 + (E p_{1i})(E p_{2i}) + (E p_{1i})(E p_{3i}) \right. \\ &\quad \left. + (E p_{1i})(E p_{4i}) + (E p_{2i})(E p_{4i}) + (E p_{3i})(E p_{4i})\} \right]. \end{aligned} \tag{3.1}$$

Note that all the quantities in the above expression are evaluated at a given time t . The expectation ED_{1i} is the value of the cytonuclear disequilibrium within the i th subpopulation. To derive the above term one uses the fact that the subpopulations are independent of each other and hence $E(p_{ki}p_{kj}) = E(p_{ki})E(p_{kj})$ for $i \neq j$. To find the expectation of p_{ki} 's one assumes the RUZ (Random Union of Zygotes) model for the mating within each cytonuclear subpopulation (Fu and Arnold, 1992). For each subpopulation under the random drift model the conditional moment generating function can be written as

$$\begin{aligned}
 M(\theta_1, \dots, \theta_6) &= E[\exp\{\sum_k \theta_k p_{ki}(t)\}|p_{ki}(t-1)] \\
 &= \sum_{f,m} Pr(Y_{fmi}(t)) \exp\left(\sum_k \theta_k p_{ki}(t)\right), \\
 (3.2) \quad &= \left(\sum_{f,m} e_{fi}(t-1) e_{mi}(t-1) \exp(\sum_k \theta_k \alpha_{fmk}/n_i(t))\right)^{n_i(t)}
 \end{aligned}$$

Here, i stands for the i th subpopulation. f, m stand for the father and the mother, respectively, and α_{fmk} are known constants (Datta *et al.*, 1996a). The count Y_{fmi} is the number of individuals in the i th subpopulation receiving gametes of type f from the father and type m from the mother. Each f and m can be one of four gametes A/M , A/m , a/M or a/m , and e_{fi} , e_{mi} are the gametic frequencies in fathers and mothers respectively, for the i th subpopulation. For details, we refer to Datta *et al.* (1996a).

From the above moment generating function one can find the expected values of the genotypic frequencies p_{ki} and also D_{1i} within a subpopulation. To that end, Datta and Arnold (1998) defined the following variables:

$$\begin{aligned}
 x_1 &= D, \quad x_2 = Dp, \quad x_3 = pq, \quad x_4 = p^2q, \\
 x_5 &= p^2, \quad x_6 = q, \quad x_7 = p.
 \end{aligned}$$

One can then verify from the moment generating function (3.2) that

$$(3.3) \quad \underline{X}(t) = \mathbf{A}(t)\underline{X}(t-1),$$

where the $\underline{X} = X_1, \dots, X_7$ is a vector containing the expectations of the x 's mentioned above and $\mathbf{A}(t)$ is a 7×7 matrix whose nonzero elements are polynomials of the subpopulation sizes at time t , $n_i(t)$. It is easy to solve the linear recursion (3.3) and obtain the X 's at a given time t knowing their initial values (at $t = 0$). Moreover, Datta and Arnold (1998) noted that the expectations of all the p_{ki} and D_{1i} can be written in terms of X 's. Consequently, from (3.1),

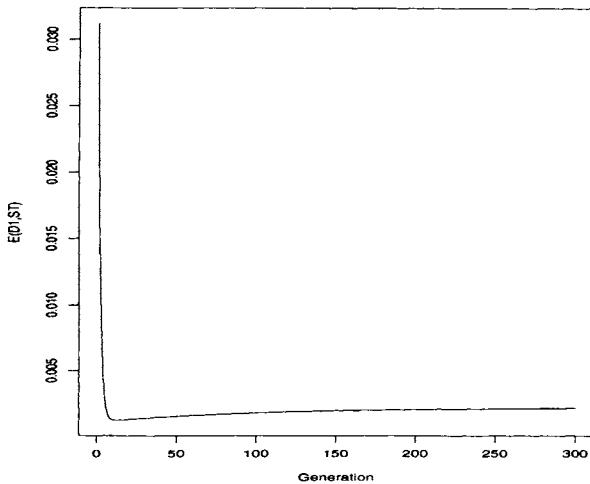


FIG. 1. *Trajectory of the expectation of overall cytonuclear disequilibria, $E(D_{1,ST})$, over time for a population consisting of 10 subpopulations. The initial values are $p(0) = q(0) = 0.25$, $D(0) = 0.125$.*

the expectations of the overall cytonuclear disequilibria $D_{1,ST}$ and $D_{2,ST}$ can be expressed in terms of the X 's.

If we assume that the sizes of the subpopulations remain constant over all generations and they are the same for all the subpopulations, then the above expressions become somewhat simpler. Graphs of $E(D_{1,ST})$ and $E(D_{2,ST})$ for this special case are given in Figures 1 and 2, respectively. We choose the initial values $p(0) = 0.25$, $q(0) = 0.25$, $D(0) = 0.125$ which are taken to be the same for all the subpopulations. In both the figures the number of subpopulations $n = 10$ and we consider that all the subpopulations are of the same size $s = 50$, which remains constant over the generations. In Figure 1, we see that the expectation of $D_{1,ST}$ drops rapidly in the first few generations and afterwards it steadily approaches its asymptotic limit given by (4.2) in Section 4. The nonzero asymptotic value is 0.0022058, for $n = 10$. In Figure 2, we see that $E(D_{2,ST})$ approaches zero (its asymptotic value, (see (4.3))) quite fast even in a subdivided population.

If the population is not subdivided i.e., $n = 1$ then the recursions for the expectations of the cytonuclear disequilibria $E(D_1)$ and $E(D_2)$ over the generations are given in Datta *et al.* (1996a). It is shown that, unlike in the presence of the subdivision in the population, the undivided population reaches equilibrium after only a few generations and all the expected disequilibria measures approach zero after just a few generations in the presence of just random drift.

For the subdivided population, the exact formulas for the variance function

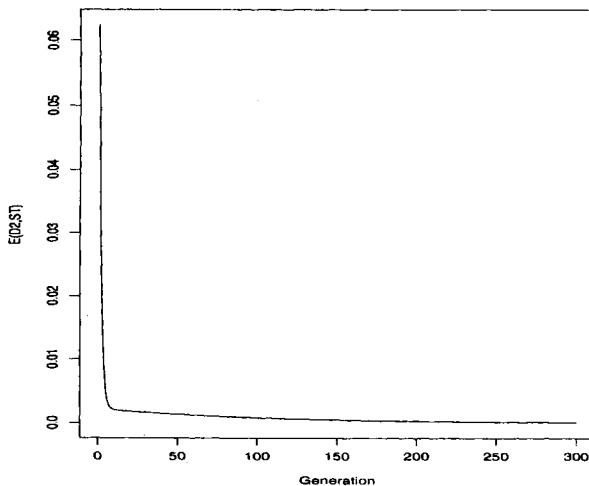


FIG. 2. Trajectory of the expectation of overall cytonuclear disequilibria, $E(D_{2,ST})$, over time for a population consisting of 10 subpopulations. The initial values are $p(0) = q(0) = 0.25$, $D(0) = 0.125$.

over time (generation) of the overall disequilibria are likely to be extremely complicated even for the case when all the subpopulations are of the same size and have the same initial condition. Therefore, we only report the results of a Monte Carlo simulation to describe the approximate trajectory of the overall variances of the disequilibria $D_{1,ST}$ and $D_{2,ST}$. We have used a smoothed version of the simulated variance curve to partially remove the simulation error. The graphs of the trajectories are given in Figures 3 and 4, respectively. For both the trajectories the number of subpopulations $n = 10$ and we assume that all the subpopulations are of the same size $s = 50$, which remains constant over generations. From Figure 5, we can see that $Var(D_{1,ST})$ decreases for just a few initial generations and after that it increases and eventually it converges to its predicted asymptotic value given in equation (4.2) in the next section. The pattern remains the same for different number of subpopulations (result not shown). It can also be shown that the magnitude of $Var(D_{1,ST})$ remains higher for smaller number of subdivisions consistently for all the generations (Datta and Arnold, 1998). Consequently, the asymptotic values are also higher for smaller number of subdivisions (Datta and Arnold, 1998). In Figure 6, we observe that unlike $Var(D_{1,ST})$, $Var(D_{2,ST})$ increases for a few initial generations and after that decreases and converges to zero. The pattern remains the same for different numbers of subdivisions (Datta and Arnold, 1998).

When the population is undivided, i.e., $n = 1$, the exact calculation of the variance under the random drift model is shown in Datta *et al.* (1996a). It is

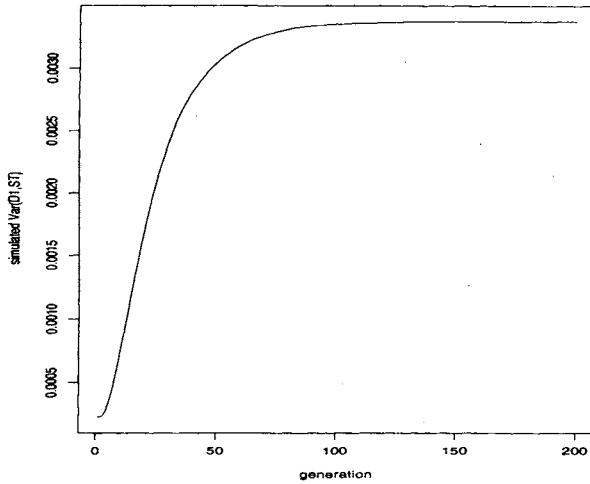


FIG. 3. Trajectories of the simulated variance of overall cytonuclear disequilibria $\text{Var}(D_{1,ST})$ over 200 generations for a population consisting of 10 subpopulations. The initial values are $p(0) = q(0) = 0.25$, $D(0) = 0.125$.

clear from the graphs of the variances that both the variances go to zero after a few generations (Datta *et al.*, 1996a).

3.2. Random drift with mutation. In this section we will consider methods to calculate the expected value of the overall cytonuclear disequilibrium $E(D_{1,ST})$ under the random drift model in the presence of mutation.

Suppose the mutation rate at the nuclear locus from A to a is μ_1 , a to A is ν_1 , and at the mtDNA locus is μ_2 for M to m and ν_2 for m to M , respectively. Following the same argument as in Ohta and Kimura (1969), we have

$$p_m(t) = (1 - \mu_1 - \nu_1)p(t) + \nu_1, \quad q_m(t) = (1 - \mu_2 - \nu_2)q(t) + \nu_2,$$

$$D_m(t) = (1 - \mu)D(t),$$

where

$$\mu = \mu_1 + \nu_1 + \mu_2 + \nu_2,$$

and the higher order terms are ignored. The recursions of the expectations $E(D_{1,ST})$ and $E(D_{2,ST})$ can be found by solving the recursive relationship given below.

$$X_m^*(t) = \left(\prod_{j=1}^t H^* A^*(t-j+1) \right) X_m^*(0)$$

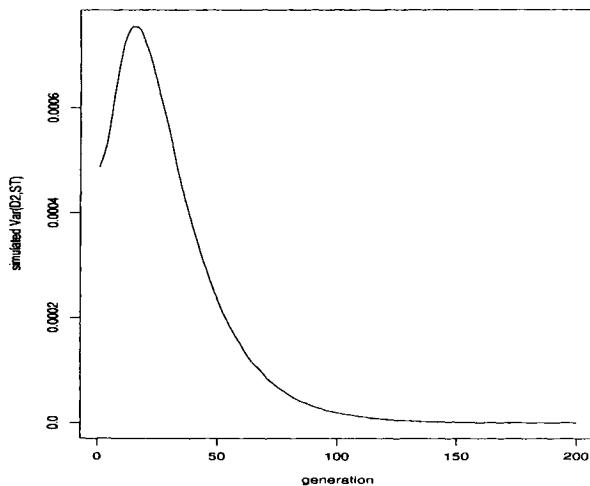


FIG. 4. Trajectories of the simulated variance of overall cytonuclear disequilibria $\text{Var}(D_{2,ST})$ over 200 generations for a population consisting of 10 subpopulations. The initial values are $p(0) = q(0) = 0.25$, $D(0) = 0.125$.

where X_m^* are the values of the X vector discussed in Section 2 in the presence of mutation. H^* and A^* are 8×8 matrices. For the details, see Datta and Arnold (1998).

We draw the trajectories of the expected values of expected values of $D_{1,ST}$ and $D_{2,ST}$ in Figures 5 and 6 respectively. From Figure 5, we notice that in the presence of mutation, the expectation of the overall D_1 decays down to zero eventually, although it takes a long time for the mutation to remove the non zero asymptotic value of the expectation under random drift alone. Hence the rate of decay under the mutation could be extremely slow. In Figure 5, we find that when the mutation rate is larger, the rate of decay is faster, as to be expected. In the case of $E(D_{2,ST})$ however, even in the presence of mutation the value converges to zero much faster than $E(D_{1,ST})$ irrespective of the magnitude of the mutation rates (Figure 6).

Datta *et al.* (1996a) obtained the asymptotic result for the expected cytonuclear disequilibria in the presence of mutation if the population is not subdivided. It can be shown that all the steady state expectations are zero in this case.

4. Asymptotic results. In this section, we discuss the asymptotic results for the expected values and the variances of the cytonuclear disequilibria under the random drift model in the case of a subdivided population. Assume that all

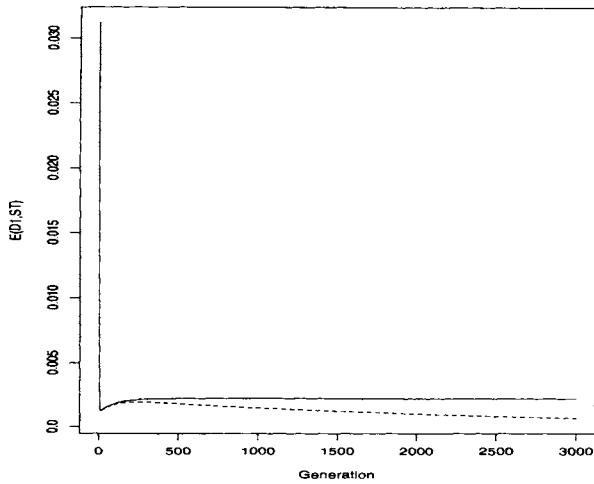


FIG. 5. Trajectories of the expectation of overall cytonuclear disequilibria $E(D_{1,ST})$ over 3000 generations for 10 subpopulations with two different mutation rates. The solid line represents mutation rates of $\mu_1 = \nu_1 = 10^{-6}$, $\mu_2 = \nu_2 = 10^{-6}$. The dashed line represents mutation rates of $\mu_1 = \nu_1 = 10^{-4}$, $\mu_2 = \nu_2 = 2 \times 10^{-4}$. The initial values in all cases are $p(0) = q(0) = 0.25$, $D(0) = 0.125$.

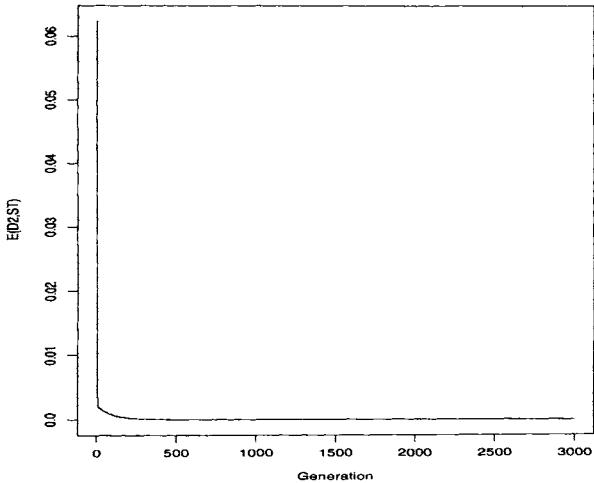


FIG. 6. Trajectories of the expectation of overall cytonuclear disequilibria $E(D_{2,ST})$ over 3000 generations for 10 subpopulations for two different mutation rates. Sets of mutation rates are $\mu_1 = \nu_1 = 10^{-4}$, $\mu_2 = \nu_2 = 2 \times 10^{-4}$, and $\mu_1 = \nu_1 = 10^{-6}$, $\mu_2 = \nu_2 = 10^{-6}$. Trajectories are almost the same for both sets of rates. The initial values in all cases are $p(0) = q(0) = 0.25$, $D(0) = 0.125$.

the subpopulations are of the same constant (in time) size s and they all have the same initial conditions. Note that, eventually, each of the subpopulations will be fixed for one of the four possible genotypes p_1, p_4, p_3 and p_6 . Let the corresponding probabilities be g_1, g_4, g_3, g_6 , respectively. The values of g 's can be determined from the initial values $D(0)$, $p(0)$ and $q(0)$ which are assumed to be the same for all the subpopulations. Thus, when all the subpopulations have reached fixation, sampling from n subpopulations is equivalent to taking a sample of size n from a multinomial distribution with parameters g_1, g_4, g_3, g_6 . From the moment generating function of the multinomial distribution we can find

$$(4.1) \quad E(D_{1,ST}(\infty)) = \left(\frac{n-1}{n}\right)D^*(0),$$

$$\begin{aligned} Var(D_{1,ST}(\infty)) &= \left(\frac{n-1}{n^3}\right)\{(n-1)D^*(0) - 2D^*(0)^2(n-1) \\ &\quad - 2D^*(0)p(0)(n-1) - 2D^*(0)q(0)(n-1) \\ &\quad - 4D^*(0)p(0)q(0)(n-1) + np(0) - np^2(0)q(0) \\ (4.2) \quad &\quad - np(0)q^2(0) + np^2(0)q^2(0)\}, \end{aligned}$$

where

$$D^*(0) = \left(\frac{1}{s+1}\right)D(0).$$

Note that when $n = 1$, i.e, when the population is not subdivided then both the above quantities are zero.

Since at time $t = \infty$, each of the sub-populations will be fixed at one of the four genotypes mentioned above, $P_2 = P_5 = 0$. Therefore, asymptotically $D_{2,ST} = 0$ and $D_{1,ST} = D_{ST}$, with probability one. Indeed it can be checked that the above asymptotic expressions for $D_{1,ST}$ agrees with those for D_{ST} as given in Fu and Arnold (1991). Furthermore,

$$(4.3) \quad E(D_{2,ST}(\infty)) = 0; \quad Var(D_{2,ST}(\infty)) = 0.$$

5. Statistical tests based on the dynamics of cytonuclear disequilibria. One can use the formulas for the expectation and the variance for D_1 and D_2 over time to construct a goodness of fit type statistical test which assesses departure from a given model. In particular, using the moment formulas under random drift, this approach yields a test of the neutrality hypothesis using the dynamics of cytonuclear disequilibria. A number of tests following this idea can be constructed depending on the sampling scheme used to obtain the necessary data.

5.1. *Test for random drift in undivided populations.* Consider a single undivided population. In this case, the formulas for the expectations and variances under the random drift model can be found by specializing to $n = 1$ in the general formulas for a subdivided population or directly from the results in Datta *et al.* (1996a). The above paper also shows how to calculate the covariance between D_1 and D_2 at a given time following the approach described in Section 3.1.

Suppose, data are available on a number of populations of the same species. We assume that the initial conditions are known for each and that the i th population is completely sacrificed at time i so that one can calculate $D_1(t)$ and $D_2(t)$ only for $t = i$ for this population. For example, this sampling scheme was used by Scribner and Avise (1994). In this scheme, the statistics for different generations are independent of each other. Note that under this sampling scheme the counts are based on a complete census of the population and there is no sampling variability (only the genetic variability). Letting $\underline{D}(t) = (D_1(t), D_2(t))$, $\underline{\mu}(t) = (ED_1(t), ED_2(t))$, and

$$\Sigma(t) = \begin{pmatrix} Var(D_1(t)), & Cov(D_1(t), D_2(t)) \\ Cov(D_1(t), D_2(t)), & Var(D_2(t)) \end{pmatrix}$$

where all the moments are calculated under the random drift model, one can construct a test statistic

$$T = \sum_{t=1}^k (\underline{D}(t) - \underline{\mu}(t))^T \Sigma^{-1}(t) (\underline{D}(t) - \underline{\mu}(t))$$

which measures the total distance across time between the observed and the expected disequilibria under random drift. The asymptotic null distribution of the above statistic was shown to be chi-square with $2k$ degrees of freedom by Datta and Arnold (1996). Therefore one would reject the random drift model if $T > \chi_\alpha^2(2k)$. Datta *et al.* (1996b) proposed a test along the same line under a different sampling scheme which results in both sampling as well as genetic variation.

5.2. *Tests of random drift in subdivided populations.* Consider a subdivided population with n components which had identical initial conditions. Then in the notation of Section 2, $ED_{1i}(t)$ and $ED_{2i}(t)$ are constant in i , up to terms that are $O(n_i^{-1})$. Denote the common value by $\mu_1(t)$ and $\mu_2(t)$ respectively. Suppose, the entire population is sacrificed at time t so that one can calculate $D_{1i}(t)$ and $D_{2i}(t)$ for $i = 1, \dots, n$. Combining these disequilibria measures from

each subpopulation one can construct an overall estimate $\hat{\mu}_j(t), j = 1, 2$ by

$$\hat{\mu}_j(t) = g_j \left(\sum_{i=1}^n \hat{v}_{ij} \lambda_i g_j^{-1}(D_{ij}(t)) / \sum_{i=1}^n \hat{v}_{ij} \lambda_i \right),$$

where g_j is a smooth function, $\lambda_i = n_i/N$, v_{ij} is the variance of D_{ij} . For example, g_j could be such that $g_j(u) = \sqrt{u}$ for $j = 1, 2$. Note that unlike $D_{j,ST}(t)$, $\hat{\mu}_j(t)$ is both asymptotically (as n_i grows) unbiased as well as efficient. A test of random drift based on $\underline{\hat{\mu}}(t)$ can be constructed using the statistic

$$T = (\underline{\hat{\mu}}(t) - \underline{\mu}(t))^T \Sigma^{-1}(t) (\underline{\hat{\mu}}(t) - \underline{\mu}(t))$$

where

$$\Sigma(t) = \begin{pmatrix} Var(\hat{\mu}_1(t)), & Cov(\hat{\mu}_1(t), \hat{\mu}_2(t)) \\ Cov(\hat{\mu}_1(t), \hat{\mu}_2(t)), & Var(\hat{\mu}_2(t)) \end{pmatrix}.$$

Note that Σ can be calculated from the variance covariance formulas for $D_{1i}(t)$, $D_{2i}(t)$ via the delta method and the independence among subpopulations. Moreover it is possible to show that T has an approximate chi-square distribution with two degrees of freedom.

If one has data from a number of independent subdivided populations as in the previous subsection then one can obtain an overall test statistic by adding the T 's obtained from each subdivided population.

6. Discussion. In this paper, we have reviewed some recent results on the exact dynamics of the expectation of two measures of overall cytonuclear disequilibrium, $D_{1,ST}$ and $D_{2,ST}$, in a subdivided population with n demes under genetic drift and genetic drift with mutation. The variance curves of these disequilibria measures are also studied and construction of statistical tests using the above results are discussed.

It is an established fact that genetic drift generates variation in linkage disequilibrium between two or more genetic loci in a subdivided population (Ohta, 1982a; Ohta, 1982b) and that mutation eventually eliminates that permanent disequilibria caused by genetic drift. In this paper, in the absence of mutation we see that the expectation of cytonuclear disequilibrium $E(D_{1,ST})$ eventually goes to a nonzero asymptotic value $[(n-1)/n]D^*(0)$, ($D^*(0)$ is defined in Section 4) which is the same as the asymptotic value of the allelic disequilibrium given in Fu and Arnold (1991). However, the exact dynamics of $E(D_{1,ST})$ are different from that of the expected allelic disequilibrium. Initially, both the within-subpopulation and between-subpopulation component of the disequilibria decrease very rapidly. Then the between-subpopulation component starts

increasing slowly and it finally settles down for the between-population component $[(n - 1)/n]D(0)/(s + 1)$. An increased subdivision (larger n , the number of subpopulations) increases the value of $E(D_{1,ST})$ (Datta and Arnold, 1998) and consequently yields a higher steady-state value. The between-population component of $E(D_{1,ST})$ is always a little larger for larger n . It is easy to see that increasing $D(0)$ and/or decreasing s will result in larger steady-state values. In the presence of mutation the permanent association is eventually removed and $E(D_{1,ST})$ decreases to zero, but it may take a very long time (Figure 2). Note that higher mutation rates imply a shorter time till the expectation of $D_{1,ST}$ goes to zero (Figure 2).

Unlike the expectation of $D_{1,ST}$, expectation of $D_{2,ST}$ goes down to zero (its asymptotic value) quite fast under the random drift model. It is interesting to see that the between-population component of $D_{2,ST}$ also goes down to zero for the heterozygote. Even with the presence of mutation it goes down to zero quite fast irrespective of the different mutation rates. Genetic drift weeds out the nuclear heterozygotes Aa/M and Aa/m , and no ' F_1 hybrids' remain to generate a nonzero D_2 .

For an undivided single population both the disequilibria measures eventually go to zero under the random drift model and in the presence of mutation along with random drift. The variance of cytonuclear disequilibria decays asymptotically under the random drift model if there is no other source of variability like mutation or migration. In the presence of mutation it has non-zero asymptotic value. For the detailed discussion on this matter we refer to Datta *et al.* (1996a).

For a subdivided population, trajectory of the simulated variance of the disequilibrium for homozygotes $Var(D_{1,ST})$ under the random drift model is shown in Figure 5. In the initial generations, value of the variance is low but it increases quite rapidly under the random drift to reach the predicted non-zero steady-state value as expected. We have discussed in Section 4 that asymptotically $D_{1,ST}$ and D_{ST} are the same. Hence the steady-state variance is also the same. The asymptotic value of the variance $Var(D_{1,ST})$ increases as number of subpopulations decrease (Datta and Arnold, 1998). Variance of $D_{2,ST}$ (Figure 6) on the other hand goes down to zero within the first 150 generations approximately for all the different number of subpopulations and that is what one would expect, because under the random drift overall disequilibria due to the heterozygotes goes to zero. The nonzero values of the variance $Var(D_{2,ST})$ are higher for smaller number of subdivisions (Datta and Arnold, 1998).

Clearly the results in this paper offer more insight into the behavior of a subdivided system under a neutral model than those using just the allelic disequilibrium. This will be reflected in the additional statistical power if one

constructs a test comparing the observed dynamics of the pair $(D_{1,ST}, D_{2,ST})$ with the expected dynamics under a neutral model. Such neutrality tests in the case of an undivided population have been proposed by Datta and Arnold (1996) and Datta *et al.* (1996b). An extension of the Datta and Arnold (1996) test to a subdivided population has been briefly described in Section 5.2 of this paper. This test is applicable if the data are collected in an ideal experimental setting as described in Section 5.2. In the case of an undivided population, data using such a sampling scheme have been collected by Scribner and Avise (1994). An extension of the test to handle a more practical setup incorporating genetic as well as statistical sampling can also be done.

Acknowledgments. I would like to thank Professor P. K. Sen for some useful discussions. Thanks are also due to two anonymous referees for their critical reviews which have improved the manuscript to a great extent.

REFERENCES

- ARNOLD, J. (1993). Cytonuclear Disequilibria in hybrid zones. *Annu. Rev. Ecol. Syst.* **24** 521–554.
- ASMUSSEN, M. A., ARNOLD, J. and ARNOLD, J. (1987). Definition and properties of disequilibrium statistics for associations between nuclear and cytoplasmic genotypes. *Genetics* **115** 1351–1363.
- BIRKEY, C. W., FUERST, P. and MARUYAMA, T. (1989). Organelle gene diversity under migration, mutation, and drift: equilibrium expectations, approach to equilibrium, effects of heteroplasmic cells, and comparison to nuclear genes. *Genetics* **121** 613–627.
- DATTA, S. and ARNOLD, J. (1996). Diagnostics and a Statistical test of neutrality hypotheses using the dynamics of cytonuclear disequilibria. *Biometrics* **52** 1042–1054.
- DATTA, S. and ARNOLD, J. (1998). Dynamics of cytonuclear disequilibria in subdivided populations. *Journal of Theoretical Biology*, **192** 99–111.
- DATTA, S., FU, Y. X. and ARNOLD, J. (1996a). Dynamics and equilibrium behavior of cytonuclear disequilibria under genetic drift. *Theor. Pop. Biol.* **50** 298–324.
- DATTA, S., KIPARSKY, M., RAND, D. M. and ARNOLD, J. (1996b). A statistical test of a neutral model using the dynamics of cytonuclear disequilibria. *Genetics* **144** 1985–1992.
- FERRIS, S. D., SAGE, R. D., HUANG, C. M., NIELSEN, J. T., RITTE, U. and WILSON, A. C. (1983). Flow of mitochondrial DNA across a species boundary. *Proc. Natl. Acad. Sci. USA* **80** 2290–2294.
- FU, Y. X. and ARNOLD, J. (1991). On the association of fragment length polymorphisms across species boundaries. *Proc. Natl. Acad. Sci. USA* **88** 3967–3971.
- FU, Y. X. and ARNOLD, J. (1992). Dynamics of cytonuclear disequilibria in finite populations and a comparison with a two-locus nuclear system. *Theor. Popul. Biol.* **41**, 1–25.
- LAMB, T. and AVISE, J. C. (1986). Directional introgression of mitochondrial DNA in a hybrid population of tree frogs: the influence of mating behavior. *Proc. Natl. Acad. Sci. USA* **83** 2526–2530.
- MALLET, J., BARTON, N., LAMAS, G., SANTISTEBAN, J., MUEDAS, M. and EELEY, H. (1990). Estimates of cline width and linkage disequilibrium in *Helicous* hybrid zones. *Genetics* **124** 921–936.
- OHTA, T. and KIMURA, M. (1969). Linkage disequilibrium at steady state determined by random genetic drift and recurrent mutation. *Genetics* **63** 229–238.
- OHTA, T. (1982). Linkage disequilibrium with the island model. *Genetics* **101** 139–155.

- OHTA, T. (1982). Linkage disequilibrium due to random genetic drift in finite subdivided populations. *Proc. Natl. Acad. Sci. USA* **79** 1940–1944.
- RAND, D. M. and HARRISON, R. G. (1989). Ecological genetics of a mosaic hybrid zone: mitochondrial, nuclear, and reproductive differentiation by soil type. *Evolution* **43** 432–449.
- SCRIBNER, K. T. and AVISE, J. C. (1994). Population cage experiments with a vertebrate: genetics of hybridization in *Gambusia* fishes. *Evolution* **48** 155–171.
- SPOLSKY, C. and UZZEL, T. (1984). Natural interspecies transfer of mitochondrial DNA in amphibians. *Proc. Natl. Acad. Sci. USA* **81** 5802–5805.
- WRIGHT, S. (1968). *Evolutions and Genetics of Populations, Vol II. The Theory of Gene Frequencies*. University Press of Chicago.

DEPARTMENT OF MATHEMATICS AND COMPUTER SCIENCE
GEORGIA STATE UNIVERSITY
UNIVERSITY PLAZA
ATLANTA GA 30303
SDATTA@CS.GSU.EDU

DIFFUSION PROCESS CALCULATIONS FOR MUTANT GENES IN NONSTATIONARY POPULATIONS

BY RUZONG FAN AND KENNETH LANGE¹

University of Michigan

Diffusion process approximations were introduced into population genetics by Fisher and Wright and perfected by Kimura. Contrary to popular scientific opinion, these pioneers did not solve all of the interesting modeling problems. For instance, none of them has much to say about the stochastic dynamics of recessive disease genes. They are also more or less silent on the stochastic aspects of evolution in the presence of exponential population growth. The current paper uses Itô's formula to derive an infinite hierarchy of integral equations satisfied by the moments of a diffusion process. These integral equations can be converted into an infinite hierarchy of ordinary differential equations and solved either exactly or numerically. We illustrate some of the possibilities for dominant, neutral, and recessive models of inheritance by computing the moments of gene frequencies in the presence of exponential population growth.

1. Introduction. The evolutionary forces governing the distribution and dynamics of human genetic diseases can be modeled in a variety of ways. The earliest and most understandable models are deterministic (Cavalli-Sforza and Bodmer 1971, Crow and Kimura 1970, Ewens 1979, Nagylaki 1992). Later models attempt to capture the more subtle stochastic effects that inevitably come into play. For autosomal dominant or X-linked diseases, branching process models are ideal (Fisher 1930, Haldane 1927, Harris 1989, Skellam 1949). By viewing each new disease mutation as the progenitor of an independently evolving clan of deleterious gene carriers, one can answer a host of interesting population genetic questions (Gladstien and Lange 1978a, Gladstien and Lange 1978b, Lange and Gladstien 1980, Lange 1982). We have recently extended these branching process models to include exponential growth of the surrounding population of normal individuals (Lange and Fan 1997, Fan and Lange 1998).

For recessive diseases, selection occurs when carrier individuals from the same or different clans mate. Thus, the independence assumption of the branching process paradigm breaks down. Although the alternative Wright-Fisher model of evolution eschews the dubious assumption of independently evolving clans, it has yielded, contrary to popular scientific opinion, little insight into the balance between selection and mutation for recessive diseases (Crow and

¹ Research supported in part by USPHS grant GM53275.

AMS 1991 subject classifications. Primary 60J25, 60J60, 60J65, 60J70, 62P10; secondary 92D10.

Key words and phrases. Brownian motion, diffusion processes, Itô integral, mutant genes, population genetics, stochastic differential equations.

Kimura 1970, Ewens 1979). Even the enormously prolific Kimura was largely silent on the question of recessive diseases and the impact of population growth on their dynamics.

The present paper crafts some new calculational tools for the Wright-Fisher model. Following the lead of Kimura (Crow and Kimura 1970), we immediately pass to the diffusion approximation of the Wright-Fisher model. This puts the considerable machinery of stochastic integration at our disposal (Chung and Williams 1990). Within this framework, we derive infinite hierarchies of integral and ordinary differential equations for the moments of a diffusion process. When the infinitesimal mean of the diffusion process is linear in its spatial variable and the infinitesimal variance is quadratic, these equations can often be solved exactly (Fan *et al.* 1998). When the infinitesimal mean and variance are polynomials in their spatial variables, we show how the hierarchy of differential equations can still be solved numerically. These results are of independent interest quite apart from their applications in population genetics. Our solutions specifically incorporate time inhomogeneities such as growth of the surrounding normal population.

With these introductory comments in mind, Section 2 briefly reviews the Wright-Fisher model and its classical diffusion approximation to the frequency of a neutral or deleterious gene. Section 3 derives via Itô's formula the aforementioned infinite hierarchies of integral and differential equations for the moments of a diffusion process. Sections 4 and 5 then discuss exact moment calculation for the neutral and dominant versions of the Wright-Fisher model. Section 6 describes the numerical techniques used for the recessive disease moments featured in the examples of Section 7. Finally, our concluding discussion suggests limitations of the genetic models and raises open problems about the rigor of the numerical methods.

2. Wright-Fisher genetic model. The Wright-Fisher model for the evolution of a deleterious or neutral gene postulates (a) discrete generations, (b) finite population size, (c) no immigration, and (d) formation of gametes by random binomial sampling. In assumption (d), each current population member contributes to an infinite pool of potential gametes in proportion to his or her fitness. Mutation from the normal allele a to the deleterious allele b takes place at this stage with mutation rate η ; backmutation is not permitted. In the neutral model we neglect mutation and treat the two alleles symmetrically. Once the pool of potential gametes is formed, actual gametes are sampled randomly. Although the three genotypes occur in the usual Hardy-Weinberg proportions just after gamete sampling, selection causes allele frequencies to change over time.

If we let $w_{a/a}$, $w_{a/b}$, and $w_{b/b}$ denote the average fitnesses of the three genotypes a/a , a/b and b/b of an autosomally determined trait, then for a dominant disease we may suppose that $w_{a/a} = 1$ and $w_{a/b} = w_{b/b} = f < 1$. For a neutral trait, $w_{a/a} = w_{a/b} = w_{b/b} = 1$, and for a recessive disease $w_{a/a} = w_{a/b} = 1$ and $w_{b/b} = f < 1$. For our purposes, the population size N_m at generation m need not be constant. The primary object of study in this paper is the frequency X_m of allele b at generation m . This frequency is the ratio of the total number Y_m of b alleles to the total number of genes $2N_m$. The Wright-Fisher model specifies that Y_m is binomially distributed with $2N_m$ trials and success probability $p(X_{m-1})$ determined by the proportion $p(X_{m-1})$ of b alleles in the pool of potential gametes for generation m . In passing to a diffusion approximation, we take one generation as the unit of time and substitute

$$\begin{aligned}\mu(m, x_m) &= \mathbb{E}(X_{m+1} - X_m \mid X_m = x_m) \\ &= p(x_m) - x_m \\ \sigma^2(m, x_m) &= \text{Var}(X_{m+1} - X_m \mid X_m = x_m) \\ &= \frac{p(x_m)[1 - p(x_m)]}{2N_{m+1}}\end{aligned}$$

for the infinitesimal mean $\mu(t, x)$ and variance $\sigma^2(t, x)$ of the diffusion process evaluated at time $t = m$ and position $x = x_m$ (Crow and Kimura 1970, Ewens 1979). Given the assumption of exponential growth in a human population, the population size at generation m is $N_m = N_0 e^{cm}$ for some growth rate c .

Under neutral evolution, the gamete success probability is $p(x) = x$. This formula for $p(x)$ entails no systematic tendency for either allele to expand at the expense of the other allele. For a dominant disease, $p(x) = \eta + fx$, which implies an equilibrium frequency of $x_\infty = \frac{\eta}{1-f}$ in the corresponding deterministic model. Finally, for a recessive disease, $p(x) = \eta + x - (1-f)x^2$, which implies an equilibrium frequency of $x_\infty = \sqrt{\frac{\eta}{1-f}}$. These formulas and the approximations made in deriving them are discussed in the references (Crow and Kimura 1970, Ewens 1979, Fan and Lange 1998, Lange 1997). Most population geneticists substitute $p(x) = x$ in the infinitesimal variance $\sigma^2(t, x)$. This action is justified for neutral and recessive inheritance, but less so for dominant inheritance where the allele frequency x is typically on the order of magnitude of the mutation rate η .

For the neutral and dominant models, the infinitesimal mean $\mu(x, t)$ is linear in x , and the infinitesimal variance $\sigma^2(t, x)$ is quadratic. As we shall see, these facts enable one to calculate the moments of the diffusion approximation X_t to the discrete process X_n exactly. For the recessive model, $\mu(t, x)$ is unfortunately

quadratic in x . This increase in the degree of $\mu(t, x)$ hinders exact calculation of moments. However, we will show how to compute moments numerically even in this harder case.

3. Moment equations for diffusion processes. We consider a general diffusion process generated by a stochastic differential equation. Let (Ω, \mathcal{F}, P) be a complete probability space with a right-continuous increasing family $(\mathcal{F}_t)_{t \geq 0}$ of sub σ -fields of \mathcal{F} , each of which contains all P -null sets. If B_t is standard Brownian motion, X_0 is an \mathcal{F}_0 -measurable random variable independent of B_t , and $\sigma(s, x)$ and $\mu(s, x)$ are sufficiently smooth functions, then the solution X_t to the stochastic differential equation

$$(3.1) \quad X_t = X_0 + \int_0^t \sigma(s, X_s) dB_s + \int_0^t \mu(s, X_s) ds,$$

exists and is an \mathcal{F}_t -adapted continuous process. The smoothness assumptions on $\sigma(s, x)$ and $\mu(s, x)$ are Lipschitz conditions that need not concern us here (Chung and Williams 1990). We will require that X_t be square integrable and, indeed, possess any higher order moments mentioned below. The stochastic integral involved in equation (3.1) can be either Itô's or Stratonovich's integral. Itô's integral can be transformed to Stratonovich's by a change of variables and vice versa. We prefer Itô's integral because it allows us to use the infinitesimal means and variances directly.

If we apply Itô's formula to equation (3.1) with the transformation function $f(x) = e^{iux}$ (Chung and Williams 1990), then we find that

$$\begin{aligned} e^{iux_t} &= e^{iux_0} + iu \int_0^t e^{iux_s} \sigma(s, X_s) dB_s + iu \int_0^t e^{iux_s} \mu(s, X_s) ds \\ &\quad - \frac{u^2}{2} \int_0^t e^{iux_s} \sigma^2(s, X_s) ds. \end{aligned}$$

Taking expectations now yields the equation

$$\begin{aligned} E(e^{iux_t}) &= E(e^{iux_0}) + iu \int_0^t E[e^{iux_s} \mu(s, X_s)] ds \\ (3.2) \quad &\quad - \frac{u^2}{2} \int_0^t E[e^{iux_s} \sigma^2(s, X_s)] ds \end{aligned}$$

for the characteristic function of X_t . Repeatedly differentiating equation (3.2) with respect to u and evaluating the results at $u = 0$ produces the hierarchy of integral equations

$$(3.3) \quad E(X_t) = E(X_0) + \int_0^t E[\mu(s, X_s)] ds$$

$$(3.4) \quad \begin{aligned} \mathbb{E}(X_t^n) &= \mathbb{E}(X_0^n) + n \int_0^t \mathbb{E}[X_s^{n-1} \mu(s, X_s)] ds \\ &\quad + \frac{n(n-1)}{2} \int_0^t \mathbb{E}[X_s^{n-2} \sigma^2(s, X_s)] ds, \quad n \geq 2. \end{aligned}$$

Finally differentiating equations (3.3) and (3.4) with respect to t gives the corresponding hierarchy of differential equations

$$(3.5) \quad \begin{aligned} \frac{d}{dt} \mathbb{E}(X_t) &= \mathbb{E}[\mu(t, X_t)] \\ \frac{d}{dt} \mathbb{E}(X_t^n) &= n \mathbb{E}[X_t^{n-1} \mu(t, X_t)] \\ (3.6) \quad &\quad + \frac{n(n-1)}{2} \mathbb{E}[X_t^{n-2} \sigma^2(t, X_t)], \quad n \geq 2. \end{aligned}$$

In some cases, these differential equations are tractable analytically. When they are intractable analytically, they may be tractable numerically.

Equation (3.4) for the second moment of X_t amounts to

$$(3.7) \quad \mathbb{E}(X_t^2) - \mathbb{E}(X_0^2) = 2 \int_0^t \mathbb{E}[X_s \mu(s, X_s)] ds + \int_0^t \mathbb{E}[\sigma^2(s, X_s)] ds.$$

To recast this as an equation for the variance (Fan *et al.* 1998), note that equation (3.3) entails $d\mathbb{E}(X_t) = \mathbb{E}[\mu(t, X_t)]dt$. Hence, the fundamental theorem of calculus implies

$$\begin{aligned} \mathbb{E}(X_t)^2 - \mathbb{E}(X_0)^2 &= 2 \int_0^t \mathbb{E}(X_s) d\mathbb{E}(X_s) \\ &= 2 \int_0^t \mathbb{E}(X_s) \mathbb{E}[\mu(s, X_s)] ds. \end{aligned}$$

Subtracting this identity from equation (3.7) gives the promised variance equation

$$(3.8) \quad \begin{aligned} \text{Var}(X_t) - \text{Var}(X_0) \\ = \int_0^t \mathbb{E}[\sigma^2(s, X_s)] ds + 2 \int_0^t \text{Cov}[X_s, \mu(s, X_s)] ds. \end{aligned}$$

If $\mu(t, x)$ is linear in x and $\sigma^2(t, x)$ is quadratic in x , then the differential equations (3.5) and (3.6) take the form

$$(3.9) \quad y'(t) = f(t) + g(t)y(t).$$

In many cases of interest, the solution

$$(3.10) \quad y(t) = e^{\int_0^t g(s)ds} \left(y(0) + \int_0^t f(s) e^{-\int_0^s g(u)du} ds \right).$$

can be explicitly calculated (Fan *et al.* 1998). This is certainly true for the neutral and dominant Wright-Fisher models. However, for the recessive Wright-Fisher model, the infinitesimal mean $\mu(t, x)$ is quadratic in x .

If $\mu(t, x)$ is quadratic in x , then the hierarchy of moment equations (3.3) and (3.4) is coupled in the sense that lower order moment functions depend on higher order moment functions. This fact prevents one from solving the equations recursively. However, numerical progress can be made if we view the hierarchy of equations as a single infinite-dimensional differential equation. For the sake of concreteness, suppose that

$$(3.11) \quad \begin{aligned} \mu(t, x) &= \mu_0(t) + \mu_1(t)x + \mu_2(t)x^2 \\ \sigma^2(t, x) &= \sigma_0(t) + \sigma_1(t)x + \sigma_2(t)x^2, \end{aligned}$$

and let $M(t)$ be the infinite-dimensional column vector whose n th entry is $m_n(t) = E(X_t^n)$ for $0 \leq n < \infty$. Then the hierarchy of equations (3.5) and (3.6) can be written as the single differential equation

$$(3.12) \quad \frac{d}{dt} M(t) = A(t)M(t),$$

where the infinite-dimensional matrix $A(t) = A_1(t) + A_2(t)$ is the sum of the two infinite-dimensional matrices

$$A_1(t) = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\ \mu_0(t) & \mu_1(t) & \mu_2(t) & 0 & 0 & 0 & \cdots \\ 0 & 2\mu_0(t) & 2\mu_1(t) & 2\mu_2(t) & 0 & 0 & \cdots \\ 0 & 0 & 3\mu_0(t) & 3\mu_1(t) & 3\mu_2(t) & 0 & \cdots \\ 0 & 0 & 0 & 4\mu_0(t) & 4\mu_1(t) & 4\mu_2(t) & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

$$A_2(s) = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & 0 & 0 & \cdots \\ \sigma_0(t) & \sigma_1(t) & \sigma_2(t) & 0 & 0 & \cdots \\ 0 & \frac{3 \cdot 2}{2} \sigma_0(t) & \frac{3 \cdot 2}{2} \sigma_1(t) & \frac{3 \cdot 2}{2} \sigma_2(t) & 0 & \cdots \\ 0 & 0 & \frac{4 \cdot 3}{2} \sigma_0(t) & \frac{4 \cdot 3}{2} \sigma_1(t) & \frac{4 \cdot 3}{2} \sigma_2(t) & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

If $\mu(t, x)$ and $\sigma^2(t, x)$ are general polynomials in x rather than quadratics, then the same differential equation (3.12) holds provided we modify $A(t)$ in the obvious manner. For the sake of simplicity, we confine our attention to the quadratic case.

4. Neutral model moments. In the neutral genetic model with exponential population growth, the infinitesimal mean and variance are

$$\mu(t, x) = 0 \quad \sigma^2(t, x) = \frac{x(1-x)}{2N_0 e^{ct}}.$$

Equation (3.3) makes it evident that $m_1(t) = E(X_t) = E(X_0)$ for all $t \geq 0$. In view of equation (3.6), the n th moment $m_n(t)$ of X_t satisfies

$$(4.1) \quad \frac{d}{dt}m_n(t) = -\frac{n(n-1)[m_n(t) - m_{n-1}(t)]}{4N_0 e^{ct}}.$$

To solve equation (4.1), we make the change of variables $s = e^{-ct} - 1$ and $\nu_n(s) = m_n(t)$. Because $ds = -ce^{-ct}dt$, this substitution yields Kimura's (1955) equation

$$(4.2) \quad \frac{d}{ds}\nu_n(s) = \frac{n(n-1)[\nu_n(s) - \nu_{n-1}(s)]}{4N_0 c}$$

for a stationary neutral process.

Now consider the trial solution

$$(4.3) \quad \nu_n(s) = \sum_{i=0}^{n-1} c_{ni} e^{\lambda_i s}.$$

The fact that this sum has upper limit $n - 1$ will determine the eigenvalues λ_i . If we differentiate the trial solution (4.3), compare the result to equation (4.2), and equate coefficients of $e^{\lambda_i s}$, then we find that

$$\lambda_i c_{ni} = \frac{n(n-1)}{4N_0 c} (c_{ni} - c_{n-1,i}).$$

This determines c_{ni} as

$$(4.4) \quad c_{ni} = \frac{n(n-1)}{n(n-1) - 4N_0 c \lambda_i} c_{n-1,i}$$

unless $n(n-1) = 4N_0 c \lambda_i$. We want this exceptional condition to occur when $i = n - 1$ because then the requirement $c_{n-1,n-1} = 0$ imposes no constraint on $c_{n,n-1}$. Thus, we take

$$\lambda_{n-1} = \frac{n(n-1)}{4N_0 c}.$$

The coefficients c_{ni} can now be found by invoking the initial conditions. Clearly, $c_{10} = m_1(0)$. Suppose we know the coefficients $c_{n-1,0}, \dots, c_{n-1,n-2}$ determining $\nu_{n-1}(s)$. Then the coefficients $c_{n0}, \dots, c_{n,n-2}$ can be computed via the recurrence relation (4.4). The final coefficient $c_{n,n-1}$ is determined by the initial condition $m_n(0) = \sum_{i=0}^{n-1} c_{ni}$. These considerations allow us to calculate, for example,

$$\begin{aligned}\nu_2(s) &= m_1(0) + [m_2(0) - m_1(0)]e^{\frac{s}{2N_0c}} \\ \nu_3(s) &= m_1(0) + \frac{3}{2}[m_2(0) - m_1(0)]e^{\frac{s}{2N_0c}} \\ &\quad + \frac{1}{2}[2m_3(0) - 3m_2(0) + m_1(0)]e^{\frac{3s}{2N_0c}} \\ \nu_4(s) &= m_1(0) + \frac{9}{5}[m_2(0) - m_1(0)]e^{\frac{s}{2N_0c}} \\ &\quad + [2m_3(0) - 3m_2(0) + m_1(0)]e^{\frac{3s}{2N_0c}} \\ &\quad + \frac{1}{5}[5m_4(0) - 10m_3(0) + 6m_2(0) - m_1(0)]e^{\frac{3s}{N_0c}}.\end{aligned}$$

Kimura (1955) gives explicit expressions for the coefficients c_{ni} when $X_0 = p$ is constant and $m_n(0) = p^n$ for all n .

Similar reasoning enables one to find not only the moments but also the density function $f(t, x)$ of the neutral Wright-Fisher process X_t . It is well known that $f(t, x)$ satisfies the Fokker-Planck or Kolmogorov forward equation (Feller 1951)

$$\frac{\partial}{\partial t}f(t, x) = \frac{1}{4N_0e^{ct}}\frac{\partial^2}{\partial x^2}[x(1-x)f(t, x)].$$

The change of variables $s = 1 - e^{-ct}$ and $g(s, x) = f(t, x)$ transforms this partial differential equation into the corresponding partial differential equation

$$\frac{\partial}{\partial s}g(s, x) = \frac{1}{4N_0c}\frac{\partial^2}{\partial x^2}[x(1-x)g(s, x)].$$

for neutral evolution in a stationary population. Crow and Kimura (1970) explain how $g(s, x)$ and therefore $f(t, x)$ can be expanded in terms of appropriate eigenfunctions.

5. Dominant model moments. For a dominant disease, the infinitesimal mean $\mu(t, x) = \eta - (1-f)x$. Hence, the differential equation (3.5) for the first moment $m_1(t) = E(X_t)$ becomes

$$\frac{d}{dt}m_1(t) = \eta - (1-f)m_1(t)$$

with solution

$$m_1(t) = \left[m_1(0) - \frac{\eta}{1-f} \right] e^{-(1-f)t} + \frac{\eta}{1-f}.$$

Because the infinitesimal variance is $\sigma^2(t, x) = \frac{(\eta+fx)(1-\eta-fx)}{2N_0e^{ct}}$ under exponential growth, the differential equation (3.6) for the second moment $m_2(t)$ amounts to

$$\begin{aligned} \frac{d}{dt}m_2(t) &= 2\eta m_1(t) - 2(1-f)m_2(t) \\ &\quad + \frac{1}{2N_0e^{ct}} [\eta(1-\eta) + f(1-2\eta)m_1(t) - f^2m_2(t)]. \end{aligned}$$

Obviously, this differential equation is the special case of equation (3.9) with $y(t) = m_2(t)$ and

$$\begin{aligned} f(t) &= 2\eta m_1(t) + \frac{1}{2N_0e^{ct}} [\eta(1-\eta) + f(1-2\eta)m_1(t)] \\ g(t) &= -2(1-f) - \frac{f^2}{2N_0e^{ct}}. \end{aligned}$$

The solution (3.10) to the differential equation (3.9) includes the integral

$$\int_0^t g(s)ds = -2(1-f)t - \frac{f^2}{2cN_0}(1 - e^{-ct})$$

and the integral $\int_0^t f(s)e^{-\int_0^s g(u)du}ds$. The latter integral decomposes as a linear combination of terms of the kind

$$\int_0^t e^{-\alpha s - \beta e^{-cs}} ds = \frac{1}{c} \int_0^{ct} e^{-\frac{\alpha}{c}u - \beta e^{-u}} du$$

for $\beta = \frac{f^2}{2cN_0}$ and various choices of the constant α . One can evaluate each such term via the special function

$$\begin{aligned} I_{\alpha,\beta}(t) &= \int_0^t e^{-\alpha s - \beta e^{-s}} ds \\ &= \sum_{k=0}^{\infty} \int_0^t e^{-\alpha s} \frac{(-\beta e^{-s})^k}{k!} ds \\ &= \sum_{k=0}^{\infty} \frac{(-\beta)^k}{k!(\alpha+k)} [1 - e^{-(\alpha+k)t}]. \end{aligned}$$

Because β is small in practice, the above series converges quickly. Somewhat tedious algebra shows that

$$\begin{aligned} & \int_0^t f(s) e^{-\int_0^s g(u) du} ds \\ &= \frac{e^\beta}{c} \left\{ 2\eta \left[m_1(0) - \frac{\eta}{1-f} \right] I_{\frac{f-1}{c}, \beta}(ct) + \frac{2\eta^2}{1-f} I_{\frac{2(f-1)}{c}, \beta}(ct) \right. \\ &+ \frac{f(1-2\eta)}{2N_0} \left[m_1(0) - \frac{\eta}{1-f} \right] I_{\frac{f-1}{c}+1, \beta}(ct) \\ &+ \left. \frac{\eta(1-\eta-f\eta)}{2N_0(1-f)} I_{\frac{2(f-1)}{c}+1, \beta}(ct) \right\}. \end{aligned}$$

6. Numerical methods. Over a short time interval dt , the differential equation (3.12) entails the Euler approximation

$$(6.1) \quad M(t + dt) \approx [I + dtA(t)]M(t),$$

where I is the identity matrix. If we partition the interval $[0, t]$ into n subintervals $[i\delta_n, (i+1)\delta_n]$ for $i = 0, \dots, n-1$ and $\delta_n = \frac{t}{n}$, then the approximation (6.1) propagates into Euler's method

$$(6.2) \quad M(t) \approx \prod_{i=0}^{n-1} [I + \delta_n A(i\delta_n)]M(0)$$

of solving for $M(t)$. With luck, the expression on the left of (6.2) will tend to $M(t)$ as n tends to ∞ . Mathematical justification of limits of this type belongs to the province of product integration (Dollard and Friedman 1979, Gill and Johansen 1990). If $A(t) = A$ does not depend on the time parameter t , then the product integral

$$M(t) = \prod_{s=0}^t e^{A(s)ds} M(0) = e^{tA} M(0)$$

coincides with multiplication by a matrix exponential.

Making the theory of product integration rigorous in the current context is difficult because $M(t)$ and $A(t)$ are infinite dimensional and $A(t)$ is unbounded. For practical purposes, we truncate $M(t)$ and $A(t)$ to their first k rows and columns and carry out all computations with the resulting finite-dimensional versions $M_k(t)$ and $A_k(t)$ of $M(t)$ and $A(t)$. The sparsity of the matrices $M_k(t)$ in our genetic examples obviously decreases both the computational complexity

and the storage requirements of the matrix-vector multiplications implied by formula (6.2).

In the case of neutral and dominant genes, where we have analytic results, the truncated system is easy to solve numerically. Unfortunately, for recessive genes with high initial frequencies, the truncated system poses more of a numerical challenge. In addition to Euler's method, we have tried a standard fourth-order Runge-Kutta scheme (Birkhoff and Rota 1978, Press *et al.* 1992) and the power series method sketched below. To achieve stable solutions, all three methods require short steps ($n \geq 20000$) and many moments ($k \geq 500$) for initial gene frequencies in excess of .005. We enhance the stability of each method by instituting three safeguards. First, we perform all computations in double precision. Second, at the end of each step, we reset all negative entries $m_j(t)$ of $M_k(t)$ to 0. In our models all moments are nonnegative, so presumably this tactic helps. Third, at the end of each step, we also exploit Hölder's moment inequality $m_j(t)^{1/j} \leq m_{j+1}(t)^{1/(j+1)}$ by replacing $m_{j+1}(t)$ by $\max\{m_{j+1}(t), m_j(t)^{(j+1)/j}\}$ recursively for $j = 1, 2, \dots, k - 1$.

To explain the series method, note that for a recessive disease in the presence of exponential population growth, the quadratic expressions (3.11) for the infinitesimal mean and variance have coefficients

$$\begin{bmatrix} \mu_0(t) \\ \mu_1(t) \\ \mu_2(t) \end{bmatrix} = \begin{bmatrix} \eta \\ 0 \\ -(1-f) \end{bmatrix}, \quad \begin{bmatrix} \sigma_0(t) \\ \sigma_1(t) \\ \sigma_2(t) \end{bmatrix} = \begin{bmatrix} 0 \\ \frac{e^{-ct}}{2N_0} \\ -\frac{e^{-ct}}{2N_0} \end{bmatrix}.$$

These expressions imply that the matrix $A(t)$ can be expanded in the power series

$$A(t) = B_1 + B_2 e^{-ct} = \sum_{i=0}^{\infty} C_i t^i,$$

where $C_0 = B_1 + B_2$ and $C_i = \frac{(-c)^i}{i!} B_2$, $i \geq 1$, are constant matrices. If we formally expand the moment vector $M(t) = \sum_{i=0}^{\infty} D_i t^i$ in a similar power series and differentiate term by term, then equating coefficients of t^j in the differential equation (3.12) yields the recurrence

$$D_{j+1} = \frac{1}{j+1} \sum_{i=0}^j C_i D_{j-i}.$$

Together with the initial condition $D_0 = M(0)$, this gives an effective method of computing the series expansion of $M(t)$ (Apostol 1969). For t reasonably small, we can terminate the expansion after a few terms, say five, and still retain a

good approximation to $M(t)$. For larger t , we choose n large and approximate $M(t/n)$, use this as an initial value to approximate $M(2t/n)$, and so forth, until we finally recover $M(t)$. Once again we must operate on truncated vectors and matrices.

The fact that all three solution methods ultimately provide similar answers increases our overall confidence that we can compute the moment vector $M(t)$ accurately. However, we are still far from fully understanding the analytic and numerical behavior of the moment differential equations. More research is clearly needed.

7. Examples. To illustrate the theory, we now turn to some concrete examples based on the demographic history of Finland [Hästbacka *et al.* (1992)]. Finland was settled around 2000 years ago by a founding population of about 1000 people. Given a current Finnish population of 5 million people and a generation time of 25 years, this implies an exponential growth rate of $c = \frac{1}{80} \ln(5000) = .1065$ per generation. In Figures 1 through 5, we consider the evolution of the *b* allele in three simple biallelic genetic models. For the sake of completeness, our graphs extrapolate 20 generations into the future.

Figure 1 plots the coefficient of variation $\kappa_2(t)^{\frac{1}{2}}/\kappa_1(t)$, skewness $\kappa_3(t)/\kappa_2(t)^{\frac{3}{2}}$, and kurtosis $\kappa_4(t)/\kappa_2(t)^2$ of the frequency X_t of the *b* allele under the neutral Wright-Fisher model. Here $\kappa_j(t)$ is the j th cumulant of X_t at time t . For a normally distributed random variable, skewness and kurtosis are both 0. Because neither selection nor mutation operate in this model, the mean of X_t and its deterministic analog remain fixed at our chosen initial value $X_0 = .1$. The figure makes it evident that the variance first increases sharply and later flattens out. This behavior confirms our intuition that most of the stochastic effects take place in the early generations when the population size is small. The nontrivial skewness and kurtosis that develop during this period are eventually frozen into place by the exponential growth of the population.

Figure 2 plots the mean of X_t and its deterministic analog for a dominant disease allele with a mutation rate $\eta = 10^{-6}$, a fitness $f = .9$, and an initial gene frequency $X_0 = 5 \times 10^{-4}$. The upper and lower bands in this Figure are the curves $\max\{0, \kappa_1(t) \pm 2\kappa_2(t)^{\frac{1}{2}}\}$. The initial gene frequency corresponds to one affected person among the 1000 founders. With such a high fitness, the effects of the affected founder persist for many generations. Eventually, however, the balance between selection and mutation asserts itself, and the mean approaches its low equilibrium level. Figure 3 shows that in the process a quasi-stochastic equilibrium develops with decreasing variance and low skewness and kurtosis. The large skewness and kurtosis seen in early generations presumably reflect the nonnegligible probability that the affected population founder generates a

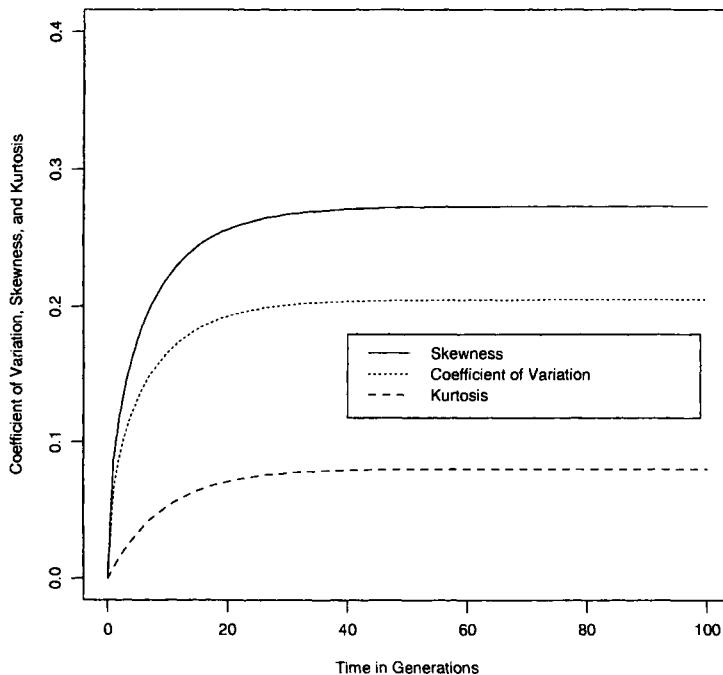


FIG. 1. *Coefficient of Variation, Skewness, and Kurtosis of the Frequency of a Neutral Gene when $X_0 = .1$.*

very large clan of affected descendants.

Finally, Figures 4 and 5 depict the dynamics of a recessive disease with a mutation rate $\eta = 10^{-6}$, a fitness $f = .5$, and an initial gene frequency $X_0 = .005$. Figure 4 displays good agreement between the mean of X_t and its deterministic analog. Under the pressure of selection, both are slowly tending to the deterministic equilibrium. Just as with a neutral gene, there are large stochastic effects in early generations that persist for the duration of Figures 4 and 5. The interesting behavior of the kurtosis curve in Figure 5 is difficult to rationalize. Possibly it is a numerical artifact, but our calculations appear stable when the step size is small enough and sufficiently many moments are taken into account.

8. Discussion. The diffusion process models familiar to population geneticists almost invariably assume a stationary population (Crow and Kimura 1970, Ewens 1979, Feller 1951). However, human populations tend to display exponential growth with episodes of decline brought on by famine, plague, and war.

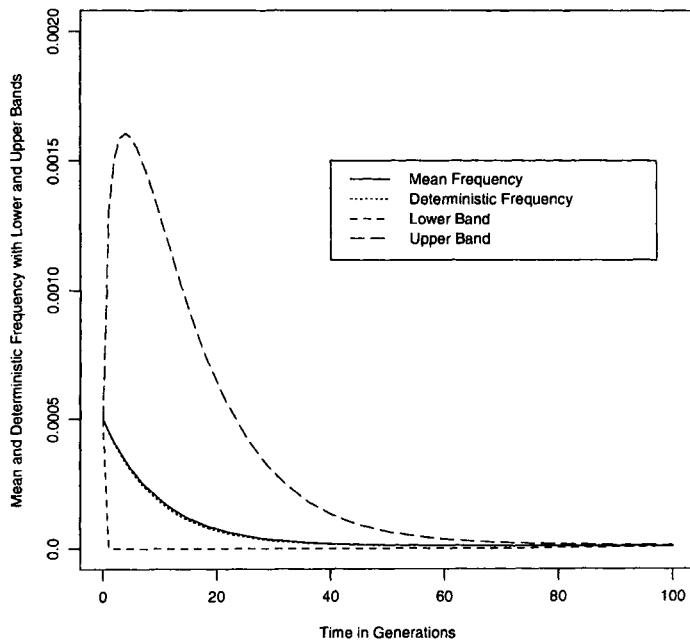


FIG. 2. *Mean and Deterministic Value for the Frequency of a Dominant Gene with $\eta = 10^{-6}$, $f = .90$, and $X_0 = 5 \times 10^{-4}$.*

In the current paper, we extend to exponentially growing populations some of the moment calculations for gene frequencies previously carried out under the stationary Wright-Fisher model. We also fill in some mathematical gaps in the treatment of recessive diseases.

We view diffusion process models as complementary to branching process models for neutral and dominant disease genes. As emphasized in our introduction, branching processes are poor vehicles for modeling recessive diseases. The Wright-Fisher model overcomes this defect, but at the cost of introducing diffusion approximations and a specific sampling framework for generating gametes. The binomial sampling premise of the Wright-Fisher effectively requires that each parent produces a Poisson number of children. This offspring assumption probably underestimates the variance in the number of children per parent.

The example featured in Figures 2 and 3 clearly illustrates how quickly the deterministic balance between selection and mutation is reached for a dominant disease. The frequency of the disease allele rapidly tends to its deterministic equilibrium with little stochastic variation left in later generations. A low mu-

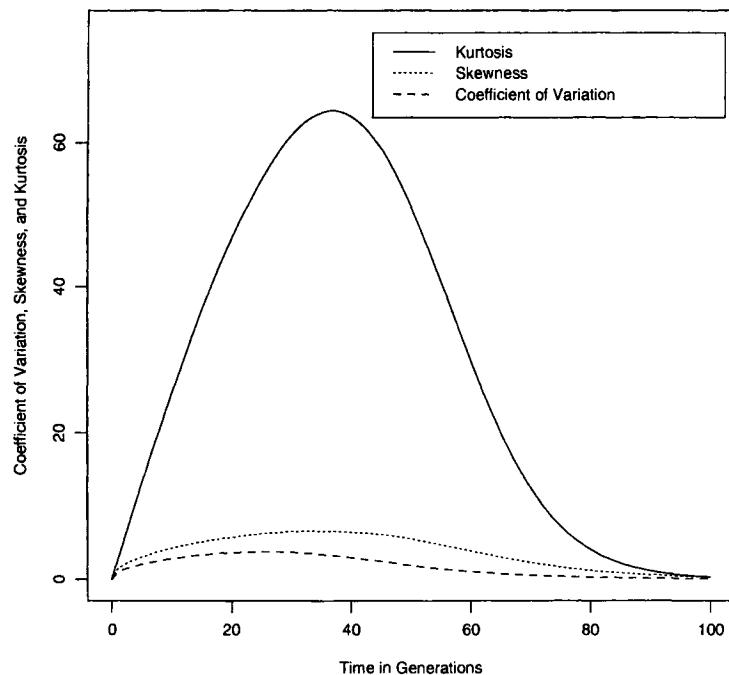


FIG. 3. *Coefficient of Variation, Skewness, and Kurtosis of the Frequency of a Dominant Gene when $\eta = 10^{-6}$, $f = .90$, and $X_0 = 5 \times 10^{-4}$.*

tation rate and a high fitness retard the approach of a dominant gene to its deterministic equilibrium. It is noteworthy that cumulants (means, variances, skewness, and kurtosis) calculated by diffusion methods and by branching process methods are comparable for dominant diseases. Indeed, the cumulant results from the two branching process models in Lange and Fan (1997) appear to bracket the cumulants results from the diffusion process model. Obviously, this check is impossible for a recessive disease.

Neutral and recessive genes operate on an entirely different time scale than dominant genes. Stochastic fluctuations are considerable in a small population, and the large variance that develops in early generations persists for many generations to come. This suggests that predictions from the standard deterministic models be treated with extreme caution. Although population geneticists are well aware of this fact, it constantly needs to be reiterated for geneticists lacking relevant training.

Our methods for calculating the moments of a diffusion process should be

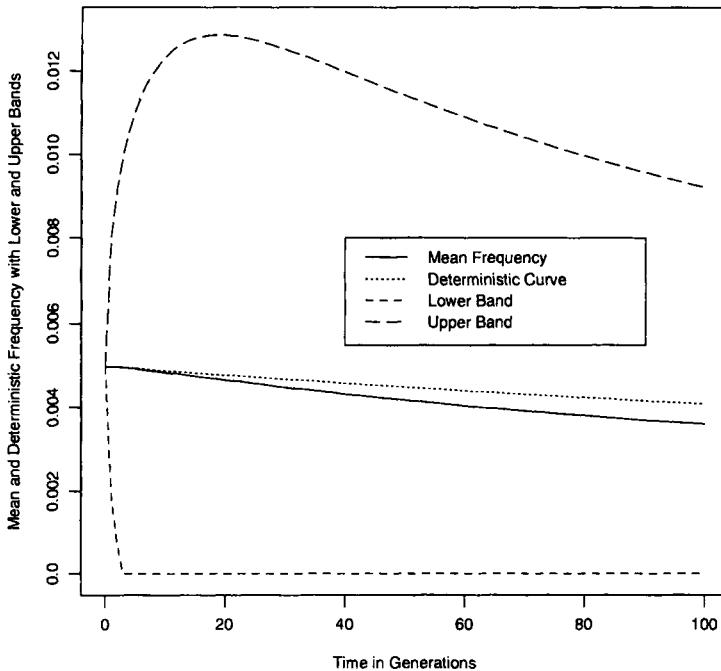


FIG. 4. *Mean and Deterministic Value for the Frequency of a Recessive Gene with $\eta = 10^{-6}$, $f = .5$, and $X_0 = .005$.*

of general interest. The versatility of the methods in the face of time inhomogeneities and polynomial dependence of the infinitesimal means and variances is a major advantage. Putting our numerical methods on a firmer theoretical foundation is clearly the next order of business. The techniques of functional analysis such as the Hille-Yosida theorem for continuous semigroups of operators offer one line of attack (Yosida 1980).

Extensions of the moment calculations to multivariate diffusion processes are also worth pursuing. In previous papers (Lange and Fan 1997, Fan and Lange 1998), we have set down branching process models that illuminate some of the issues in haplotype mapping of disease genes. To extend these calculations to recessive diseases, we must contend with multivariate versions of the Wright-Fisher model. The necessary particle types are intact and recombined chromosomes that carry ancestral disease mutations.

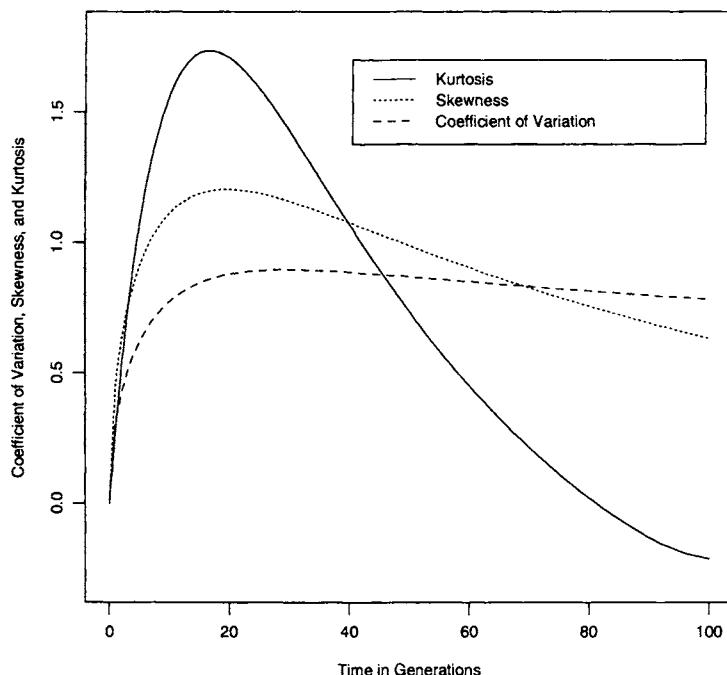


FIG. 5. *Coefficient of Variation, Skewness, and Kurtosis of the Frequency of a Recessive Gene when $\eta = 10^{-6}$, $f = .5$, and $X_0 = 0.005$.*

REFERENCES

- APOSTOL, T.M. (1969). *Calculus, Vol. II.* 2nd ed., pp 217–221. Wiley, New York.
- BIRKHOFF, G. and ROTA, G.-C. (1978). *Ordinary Differential Equations.* 3rd ed. Wiley, New York.
- CAVALLI-SFORZA, L.L. and BODMER, W.F. (1971). *The Genetics of Human Populations.* Freeman, San Francisco.
- CHUNG, K.L. and WILLIAMS, R.J. (1990). *Introduction to Stochastic Integration.* 2nd ed., p 94. Birkhäuser, Boston.
- CROW, J.F. and KIMURA, M. (1970). *An Introduction to Population Genetics Theory.* pp 367–432. Harper & Row, New York.
- DOLLARD, J.D. and FRIEDMAN, C.N. (1979). *Product Integration with Application to Differential Equations.* Addison-Wesley, Reading, MA.
- EWENS, W.J. (1979). *Mathematical Population Genetics.* pp 138–175. Springer-Verlag, New York.
- FAN, R.Z. and LANGE, K. (1998). Models for haplotype evolution in a nonstationary population. *Theor. Pop. Biol.* (in press).
- FAN, R.Z., LANGE, K. and PENA, E. (1998). A note on variation in point processes and damage models (submitted).
- FELLER, W. (1951). Diffusion processes in genetics. *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability.* pp 227–246. University of California Press,

Berkeley, CA.

- FISHER, R.A. (1930). The distribution of gene ratios for rare mutations. *Proc. Roy. Soc. Edinburgh* **50** 205–220.
- GILL, R.D. and JOHANSEN, S. (1990). A survey of product-integration with a view toward application in survival analysis. *Annals Stat.* **18** 1501–1555.
- GLADSTIEN, K. and LANGE, K. (1978a). Number of people and number of generations affected by a single deleterious mutation. *Theor. Pop. Biol.* **14** 313–321.
- GLADSTIEN, K. and LANGE, K. (1978b). Equilibrium distributions for deleterious genes in large stationary populations. *Theor. Pop. Biol.* **14** 322–328.
- HALDANE, J.B.S. (1927). A mathematical theory of natural and artificial selection, Part V: Selection and mutation. *Proc. Camb. Phil. Soc.* **23** 838–844.
- HARRIS, T.E. (1989). *The Theory of Branching Processes*. Dover, New York, p 99.
- HÄSTBACKA, J., de la CHAPELLE, A., KAITILA, I., SISTONEN, P., WEAVER, A. and LANDER, E. (1992). Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland. *Nature Genet.* **2** 204–211.
- KIMURA, M. (1955). Solution of a process of random genetic drift with a continuous model. *Proc. Natl. Acad. Sci. USA* **41** 144–50.
- LANGE, K. (1982). Calculation of the equilibrium distribution for a deleterious gene by the finite Fourier transform. *Biometrics* **38** 79–86.
- LANGE, K. (1997). *Mathematical and Statistical Methods for Genetic Analysis*. Springer-Verlag, New York.
- LANGE, K. and FAN, R.Z. (1997). Branching process models for mutant genes in nonstationary populations. *Theor. Pop. Biol.* **51** 118–133.
- LANGE, K. and GLADSTIEN, K. (1980) Further characterization of the long-run population distribution of a deleterious gene. *Theor. Pop. Biol.* **18** 31–43.
- NAGYLAKI, T. (1992). *Introduction to Theoretical Population Genetics*. Springer-Verlag, Berlin.
- PRESS, W.H., TEUKOLSKY, S.A., VETTERLING, W.T. and FLANNERY, B.P. (1992). *Numerical Recipes in Fortran: The Art of Scientific Computing*. 2nd ed. Cambridge University Press, Cambridge.
- SKELLAM, J.G. (1949). The probability distribution of gene-differences in relation to selection, mutation, and random extinction. *Proc. Camb. Phil. Soc.* **45** 364–367.
- YOSIDA, K. (1980). *Functional Analysis*. 6th ed. Springer-Verlag, New York.

DEPARTMENT OF BIOSTATISTICS
SCHOOL OF PUBLIC HEALTH
THE UNIVERSITY OF MICHIGAN
ANN ARBOR, MI 48109-2029
RFAN@UMICH.EDU

DEPARTMENT OF MATHEMATICS
THE UNIVERSITY OF MICHIGAN
ANN ARBOR, MI 48109-2029
KLANGE@UMICH.EDU

THE COALESCENT WITH PARTIAL SELFING AND BALANCING SELECTION: AN APPLICATION OF STRUCTURED COALESCENT PROCESSES

BY MAGNUS NORDBORG

Lund University

As a demonstration of a generally applicable technique, a theorem based on separation of time scales in the structured coalescent is used to extend results for the coalescent process with balancing selection to allow partial selfing. The resulting model behaves like the random-mating one, but with different rates of coalescence and recombination. This result has important implications for attempts to locate selectively maintained polymorphisms. Such polymorphisms can in principle be detected through their effect on the pattern of polymorphism in the genomic region surrounding the site under selection, however this is not practically feasible unless the effected region is sufficiently large. An implication of the present results is that the region is expected to be much larger in partially selfing organisms than in outcrossing ones, suggesting that studies attempting to locate selectively maintained polymorphisms should utilize selfing organisms.

1. Introduction. If natural selection has maintained polymorphism at a certain site or locus for a long period of time, the evolutionary dynamics in closely linked regions of the chromosome will be effected. In particular, the expected pattern of neutral polymorphism will be altered in a manner that allows inference about the action of selection directly from molecular polymorphism data, without phenotypic observation [Hudson and Kaplan (1988)]. This phenomenon has been observed in a few cases [Kreitman and Akashi (1995), Hudson (1996)]: MHC (immune system) loci in human; the *S* (self-incompatibility) locus in plants; and *Adh* (alcohol dehydrogenase) in *Drosophila melanogaster*. Although these loci were already known to harbor selectively maintained polymorphisms, the same phenomenon could, in principle, be used to locate such polymorphisms without prior information. For this to be practical, however, the region of the chromosome in which the effects of selection are noticeable must be large enough. One of the aims of this paper is to prove earlier claims [Nordborg et al. (1996), Nordborg (1997), Charlesworth et al. (1997)] that the effected regions will be much wider in partially selfing organisms than in outcrossing

Most of the research that led to this paper was done while the author was in the Department of Ecology & Evolution at the University of Chicago, initially as a research associate, and later as a visitor. Partial financial support was provided by National Science Foundation grants DEB-9217683 to D. Charlesworth and DEB-9350363 to J. Bergelson.

AMS 1991 subject classifications. Primary 60G35, 92D10; secondary 60F05.

Key words and phrases. Population genetics, molecular evolution, structured coalescent, selection, selfing, time scales.

ones.

The main result is a generalization of a previous treatment of the coalescent with selection [Kaplan et al. (1988), Hudson and Kaplan (1988)] to allow partial selfing. It is based on an argument about separation of time scales in the structured coalescent [Nordborg and Donnelly (1997), Nordborg (1997)], and utilizes a convergence theorem developed for this purpose by Möhle (1998). A second aim of this paper is to demonstrate the power of this approach to greatly simplify the analysis of quite complicated coalescent models. Connections with other approaches are discussed in Section 5.1.

2. The model. The model described in this section is effectively equivalent to previously used models [Kaplan et al. (1988), Hudson and Kaplan (1988), Hey (1991)], except for allowing partial selfing. We assume a population of N (assumed to be large and constant) diploid individuals that are hermaphroditic (*i. e.*, each individual produces gametes of both types, *e. g.*, pollen and ovules) and partially selfing (individuals can fertilize themselves; a more precise definition will be given below). The population has discrete generations. In each generation, all individuals produce infinitely many gametes which unite to form infinitely many zygotes. Mutation and recombination occur during gamete formation. The zygotes are then subject to selection, after which N of them are chosen to form the next generation of adults.

2.1. Forward dynamics at the selected locus. Consider a locus with two alleles \mathcal{A}_1 and \mathcal{A}_2 . The mutation probability (per gamete per generation) from \mathcal{A}_i to \mathcal{A}_j is u_{ij} . There are three possible genotypes $\mathcal{A}_1\mathcal{A}_1$, $\mathcal{A}_1\mathcal{A}_2$, and $\mathcal{A}_2\mathcal{A}_2$. Let $N_{ij}(t)$ be the (random) number of individuals in the adult population of genotype $\mathcal{A}_i\mathcal{A}_j$ in generation t : because $\sum N_{ij}(t) = N$, the $N_{ij}(t)$ are not independent. Define the genotype frequencies as $X_{ij}(t) = N_{ij}(t)/N$, and the allele frequencies as $Y_1(t) = X_{11}(t) + X_{12}(t)/2$ and $Y_2(t) = X_{22}(t) + X_{12}(t)/2$.

Let $y_i(t)$, $i \in \{1, 2\}$ be the allele frequencies among the gametes produced by the adults in generation t (*i. e.*, $y_i(t)$ is the proportion of gametes carrying \mathcal{A}_i). Because infinitely many gametes are produced, these are functions of the $Y_i(t)$. From standard population genetics theory, we have

$$(2.1) \quad y_1(t) = (1 - u_{12})Y_1(t) + u_{21}Y_2(t),$$

$$(2.2) \quad y_2(t) = (1 - u_{21})Y_2(t) + u_{12}Y_1(t).$$

Similarly, let $x_{ij}(t)$ be the zygotic frequencies. If mating were random, these frequencies could be found by simply multiplying the gamete frequencies. For a partially selfing population, however, it is assumed that only a fraction $1 - s$ of

the available female gametes are fertilized through random mating (outcrossed), and that the remaining fraction are fertilized by male gametes from the same individual. In other words, a fraction $1 - s$ of the zygotes are produced by random union of gametes within the population, where the gamete frequencies are $y_1(t)$ and $y_2(t)$, and a fraction s are produced by random union of gametes within individuals, so that

$$(2.3) \quad \begin{aligned} x_{11}(t) &= (1 - s)y_1^2(t) \\ &+ s[(1 - u_{12})^2 X_{11}(t) + (1 - u_{12} + u_{21})^2 X_{12}(t)/4 + u_{21}^2 X_{22}(t)], \end{aligned}$$

$$(2.4) \quad \begin{aligned} x_{12}(t) &= 2(1 - s)y_1(t)y_2(t) + s[(1 - u_{12} + u_{21})(1 - u_{21} + u_{12})X_{12}(t)/2 \\ &+ 2(1 - u_{12})u_{12}X_{11}(t) + 2(1 - u_{21})u_{21}X_{22}(t)], \end{aligned}$$

$$(2.5) \quad \begin{aligned} x_{22}(t) &= (1 - s)y_2^2(t) \\ &+ s[(1 - u_{21})^2 X_{22}(t) + (1 - u_{21} + u_{12})^2 X_{12}(t)/4 + u_{12}^2 X_{11}(t)]. \end{aligned}$$

The *viability* (relative chance of surviving to adulthood) of a zygote with genotype $\mathcal{A}_i\mathcal{A}_j$ is $1 - w_{ij}(t)$, where we do not exclude the possibility that $w_{ij}(t)$ is a function of the $x_{ij}(t)$ (as in frequency-dependent selection). Let $x_{ij}^*(t)$ denote the genotype frequencies after selection. Then

$$(2.6) \quad x_{ij}^*(t) = \frac{x_{ij}(t)(1 - w_{ij}(t))}{\bar{w}(t)},$$

where $\bar{w}(t) = 1 - \sum w_{ij}(t)x_{ij}(t)$.

Generation $t + 1$ is formed by drawing N individuals from the surviving zygotes. Conditional on the $N_{ij}(t)$, the $N_{ij}(t + 1)$ are thus multinomially distributed with parameters N and $x_{ij}^*(t)$.

2.2. Genealogy at a linked neutral locus. Our aim is to describe the gene genealogy at a locus linked to the selected site with recombination rate r . The locus is assumed to be neutral, *i.e.*, the only selection in the model is that on the locus described in Section 2.1. Recombination is only allowed *between* the loci. From a biological point of view, this requires that the length of the DNA sequences defined as “loci” be small relative to the distance between them. As is usual when tracing genealogies, time will be run backwards (the reverse direction from that in the previous section), so that generation ancestral to t is $t + 1$.

Consider a single chromosome, sampled from the adult population. With respect to the genotype of this adult, the sampled *instance* of the neutral gene is in one of three *genotypic* states. It is of course also characterized by being

linked to either an \mathcal{A}_1 or an \mathcal{A}_2 allele, which we will refer to as its *haplotypic* state. With respect to both classifications jointly, there are four possible states. Given this joint state in the present generation, what was the state of its ancestor in the previous generation? It should be clear that the transition probabilities with respect to the genotype of the ancestral individual can be calculated exactly from equations (2.3)–(2.5). To account for the haplotypic state in the previous generation we also need that if the current haplotypic state is $i \in \{1, 2\}$, then

1. if the ancestral individual was an \mathcal{A}_i homozygote, the haplotypic state cannot have changed;
2. if the ancestral individual was a \mathcal{A}_j homozygote ($j \neq i$), the haplotypic state must have changed (because a mutation occurred at the selected locus);
3. if the ancestral individual was a heterozygote, then the haplotypic state changed if there was either a mutation or a recombination event, but not both.

From this, the exact transition probabilities can be calculated.

Next, consider a sample of size n such instances. Each occupies one of the four states just described. In addition, they may or may not occupy the same individual as another instance. The possible genotypes for any individuals that harbor two instances is determined by the joint genotypic and haplotypic classification: for example, it is not possible for two instances to occupy the same heterozygote unless the sample contains two instances in heterozygotes, one linked to an \mathcal{A}_1 allele and the other linked to an \mathcal{A}_2 allele. The total number of states for a sample of size n is

$$\sum_{i=0}^{\lfloor n/2 \rfloor} \binom{n-2i+3}{3} \binom{i+2}{2}.$$

What about the transition probabilities from such a state in the current generation t to the possible states in the previous generation $t+1$? Under the assumptions of the model, any instance that occurs singly in an individual will “pick” its ancestral state independently of all other instances. If j , $2 \leq j \leq n$ instances pick the same genotype $\mathcal{A}_i \mathcal{A}_j$ in the generation $t+1$, then, by standard arguments, the probability that two of them pick the same parental individual is

$$(2.7) \quad \frac{\binom{j}{2}}{N_{ij}(t+1)} + O\left(\frac{1}{N_{ij}^2(t+1)}\right),$$

the probability that more than two pick the same individual is $O(1/N_{ij}^2(t+1))$, and the probability that they all pick different individuals is simply one minus

expression (2.7). Whenever two instances pick the same individual, one of two things happen. They either pick distinct haplotypes and thus their ancestors occupy the same individual in generation $t + 1$, or they pick the same haplotype, in which case they *coalesce*, and the number of distinct ancestors of the sample decreases permanently by one.

Two instances that currently occupy the same individual also pick their ancestral states independently of each other and all other ancestors, if we condition on that individual having resulted from outcrossing, which by definition is equivalent to random mating. If, on the other hand, an individual in the current generation that harbors two instances was the product of selfing, the two instances will behave precisely as if they had chosen a common parental individual through random mating, and they will thus either coalesce or continue to occupy the same individual as just described.

We will return to the transition probabilities below, however, two important remarks should be made in this context. The first is that, if $s = 0$, all individuals will always have resulted from random mating, and there is therefore no need to keep track of whether instances occupy the same individual or not. The second is that, going backwards in time, the ancestors of two instances that currently occupy the same individual will not continue doing so very long. For homozygotes this is so because if they result from selfing, then with probability one half the ancestors coalesce, and if they result from outcrossing, then the ancestors no longer occupy the same individual. For heterozygotes coalescence is considerably less likely, because it necessitates a recombination or mutation event, but it will become clear that heterozygotes are almost always the result of recent outcrossing.

2.3. Approximate model. Conditional on $\{N_{ij}(t)\}_{t \in \{0, 1, \dots\}}$, the genealogy of the n sampled copies can clearly be described by a discrete-time Markov chain with finite state space S_n of size

$$(2.8) \quad |S_n| = \sum_{k=1}^n \sum_{i=0}^{[k/2]} \binom{k-2i+3}{3} \binom{i+2}{2}.$$

How to describe the genealogy without conditioning on the allele frequencies is considerably less clear, and I refer to Section 5.1 for further discussion of this.

The issue can be avoided by assuming that selection and/or mutation are/is strong enough relative to drift (*i.e.*, relative to $1/N$) and of the correct form (*i.e.*, allowing a stable point equilibrium) for the allele frequencies to be treated as constant. In this paper, I will assume that it is selection alone that maintains constant frequencies. I will furthermore be assuming that selection is *balancing*,

by which I simply mean that the equilibrium is polymorphic. More precisely then, the model I will investigate is the discrete-time Markov chain that results from assuming that $Y_i(t) = \hat{Y}_i > 0, \forall i, t$. Define its transition matrix $\Pi_N = (\pi_{ij})_{i,j \in S_n}$. The state space S_n will be described in Section 3.2.

The model may be characterized as a generalized structured n -coalescent with discrete generations [Notohara (1990), Herbots (1994), Nordborg (1997)]. Selection has “disappeared”, and only enters the model indirectly by determining the size of the “subpopulations” and the rates of “migration” between them. Although simple in principle, the model is difficult to work with because of the size and complexity of the state space. For example, for $n = 1, 2, 3, 4, 5$, and 6, we have $|S_n| = 4, 17, 49, 120, 260$, and 519. The remainder of this paper will be devoted to demonstrating that, through the use of a time-scales approximation, the model can be reduced to a much simpler, continuous-time structured n -coalescent with a state space of size

$$(2.9) \quad \sum_{m=1}^n (m+1) = \frac{1}{2}n(n+3),$$

which for the sample sizes just given equals 2, 5, 9, 14, 20 and 27. Furthermore, the new process will be shown to be identical to the one previously obtained for random mating [Kaplan et al. (1988), Hudson and Kaplan (1988)], except that the coalescent rate is increased by a constant factor, and the rate of exchange between the haplotypic states (through recombination and mutation), is decreased by another constant factor, where both factors have useful intuitive interpretations.

After obtaining these results (in Sections 3 and 4), I will argue (Section 5.1) that the results from the analysis will also hold to a good approximation for the original model, where the allele frequencies are random variables, if only selection is of the appropriate strength and kind.

3. Continuous time. The analysis will proceed in two steps. In this section, I will switch to a continuous time scale with time measured in units of N generations, as in the standard coalescent approximation [Kingman (1982)]. To do so, I will utilize a convergence theorem that was developed by Möhle (1998) for coalescent models with transitions on several time scales, in particular for the neutral coalescent with selfing [Nordborg and Donnelly (1997)]. The resulting process is considerably simpler, however, I will show how it can be simplified further in Section 4, where I describe the final model.

3.1. Scaling the parameters. The rationale behind the main results of this paper is that some transitions, namely those that involve two ancestors picking

the same individual in the previous generation, always have probability $O(1/N)$, whereas others, *e.g.*, those that simply involve a transition to a new genotypic configuration, always have probability $O(1)$. As we shall see, the states can be grouped into sets in a manner so that transitions within sets occur on time scale that is $O(N)$ faster than transitions between sets. Exactly how the state space is partitioned depends on what assumptions we place on the parameters of the model, and this, in turn, depends on the selective scenario we are interested in [Nordborg (1997)]. For the case of balancing selection, it is appropriate to assume that recombination and mutation are weak forces, and scale these parameters with N . We thus assume that the finite limits

$$(3.10) \quad R = \lim_{N \rightarrow \infty} Nr$$

and

$$(3.11) \quad U_{ij} = \lim_{N \rightarrow \infty} Nu_{ij}$$

exist. We choose not to scale the selection parameters in the same manner, *cf.* Sections 4.3 and 5.1.

With these assumptions, any transition that implies a recombination or mutation event, or two instances independently picking the same parental individual has probability $O(1/N)$, and transitions that necessitate more than one such event have probability $O(1/N^2)$ or smaller.

3.2. Partitioning the state space. We now turn to the state space S_n . It is convenient to partition and arrange the states as follows. First, group the states in n sets by the number of ancestors that remain of the original sample. Arrange the sets in increasing order by the number of ancestors. For example, the first four states will correspond to the four possible configurations for a single instance described in Section 2.2. Since the number of ancestors in the genealogy can only decrease, this arrangement ensures that Π_N is lower block-diagonal.

Second, consider each set with a given number of ancestors, m say, in turn. The m ancestors can be arranged into $m + 1$ sets with respect to haplotypic configuration. Arrange the sets in increasing order by the number of ancestors linked to \mathcal{A}_2 (the order here is arbitrary).

Third, consider the set with i ancestors linked to \mathcal{A}_1 and $j = m - i$ ancestors linked to \mathcal{A}_2 . This set can be divided into subsets by the number of individuals that harbor two ancestors. Write $\alpha_{i,j}$ for the set of states with all ancestors in distinct individuals, $\beta_{kl,i,j}$ for the set of states with a single $\mathcal{A}_k\mathcal{A}_l$ individual that harbors two ancestors, and $\gamma_{i,j}$ for the remaining states (that have two

or more individuals harboring two ancestors). We have $|\alpha_{i,j}| = (i+1)(j+1)$, $|\beta_{12,i,j}| = ij$,

$$|\beta_{11,i,j}| = \begin{cases} (i-1)(j+1), & \text{if } i > 1, \\ 0, & \text{otherwise,} \end{cases}$$

$$|\beta_{22,i,j}| = \begin{cases} (i+1)(j-1), & \text{if } j > 1, \\ 0, & \text{otherwise,} \end{cases}$$

and a potentially very large number of states in $\gamma_{i,j}$. Arrange these five sets in order $\alpha_{i,j}$, $\beta_{22,i,j}$, $\beta_{12,i,j}$, $\beta_{11,i,j}$, and $\gamma_{i,j}$.

We now turn to the transition probabilities within and between these sets. All transitions within $\alpha_{i,j}$ have probability $O(1)$ because they do not necessitate recombination, mutation, or two ancestors choosing the same parental individual, but simply a transition to another genotypic configuration. A transition to any state outside this set, however, has probability $O(1/N)$ or smaller. The following sets can be reached with probability $O(1/N)$ (*i.e.*, there is at least one transition to the set with that probability):

- whenever $i > 1$, transitions to $\beta_{11,i,j}$ or $\alpha_{i-1,j}$ may occur because two ancestors linked to \mathcal{A}_1 pick a common parental individual;
- whenever $j > 1$, transitions to $\beta_{22,i,j}$ or $\alpha_{i,j-1}$ may occur because two ancestors linked to \mathcal{A}_2 pick a common parental individual;
- whenever $i > 0$ and $j > 0$, transitions to $\beta_{12,i,j}$ may occur because two ancestors, one linked to \mathcal{A}_1 , one linked to \mathcal{A}_2 , pick a common parental individual;
- whenever $i > 0$, transitions to $\alpha_{i-1,j+1}$ may occur through recombination or mutation;
- whenever $j > 0$, transitions to $\alpha_{i+1,j-1}$ may occur through recombination or mutation.

All other sets can only be reached through multiple independent events with probability $O(1/N)$, and therefore have probability $O(1/N^2)$ or smaller.

All transitions within $\beta_{11,i,j}$ have probability $O(1)$, and so do transitions “back to” $\alpha_{i,j}$ (these occur when the individual harboring two ancestors is the product of random mating), and to $\alpha_{i-1,j}$ (which occur with probability one half when the individual harboring two ancestors is the product of selfing). All other transitions have probability $O(1/N)$ or smaller. Exactly the same is true for $\beta_{22,i,j}$, except that a coalescence of course leads to $\alpha_{i,j-1}$ instead of $\alpha_{i-1,j}$.

All transitions within $\beta_{12,i,j}$ have probability $O(1)$, and so do transitions back to $\alpha_{i,j}$. All other transitions have probability $O(1/N)$ or smaller.

We let the transitions from $\gamma_{i,j}$ remain unspecified because they will be shown to be unimportant.

3.3. *Applying Möhle's theorem.* Rewrite the transition matrix $\Pi_N = \mathbf{A} + \mathbf{B}/N + O(1/N^2)$, where $\mathbf{A} = \lim_{N \rightarrow \infty} \Pi_N$ and $\mathbf{B} = \lim_{N \rightarrow \infty} N(\Pi_N - \mathbf{A})$. From Möhle's (1998) results, it follows that if $\mathbf{P} = \lim_{m \rightarrow \infty} \mathbf{A}^m$ exists, then the finite-dimensional distributions of the process converge to those of a continuous-time Markov process with time measured in units of N generations, and infinitesimal generator $\mathbf{G} = \mathbf{P}\mathbf{B}\mathbf{P}$.

3.3.1. *Finding \mathbf{A} and \mathbf{P} .* From Section 3.2, it follows that \mathbf{A} has the form

$$\mathbf{A} = \left(\begin{array}{c|cc|cc|cc|cc|cc|cc} \cdot & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \mathbf{A}_{i,j-1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \cdot & \cdot & \cdot & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \mathbf{A}_{i-1,j} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \mathbf{A}_{i+1,j-1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{A}_{i,j} & 0 & 0 & 0 & 0 & 0 \\ 0 & \mathbf{A}'_{22,\wedge} & 0 & 0 & 0 & 0 & 0 & \mathbf{A}'_{22,v} & \mathbf{A}'_{22} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{A}'_{12,v} & 0 & \mathbf{A}'_{12} & 0 & 0 & 0 \\ 0 & 0 & 0 & \mathbf{A}'_{11,\wedge} & 0 & 0 & 0 & \mathbf{A}'_{11,v} & 0 & 0 & \mathbf{A}'_{11} & 0 & 0 \\ \cdot & \mathbf{A}'_\gamma & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{A}_{i-1,j+1} & 0 \\ \cdot & \cdot \end{array} \right),$$

where $\mathbf{A}_{i,j}$ contains all transitions within $\alpha_{i,j}$, and the \mathbf{A}'_{\dots} denote the transitions involving states with one or more individuals harboring two ancestors, with the dependence on i and j suppressed to save space. Thus \mathbf{A}'_{kl} contains all transitions within $\beta_{kl,i,j}$, where i and j should be understood to be the ones of the first $\mathbf{A}_{i,j}$ encountered when following the diagonal towards the upper left corner, and analogously for \mathbf{A}'_γ . The meaning of the off-diagonal elements is evident from their position. The dots represent blocks that may contain non-zero elements. It is clear that \mathbf{A} is stochastic because it can be interpreted as Π_N conditional on no events with probability $O(1/N)$ or lower taking place.

It is easy to show through induction that all blocks of zeros in \mathbf{A} remain blocks of zeros in \mathbf{A}^m . Furthermore, the diagonal blocks of \mathbf{A}^m are simply the diagonal blocks of \mathbf{A} raised to the power m . Since $\mathbf{A}_{i,j}$ is a positive stochastic matrix containing the transition probabilities within $\alpha_{i,j}$ conditional on not leaving that set, $\mathbf{P}_{i,j} = \lim_{m \rightarrow \infty} \mathbf{A}_{i,j}^m$ is a matrix with all rows identical and equal to the stationary distribution within $\alpha_{i,j}$. The diagonal blocks for states involving one or more individuals harboring two ancestors are even easier: these blocks are all non-negative matrices with modulus less than one and thus vanish as $m \rightarrow \infty$.

\mathbf{A}^m also contains two types of sub-diagonal blocks, namely those that correspond to transitions from states with a single individual harboring two ancestors to states with all ancestors in different individuals, and those that correspond to transitions from states with more than one individual harboring two ancestors to states with a single such ancestor. Both types of blocks have the general form

$$(3.12) \quad \sum_{k=0}^{m-1} \mathbf{Q}_{22}^k \mathbf{Q}_{21} \mathbf{Q}_{11}^{m-1-k},$$

where \mathbf{Q}_{11} stands for the diagonal block containing transitions within the set with the lower number of individuals harboring two ancestors, \mathbf{Q}_{22} stands for the diagonal block containing transitions within the set with the higher number of such individuals, and \mathbf{Q}_{21} contains the transitions from higher to lower. It can be shown that

$$(3.13) \quad \lim_{m \rightarrow \infty} \sum_{k=0}^{m-1} \mathbf{Q}_{22}^k \mathbf{Q}_{21} \mathbf{Q}_{11}^{m-1-k} = (\mathbf{I} - \mathbf{Q}_{22})^{-1} \mathbf{Q}_{21} \mathbf{Q}_{11}^{\infty}.$$

Thus, when \mathbf{Q}_{11} stands for a diagonal block of transitions within a set with all ancestors in distinct individuals, \mathbf{Q}_{11}^{∞} is a matrix with identical rows containing the stationary distribution, as described in the previous paragraph, whereas when \mathbf{Q}_{11} stands for a diagonal block of transitions within a set with a single individual containing two ancestors, we have $\mathbf{Q}_{11}^{\infty} = \mathbf{0}$ so that the right hand side of equation (3.13) vanishes.

We thus have

$$\mathbf{P} = \left(\begin{array}{c|cc|cc|cc} \cdot & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \mathbf{0} \mathbf{P}_{i,j-1} \mathbf{0} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \cdot & \cdot & \cdot & 0 & 0 & 0 & 0 & 0 & 0 \\ \mathbf{0} & 0 & 0 \mathbf{P}_{i-1,j} \mathbf{0} & 0 & 0 & 0 & 0 & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & 0 & 0 & 0 & 0 & 0 \\ \mathbf{0} & 0 & 0 & 0 & 0 \mathbf{P}_{i+1,j-1} \mathbf{0} & 0 & 0 & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{P}_{i,j} & 0 \\ \mathbf{0} \mathbf{P}'_{22,\wedge} \mathbf{0} & 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{P}'_{22,\vee} & 0 \\ \mathbf{0} & 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{P}'_{12,\vee} & 0 \\ \mathbf{0} & 0 & 0 & \mathbf{P}'_{11,\wedge} & 0 & 0 & 0 & \mathbf{P}'_{11,\vee} & 0 \\ \cdot & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{P}_{i-1,j+1} \mathbf{0} \\ \cdot & \cdot \end{array} \right),$$

where the blocks are labeled as in \mathbf{A} , and the off-diagonal elements are obtained through equation (3.13).

3.3.2. *Finding \mathbf{B} and \mathbf{G} .* The matrix \mathbf{B} contains the coefficients of the terms $O(1/N)$ from a series expansion of Π_N in N^{-1} . It is neither stochastic nor non-negative. From Section 3.2, it follows that it has the structure

$$\mathbf{B} = \left(\begin{array}{cccc|cccc|cccc|cc} \cdot & \cdot & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \cdot & \mathbf{B}_{i,j-1} & \cdot & \cdot & \cdot & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \cdot & \cdot & \cdot & \mathbf{B}_{i-1,j} & \cdot & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \mathbf{B}_{i+1,j-1} & \cdot & 0 & 0 & 0 & 0 & 0 & 0 \\ \cdot & 0 & 0 \\ \hline 0 & \mathbf{B}_{0,-1} & 0 & \mathbf{B}_{-1,0} & 0 & \mathbf{B}_{+1,-1} & 0 & \mathbf{B}_{i,j} & \mathbf{B}_{22} & \mathbf{B}_{12} & \mathbf{B}_{11} & 0 & \mathbf{B}_{-1,+1} & 0 \\ \cdot & 0 \\ \cdot & 0 \\ \cdot & 0 \\ \cdot & 0 \\ \hline \cdot & 0 & 0 & 0 & 0 & \mathbf{B}_{i-1,j+1} & \cdot \\ \cdot & \cdot \end{array} \right),$$

where the diagonal blocks are denoted as before, and the meaning of the off-diagonal blocks can be inferred from their position in the matrix. Again, the dependence of these blocks on i and j has been suppressed to save space.

Turning to $\mathbf{G} = \mathbf{P}\mathbf{B}\mathbf{P}$, we first note that the diagonal blocks for all sets that contain one or more individuals harboring two ancestors are zero, as are the blocks containing transitions into such sets from sets with all ancestors in distinct individuals. Since \mathbf{G} is the infinitesimal generator for the continuous-time version of the Markov process describing the genealogy of the sample, this means that, as one would intuitively expect, all such states are instantaneous on a time-scale measured in units of $O(N)$. These states can therefore be eliminated, leaving us with a process involving only the sets $\{\alpha_{i,j}\}_{i,j \in \{1, \dots, m\}, i+j=m, m \in \{1, \dots, n\}}$. The blocks of \mathbf{G} that correspond to transitions within and between these sets are as follows.

In general, the rows of \mathbf{G} corresponding to transitions from $\alpha_{i,j}$ has five non-zero blocks. Two of these correspond to a single recombination or mutation event, namely:

$$(3.14) \quad \mathbf{G}_{\alpha_{i,j}, \alpha_{i+1,j-1}} = \mathbf{P}_{i,j} \mathbf{B}_{+1,-1} \mathbf{P}_{i+1,j-1};$$

$$(3.15) \quad \mathbf{G}_{\alpha_{i,j}, \alpha_{i-1,j+1}} = \mathbf{P}_{i,j} \mathbf{B}_{-1,+1} \mathbf{P}_{i-1,j+1},$$

two correspond to single coalescence events, namely:

$$(3.16) \quad \mathbf{G}_{\alpha_{i,j}, \alpha_{i,j-1}} = \mathbf{P}_{i,j} (\mathbf{B}_{0,-1} \mathbf{P}_{i,j-1} + \mathbf{B}_{22} \mathbf{P}'_{22,\wedge});$$

$$(3.17) \quad \mathbf{G}_{\alpha_{i,j}, \alpha_{i-1,j}} = \mathbf{P}_{i,j} (\mathbf{B}_{-1,0} \mathbf{P}_{i-1,j} + \mathbf{B}_{11} \mathbf{P}'_{11,\wedge}),$$

and the final one is the diagonal block

$$(3.18) \quad \mathbf{G}_{\alpha_{i,j}, \alpha_{i,j}} = \mathbf{P}_{i,j} (\mathbf{B}_{i,j} \mathbf{P}_{i,j} + \mathbf{B}_{22} \mathbf{P}'_{22,v} + \mathbf{B}_{12} \mathbf{P}'_{12,v} + \mathbf{B}_{11} \mathbf{P}'_{11,v}).$$

4. The collapsed process. We have demonstrated that all states with one or more individuals harboring two ancestors are instantaneous and can be eliminated. The same result is obtained for the neutral coalescent with selfing [Nordborg and Donnelly (1997), Möhle (1998)], and it greatly simplifies the process. However, it turns out that further simplifications can be made to the present model, because all transitions within the remaining sets $\{\alpha_{i,j}\}_{i,j \in \{1, \dots, m\}, i+j=m}$ ($m \in \{1, \dots, n\}$) have probability $O(1)$, whereas transitions between these sets have probability $O(1/N)$ or smaller. Loosely speaking, the individual states within these sets are instantaneous, but the sets themselves are not, and we would expect the processes governing transitions within the sets to be at stationarity on the time scale on which transitions between the sets occur.

It is evident from \mathbf{G} that this intuition is correct. Notice that each of the blocks given by equations (3.14)–(3.18) is multiplied from the left by $\mathbf{P}_{i,j}$, which, as we have seen, consists of identical rows, each equal to the stationary distribution for $\mathbf{A}_{i,j}$. Thus all blocks of \mathbf{G} also have all rows equal, so that each state within $\alpha_{i,j}$ behaves identically, and the starting condition with respect to these states is irrelevant.

We can therefore simplify the process further by collapsing each $\alpha_{i,j}$ into a single state. This is done by summing over the rows and columns in the appropriate manner, so that in the generator for the collapsed process each block $\mathbf{G}_{\alpha_{i,j}, \alpha_{k,l}}$ is replaced by a corresponding element $g_{\alpha_{i,j}, \alpha_{k,l}}$. Let $\mathbf{p}_{i,j}$ be the stationary distribution for $\mathbf{A}_{i,j}$. The elements corresponding to the blocks (3.14)–(3.18) are

$$(4.19) \quad g_{\alpha_{i,j}, \alpha_{i+1,j-1}} = \mathbf{p}_{i,j} \mathbf{B}_{+1,-1} \mathbf{P}_{i+1,j-1} \mathbf{1}^T,$$

$$(4.20) \quad g_{\alpha_{i,j}, \alpha_{i-1,j+1}} = \mathbf{p}_{i,j} \mathbf{B}_{-1,+1} \mathbf{P}_{i-1,j+1} \mathbf{1}^T,$$

$$(4.21) \quad g_{\alpha_{i,j}, \alpha_{i,j-1}} = \mathbf{p}_{i,j} (\mathbf{B}_{0,-1} \mathbf{P}_{i,j-1} + \mathbf{B}_{22} \mathbf{P}'_{22,\wedge}) \mathbf{1}^T,$$

$$(4.22) \quad g_{\alpha_{i,j}, \alpha_{i-1,j}} = \mathbf{p}_{i,j} (\mathbf{B}_{-1,0} \mathbf{P}_{i-1,j} + \mathbf{B}_{11} \mathbf{P}'_{11,\wedge}) \mathbf{1}^T,$$

$$(4.23) \quad g_{\alpha_{i,j}, \alpha_{i,j}} = \mathbf{p}_{i,j} (\mathbf{B}_{i,j} \mathbf{P}_{i,j} + \mathbf{B}_{22} \mathbf{P}'_{22,v} + \mathbf{B}_{12} \mathbf{P}'_{12,v} + \mathbf{B}_{11} \mathbf{P}'_{11,v}) \mathbf{1}^T,$$

where $\mathbf{1}$ is a unit vector of appropriate length.

We thus have a new, “collapsed” continuous-time Markov process with states consisting of the number of ancestors and their haplotypic configuration. The size of the state space is that given by expression (2.9). The new process is equivalent to the one previously obtained for random mating [Kaplan et al. (1988), Hudson and Kaplan (1988)], except for the precise values of the non-zero transitions rates [and the possible initial coalescence events for chromosomes sampled from the same individual [Nordborg and Donnelly (1997)]]. We now turn to these rates, noting that because the row sums of the generator matrix must be zero, and because of symmetry, it suffices to find $g_{\alpha_{i,j}, \alpha_{i+1,j-1}}$ and $g_{\alpha_{i,j}, \alpha_{i-1,j}}$, *i.e.*, the rate of recombination/mutation and the rate of coalescence, respectively.

4.1. The rate of recombination/mutation. Equation (4.19) turns out to have a simple interpretation. We have

$$g_{\alpha_{i,j}, \alpha_{i+1,j-1}} = \mathbf{p}_{i,j} \mathbf{B}_{+1,-1} \mathbf{P}_{i+1,j-1} \mathbf{1}^T = \mathbf{p}_{i,j} \mathbf{B}_{+1,-1} \mathbf{1}^T.$$

Each element in the column vector $\mathbf{B}_{+1,-1} \mathbf{1}^T$ is the sum, over all states in $\alpha_{i+1,j-1}$, of the transition rates from a particular state in $\alpha_{i,j}$ to the states in $\alpha_{i+1,j-1}$. To put it another way, it is the transition rate from $\alpha_{i,j}$ to $\alpha_{i+1,j-1}$ conditional on the process currently being in a particular state in $\alpha_{i,j}$. Multiplication with the stationary distribution over $\alpha_{i,j}$ gives the total transition rate from $\alpha_{i,j}$ to $\alpha_{i+1,j-1}$. Thus, transitions between these sets due to recombination or mutation occur according to the stationary distribution within them, as predicted.

The resulting rate can be found as follows. Consider the set $\alpha_{i,j}$. The states within this set can be described by (k, l) , where $k \in \{0, \dots, i\}$ ($l \in \{0, \dots, j\}$) is the number of ancestors linked to \mathcal{A}_1 (\mathcal{A}_2) in heterozygotes. The transition probabilities between these states are those found in $\mathbf{A}_{i,j}$. As explained in Section 2.2, ancestors pick their state in the previous generation independently of one another. In particular, the probability of the transition (k_0, l_0) to (k, l) can be written $\psi_{k_0, k} \omega_{l_0, l}$. Furthermore, we see from equation (2.1) that the probability that a given ancestor, currently linked to \mathcal{A}_1 was linked to \mathcal{A}_2 in the previous generation but switched because of mutation is

$$\frac{\hat{Y}_2 u_{21}}{\hat{Y}_2 u_{21} + \hat{Y}_1(1 - u_{12})} = \frac{\hat{Y}_2 U_{21}}{\hat{Y}_1 N} + O\left(\frac{1}{N^2}\right).$$

Similarly, the probability that a given ancestor switched because of recombination is

$$\frac{\hat{Y}_2 r}{\hat{Y}_2 r + \hat{Y}_1(1 - r)} = \frac{\hat{Y}_2 R}{\hat{Y}_1 N} + O\left(\frac{1}{N^2}\right),$$

conditional on the ancestor picking a heterozygous parent (the probability is zero otherwise). The total probability of a transition from $(k_0, l_0) \in \alpha_{i,j}$ to *some* state in $\alpha_{i+1,j-1}$ is therefore

$$\sum_{k=0}^i \sum_{l=0}^j \frac{\hat{Y}_2}{\hat{Y}_1} \left(\frac{U_{21}}{N} j + \frac{R}{N} l \right) \psi_{k_0, k} \omega_{l_0, l} + O\left(\frac{1}{N^2}\right),$$

and the element of $\mathbf{B}_{+1,-1} \mathbf{1}^T$ that corresponds to a transition from (k_0, l_0) is

$$\begin{aligned} \sum_{k=0}^i \sum_{l=0}^j \frac{\hat{Y}_2}{\hat{Y}_1} (U_{21}j + Rl) \psi_{k_0, k} \omega_{l_0, l} &= \frac{\hat{Y}_2}{\hat{Y}_1} \sum_{l=0}^j (U_{21}j + Rl) \omega_{l_0, l} \\ &= \frac{\hat{Y}_2}{\hat{Y}_1} \left(U_{21}j + R \sum_{l=0}^j l \omega_{l_0, l} \right) \\ &= \frac{\hat{Y}_2}{\hat{Y}_1} [U_{21}j + R \mathbb{E}(l|l_0)], \end{aligned}$$

where $\mathbb{E}(l|l_0)$ is the expectation, over the transitions in $\mathbf{A}_{i,j}$, of the number of ancestors linked to \mathcal{A}_2 occupying heterozygotes in the previous generation, given that the number is l_0 in the present generation. By multiplying $\mathbf{B}_{+1,-1} \mathbf{1}^T$ with the stationary distribution $\mathbf{p}_{i,j}$, we are in effect calculating the unconditional expectation

$$\mathbb{E}\left(\frac{\hat{Y}_2}{\hat{Y}_1} [U_{21}j + R \mathbb{E}(l|l_0)]\right) = \frac{\hat{Y}_2}{\hat{Y}_1} [U_{21}j + R \mathbb{E}(l)],$$

where l is binomially distributed with parameters j and H_2 , the probability that, in the absence of recombination and mutation, an ancestor linked to an \mathcal{A}_2 in the present generation occupied a heterozygote in the previous generation (this probability can be calculated exactly from the equations in Section 2.1, but does not have a simple form). Using this, the sought rate becomes

$$(4.24) \quad g_{\alpha_{i,j}, \alpha_{i+1,j-1}} = \left(\frac{\hat{Y}_2}{\hat{Y}_1} U_{21} + H_2 R \right) j.$$

By symmetry,

$$(4.25) \quad g_{\alpha_{i,j}, \alpha_{i-1,j+1}} = \left(\frac{\hat{Y}_1}{\hat{Y}_2} U_{12} + H_1 R \right) i,$$

where H_1 is defined analogously to H_2 .

4.2. *The rate of coalescence.* From (4.22), and utilizing (3.13), we have

$$\begin{aligned} g_{\alpha_{i,j}, \alpha_{i-1,j}} &= \mathbf{p}_{i,j}(\mathbf{B}_{-1,0}\mathbf{P}_{i-1,j} + \mathbf{B}_{11}\mathbf{P}'_{11,\wedge})\mathbf{1}^T, \\ &= \mathbf{p}_{i,j}(\mathbf{B}_{-1,0}\mathbf{P}_{i-1,j} + \mathbf{B}_{11}(\mathbf{I} - \mathbf{A}'_{11})^{-1}\mathbf{A}'_{11,\wedge}\mathbf{P}_{i,j})\mathbf{1}^T \\ &= \mathbf{p}_{i,j}\mathbf{B}_{-1,0}\mathbf{1}^T + \mathbf{p}_{i,j}\mathbf{B}_{11}(\mathbf{I} - \mathbf{A}'_{11})^{-1}\mathbf{A}'_{11,\wedge}\mathbf{1}^T. \end{aligned}$$

Consider first $\mathbf{p}_{i,j}\mathbf{B}_{-1,0}\mathbf{1}^T$. Using the arguments and notation of the previous section, and referring to equation (2.7) and the discussion following it, we note that the total probability of a single-generation transition from $(k_0, l_0) \in \alpha_{i,j}$ to *some* state in $\alpha_{i-1,j}$ can be shown to be

$$(4.26) \quad \sum_{k=0}^i \sum_{l=0}^j \left(\frac{\binom{k}{2}}{N\hat{X}_{12}} + \frac{\binom{i-k}{2}}{N\hat{X}_{11}} \frac{1}{2} \right) \psi_{k_0, k} \omega_{l_0, l} + O\left(\frac{1}{N^2}\right).$$

The corresponding element of $\mathbf{B}_{-1,0}\mathbf{1}^T$ is thus

$$\begin{aligned} \sum_{k=0}^i \sum_{l=0}^j \left(\frac{\binom{k}{2}}{\hat{X}_{12}} + \frac{\binom{i-k}{2}}{\hat{X}_{11}} \frac{1}{2} \right) \psi_{k_0, k} \omega_{l_0, l} &= \sum_{k=0}^i \left(\frac{\binom{k}{2}}{\hat{X}_{12}} + \frac{\binom{i-k}{2}}{2\hat{X}_{11}} \right) \psi_{k_0, k} \\ &= \mathbb{E} \left(\frac{\binom{k}{2}}{\hat{X}_{12}} + \frac{\binom{i-k}{2}}{2\hat{X}_{11}} \mid k_0 \right). \end{aligned}$$

By the arguments of the previous section, multiplication from the left by the stationary distribution $\mathbf{p}_{i,j}$ is equivalent to calculating the unconditional expectation

$$\mathbb{E} \left(\frac{\binom{k}{2}}{\hat{X}_{12}} + \frac{\binom{i-k}{2}}{2\hat{X}_{11}} \right) = \frac{1}{\hat{X}_{12}} \mathbb{E} \left[\binom{k}{2} \right] + \frac{1}{2\hat{X}_{11}} \mathbb{E} \left[\binom{i-k}{2} \right],$$

where k is binomially distributed with parameters i and H_1 . The sought rate is thus

$$(4.27) \quad \mathbf{p}_{i,j}\mathbf{B}_{-1,0}\mathbf{1}^T = \binom{i}{2} \left(\frac{1}{\hat{X}_{12}} H_1^2 + \frac{1}{2\hat{X}_{11}} (1 - H_1)^2 \right).$$

It remains to deal with $\mathbf{p}_{i,j} \mathbf{B}_{11} (\mathbf{I} - \mathbf{A}'_{11})^{-1} \mathbf{A}'_{11,\wedge} \mathbf{1}^T$. Consider the set $\beta_{11,i,j}$. This set contains all states with i ancestors linked to \mathcal{A}_1 and j ancestors linked to \mathcal{A}_2 such that two of the ancestors linked to \mathcal{A}_1 jointly occupy a homozygote. The remaining ancestors belong to a set of the form $\alpha_{i-2,j}$. For each transition on this set, and independently of that transition, one of three things may happen to the $\mathcal{A}_1\mathcal{A}_1$ homozygote harboring two ancestors:

1. it may have resulted from outcrossing;
2. it may have resulted from the selfing of a heterozygote;
3. it may have resulted from the selfing of a homozygote.

Ignoring probabilities of $O(1/N)$ and smaller, the following then applies for these alternatives, respectively:

1. we have a transition to a state in $\alpha_{i,j}$;
2. we have a transition to $\alpha_{i-1,j}$ because the ancestors coalesce;
3. with probability one half, we have a transition to $\alpha_{i-1,j}$ because the ancestors coalesce, and with probability one half the process remains in $\beta_{11,i,j}$.

Now consider $\mathbf{A}'_{11,\wedge} \mathbf{1}^T$. Each element of this vector is the total probability that a transition from $k_0 \in \beta_{11,i,j}$ is a transition to $\alpha_{i-1,j}$. As we have just seen, this is simply the probability that the $\mathcal{A}_1\mathcal{A}_1$ homozygote was the product of a selfed heterozygote plus one half times the probability that it was the product of a selfed homozygote. Denote the former quantity q_{12} and the latter q_{11} . We thus have

$$\mathbf{A}'_{11,\wedge} \mathbf{1}^T = \mathbf{1}^T \left(q_{12} + \frac{1}{2} q_{11} \right).$$

Next consider the matrix \mathbf{A}'_{11} . By the above, we must have

$$\mathbf{A}'_{11} = \frac{1}{2} q_{11} \mathbf{A}_{i-2,j},$$

and using this it is easy to show that

$$(\mathbf{I} - \mathbf{A}'_{11})^{-1} \mathbf{1}^T = \mathbf{1}^T \frac{1}{1 - \frac{1}{2} q_{11}},$$

so that

$$(\mathbf{I} - \mathbf{A}'_{11})^{-1} \mathbf{A}'_{11,\wedge} \mathbf{1}^T = \mathbf{1}^T \frac{2q_{12} + q_{11}}{2 - q_{11}}.$$

It follows from equation (2.3) that

$$\begin{aligned} q_{11} &= s \frac{\hat{X}_{11}}{\hat{x}_{11}}, \\ q_{12} &= s \frac{\hat{X}_{12}}{4\hat{x}_{11}}. \end{aligned}$$

Using this, and repeating the arguments leading to equation (4.27), we finally obtain

$$\mathbf{p}_{i,j} \mathbf{B}_{11} (\mathbf{I} - \mathbf{A}'_{11})^{-1} \mathbf{A}'_{11,\wedge} \mathbf{1}^T = \binom{i}{2} \frac{(1 - H_1)^2}{2\hat{X}_{11}} \frac{s\hat{Y}_1}{2\hat{x}_{11} - s\hat{X}_{11}}.$$

Combining this with equation (4.27) leads to

$$(4.28) \quad g_{\alpha_{i,j}, \alpha_{i-1,j}} = \binom{i}{2} \left(\frac{H_1^2}{\hat{X}_{12}} + \frac{(1 - H_1)^2}{2\hat{X}_{11}} + \frac{(1 - H_1)^2}{2\hat{X}_{11}} \frac{s\hat{Y}_1}{2\hat{x}_{11} - s\hat{X}_{11}} \right).$$

By symmetry,

$$(4.29) \quad g_{\alpha_{i,j}, \alpha_{i,j-1}} = \binom{j}{2} \left(\frac{H_2^2}{\hat{X}_{12}} + \frac{(1 - H_2)^2}{2\hat{X}_{22}} + \frac{(1 - H_2)^2}{2\hat{X}_{22}} \frac{s\hat{Y}_2}{2\hat{x}_{22} - s\hat{X}_{22}} \right).$$

4.3. Weak selection. The purpose of this section is to show that the rates calculated in Sections 4.1 and 4.2 have very simple and intuitive forms if we assume that selection is weak enough for terms of the order of the selection coefficients to be ignored. Note that this is a statement about the absolute magnitude of the selection coefficients: they are still assumed to be large relative to $1/N$.

Under this approximation, then, we have from equation (2.6) that $\hat{x}_{ij} = \hat{X}_{ij}$. Furthermore, it is a classical result that

$$\begin{aligned} \hat{X}_{11} &= \hat{Y}_1^2 + \hat{Y}_1 \hat{Y}_2 F, \\ \hat{X}_{12} &= 2\hat{Y}_1 \hat{Y}_2 (1 - F), \\ \hat{X}_{22} &= \hat{Y}_2^2 + \hat{Y}_1 \hat{Y}_2 F, \end{aligned}$$

where $F = s/(2 - s)$. Using this, it can be shown that

$$\begin{aligned} H_1 &= \hat{Y}_2 (1 - F), \\ H_2 &= \hat{Y}_1 (1 - F), \end{aligned}$$

so that the recombination/mutation rates (4.24)–(4.25) become

$$(4.30) \quad g_{\alpha_i,j,\alpha_{i+1},j-1} = \left(\frac{\hat{Y}_2}{\hat{Y}_1} U_{21} + \hat{Y}_2(1-F)R \right) j,$$

$$(4.31) \quad g_{\alpha_i,j,\alpha_{i-1},j+1} = \left(\frac{\hat{Y}_1}{\hat{Y}_2} U_{12} + \hat{Y}_1(1-F)R \right) i,$$

and the coalescent rates (4.28)–(4.29) become

$$(4.32) \quad g_{\alpha_i,j,\alpha_{i-1},j} = \binom{i}{2} \frac{1}{2\hat{Y}_1} (1+F),$$

$$(4.33) \quad g_{\alpha_i,j,\alpha_{i+1},j-1} = \binom{j}{2} \frac{1}{2\hat{Y}_2} (1+F).$$

Thus, under this additional approximation, the coalescent with balancing selection and partial selfing looks almost identical to the coalescent with balancing selection and random mating [Kaplan et al. (1988); Hudson and Kaplan (1988)], the only difference being that the rate of coalescence within each allelic class is sped up by a factor $1+F$, and the rate of exchange between allelic classes due to recombination is decreased by a factor $1-F$. The former factor can be interpreted as the decrease in variance effective population size, and the latter as the decrease in heterozygosity.

5. Discussion. I have demonstrated that essentially all the extra complexity caused by allowing partial selfing in a coalescent model with balancing selection can be removed through a time-scales approximation. The results are interesting from a theoretical as well as from a biological point of view.

5.1. Theoretical issues. The results of this paper demonstrate the utility of combining the structured coalescent with time-scales approximations to model complex situations. Further examples are given, albeit with much less detail, in Nordborg (1997).

A few issues related to selection in the coalescent should be commented on. As described in Section 2.3, I have assumed that selection is strong enough for the allele frequencies to be treated as deterministic, so that the coalescent with selection becomes a structured coalescent. Although I have not supplied a formal proof of convergence, this approach seems justified by the fact that the calculations can be carried out with constant selection coefficients even as we let $N \rightarrow \infty$. In their original analysis of this problem, Kaplan et al. (1988) derived the coalescent conditional on the allele frequencies in all generations,

and assumed that these obeyed a limiting diffusion. However, most results were then obtained assuming that the allele frequencies were “tightly controlled”, which is similar to the assumption used in this paper.

Recently, theory has been developed for sample genealogies with “true” selection [Neuhauser and Krone (1997), Krone and Neuhauser (1997)], *i. e.*, without conditioning on the allele frequencies in all generations. It would be very interesting to investigate the limiting behavior of such models as selection becomes stronger. It seems clear that they will converge to structured coalescent models, however, knowing more about the conditions under which they converge could be quite important, because the exact models are considerably more difficult to analyze than the type of model analyzed here. Furthermore, although the exact models yield convenient computational algorithms for simulating samples with selection, the computational time depends exponentially on the strength of selection, whereas simulations using the structured-coalescent approach of course are independent of the strength of selection.

It is thus the view of the present author that the results in this paper, and similar results, should be seen as a “strong-selection limit” for the coalescent with selection. In this context, it is illuminating to consider the simplifications in Section 4.3. Compare equations (4.28) and (4.32). The main reason for the difference in complexity between these two expressions is that when selection is sufficiently weak, the stationary probability that an ancestor occupies a certain state is proportional to the “population size” of that state, and thus inversely proportional to the coalescent rate. As shown in Nordborg (1997), this condition is a generalized version of Nagylaki’s (1980) “conservative migration” criterion (by which migration is conservative if it does not effect subpopulation sizes). Nagylaki showed that, in the strong-migration limit, a subdivided population behaves as an unstructured one if and only if migration is conservative, whereas if migration is not conservative, the population will behave as an unstructured population with a lowered variance effective population size. In the present case, we have “strong migration” between the genotypic classes within haplotypic classes, and this is clearly not conservative in general. Under some forms of balancing selection, for instance, we expect heterozygotes to be fitter than homozygotes. If we think of the genotypes as demes, then, forward in time, heterozygotes are net sources of “migrant” gametes, whereas backwards in time, “migrating” ancestors will spend a disproportional amount of time in heterozygotes. This leads to a decreased variance effective population size, and, unfortunately, to rather messy expressions.

Taking the analogy with population structure and demography further, perhaps some forms of selection will result in a non-linear change of time-scale, just like some forms of variation in population size does? Clearly much work

remains to be done in this area of population genetics theory.

5.2. Biological issues. The main biological implication of the work presented here is simply stated: the dynamics of linkage disequilibrium and other forms of allelic associations in partially selfing organisms are governed (to a reasonable approximation) by $Nr(1 - F)$ rather than Nr . This is perhaps not surprising, but may be under-appreciated. Perhaps the most exciting consequence of this result is that the traces of some form of balancing selection may be detectable at a much greater distance from the actual site of selection than in a comparable outcrossing species. This suggests that studies aiming to detect selection in this way should consider using partially selfing organisms: indeed, it may even be possible to scan the genome directly for regions that show traces of balancing selection [Nordborg et al. (1996), Nordborg (1997)].

Acknowledgments. I thank T. Nagylaki and H. Nordborg for the many discussions and explanations that were essential for this paper to come into being, M. Möhle for comments on the manuscript, the American Mathematical Society for funding this conference and subsidizing my attending it, and F. Seillier-Moiseiwitsch for organizing the conference as well as for her patience.

REFERENCES

- CHARLESWORTH, B., NORDBORG, M. and CHARLESWORTH, D. (1997). The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. *Genet. Res., Camb.* **70** 155–174.
- HERBOTS, H. M. (1994). "Stochastic Models in Population Genetics: Genealogy and Genetic Differentiation in Structured Populations". PhD thesis, University of London.
- HEY, J. (1991). A multi-dimensional coalescent process applied to multi-allelic selection models and migration models. *Theor. Pop. Biol.* **39** 30–48.
- HUDSON, R. R. (1996). Molecular population genetics of adaptation. In *Adaptation*, M. R. Rose and G. V. Lauder, eds, pp. 291–309, Academic Press, San Diego.
- HUDSON, R. R. and KAPLAN, N. L. (1988). The coalescent process in models with selection and recombination. *Genetics* **120** 831–840.
- KAPLAN, N. L., DARSEN, T. and HUDSON, R. R. (1988). The coalescent process in models with selection. *Genetics* **120** 819–829.
- KINGMAN, J. F. C. (1982). The coalescent. *Stochast. Proc. Appl.* **13** 235–248.
- KREITMAN, M. and AKASHI, H. (1995). Molecular evidence for natural selection. *Annu. Rev. Ecol. Syst.* **26** 403–422.
- KRONE, S. M. and NEUHAUSER, C. (1997). Ancestral processes with selection. *Theor. Pop. Biol.* **51** 210–237.
- MÖHLE, M. (1998). A convergence theorem for Markov chains arising in population genetics and the coalescent with selfing. *Adv. Appl. Prob.*, to appear (available via <http://www.mathematik.uni-mainz.de/Stochastik/Arbeitsgruppe/moehle.html>).
- NAGYLAKI, T. (1980). The strong-migration limit in geographically structured populations. *J. Math. Biol.* **9** 101–114.
- NEUHAUSER, C. and KRONE, S. M. (1997). The genealogy of samples in models with selection. *Genetics* **145** 519–534.

- NORDBORG, M. (1997). Structured coalescent processes on different time scales. *Genetics* **146** 1501–1514.
- NORDBORG, M., CHARLESWORTH, B. and CHARLESWORTH, D. (1996). Increased levels of polymorphism surrounding selectively maintained sites in highly selfing species. *Proc. R. Soc. Lond. B* **263** 1033–1039.
- NORDBORG, M. and DONNELLY, P. (1997). The coalescent process with selfing. *Genetics* **146** 1185–1195.
- NOTOHARA, M. (1990). The coalescent and the genealogical process in geographically structured populations. *J. Math. Biol.* **29** 59–75.

DEPARTMENT OF GENETICS
LUND UNIVERSITY
SÖLVEGATAN 29
223 62 LUND, SWEDEN
MAGNUS.NORDBORG@GEN.LU.SE

ESTIMATION OF CONDITIONAL MULTILocus GENE IDENTITY AMONG RELATIVES

BY ELIZABETH A. THOMPSON AND SIMON C. HEATH

University of Washington

Genetic Analysis Workshop 10 identified five key factors contributing to the resolution of the genetic factors affecting complex traits. These include analysis with multipoint methods, use of extended pedigrees, and selective sampling of pedigrees. By sampling the affected individuals in an extended pedigree, we obtain individuals who have an increased probability of sharing genes identical by descent (IBD) at marker loci that are linked to the trait locus or loci. Given marker data on specified members of a pedigree, the conditional IBD status among relatives can be assessed, but exact computation is often impractical for multiple linked markers on complex pedigrees. The use of Markov chain Monte Carlo (MCMC) methods greatly extends the range of models and data sets for which analysis is computationally feasible. Many forms of MCMC have now been implemented in the context of genetic analysis. Here we propose a new sampler, which takes as latent variables the segregation indicators at marker loci, and jointly updates all indicators corresponding to a given meiosis. The sampler has good mixing properties. Questions of irreducibility are also addressed.

1. Introduction. Relatives share common ancestors. A single gene in such an ancestor may therefore descend via repeated segregations to each of the relatives. Such genes, which are copies of a single ancestral gene within a defined pedigree, are said to be *identical by descent* (IBD). Disregarding mutation, IBD genes must be of like type. It is the sharing of IBD genes that underlies phenotypic similarities among relatives. The probabilities of patterns of gene identity by descent are determined by the pedigree structure, and in turn determine the probability distribution of observed data on individuals of the pedigree.

Genetic linkage is the dependent cosegregation of genes at different loci on the same chromosome. Linkage detection and linkage analysis on the basis of data observed on related individuals require the computation of multilocus probabilities of observed phenotypic data on pedigree structures. Genetic Analysis Workshop 10 identified five key factors contributing to the resolution of the genetic factors affecting complex traits (Wijsman and Amos 1997). These include analysis with multipoint methods, use of extended pedigrees, and selective sampling of pedigrees. Here we consider an approach to linkage detection which uses only data on affected individuals. However, calculation of multilocus probabili-

Work supported in part by NIH grant GM-46255 and NSF grant BIR-9305835.

AMS 1991 subject classifications. Primary 62F03 ; secondary 92D10.

Key words and phrases. Markov chain Monte Carlo, meiosis samplers, linkage detection, irreducibility, segregation indicators.

ties on extended pedigrees is computationally intensive, particularly when there are many unobserved individuals. In this paper we present a sampling-based approach for linkage detection which is well suited to sparse data at multiple loci on individuals in an extended complex pedigree.

2. Gene identity and linkage likelihoods. There are many ways to partition linkage likelihoods, the probability $\Pr_\psi(\mathbf{Y})$ of phenotypic data \mathbf{Y} under a genetic linkage model ψ . Let \mathbf{Y} consist of trait data \mathbf{Y}_T and marker data \mathbf{Y}_M . The model (genetic map positions and marker alleles frequencies) is assumed known for the data, \mathbf{Y}_M , at marker loci. In this paper, we shall focus on the problem of linkage detection, in which no trait specific genetic model for the trait data \mathbf{Y}_T is assumed. However, the development is similar in the case where hypothesized trait loci are explicitly modeled and linkage estimation is the goal of the analysis (Thompson 1994b).

Let B_M denote the pattern of gene IBD at marker loci among observed individuals. The likelihood for the genetic model on the basis of data $\mathbf{Y} = (\mathbf{Y}_T, \mathbf{Y}_M)$ is

$$(2.1) \quad \begin{aligned} \Pr_\psi(\mathbf{Y}) &= \Pr_\psi(\mathbf{Y}_T, \mathbf{Y}_M) \propto \Pr_\psi(\mathbf{Y}_T \mid \mathbf{Y}_M) \\ &= \sum_{B_M} \Pr_\psi(\mathbf{Y}_T \mid B_M) \Pr(B_M \mid \mathbf{Y}_M) \end{aligned}$$

where the model ψ relates to the trait parameters and loci positions relative to the known marker map. If desired, we may consider also IBD status B_T at putative trait loci, and partition the probability further:

$$\Pr_\psi(\mathbf{Y}_T \mid B_M) = \sum_{B_T} \Pr_\psi(\mathbf{Y}_T \mid B_T) \Pr_\psi(B_T \mid B_M).$$

Even where no explicit trait model is assumed, there is an implicit assumption in linkage analysis that a trait is genetically determined. Thus individuals of like phenotype have higher probabilities of sharing genes IBD at trait loci, and hence also at linked marker loci. Thus evidence for linkage is provided by marker data \mathbf{Y}_M that give high posterior probability $\Pr(B_M \mid \mathbf{Y}_M)$ to patterns of gene identity B_M which specify greater than expected gene sharing among affected individuals.

A simple example may clarify this perspective. In homozygosity mapping (Lander and Botstein 1987), data on unrelated inbred affected individuals are used to map rare recessive traits. Since the individuals are unrelated, we may consider separately the IBD pattern for each. An example pedigree is shown in Figure 1; this pedigree resulted from a study of a rare recessive disease (Goddard

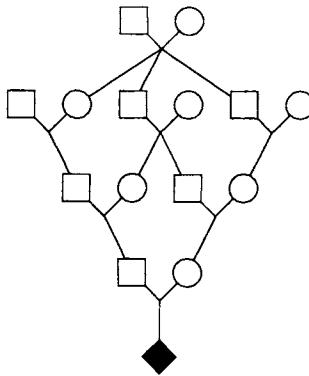


FIG. 1. Example pedigree, showing the original trait data of a single inbred affected individual.

et al. 1997). The final individual was ascertained as being affected and the offspring of a marriage between first cousins; it was later recognized that each of his parents was also the child of a first-cousin marriage, as shown. The inbreeding coefficient of the affected individual is $f = 0.109375$.

The IBD patterns of interest are whether the two genes of the inbred affected individual are IBD ($B_T = 1$) or not ($B_T = 0$). Since the trait is a rare recessive

$$\Pr_{\psi}(\mathbf{Y}_T \mid B_T = 1) \gg \Pr_{\psi}(\mathbf{Y}_T \mid B_T = 0)$$

and if a marker M is closely linked to the trait locus T

$$\Pr_{\psi}(B_T = 1 \mid B_M = 1) \gg \Pr(B_T = 1) = f$$

where $\Pr(B_T = 1)$ is the prior probability of gene identity at a locus implied by the pedigree structure, which in this case is simply the inbreeding coefficient (f) of the affected individual. Finally, if the data \mathbf{Y}_M specify homozygosity of the affected individual at a polymorphic marker locus

$$\Pr(B_M = 1 \mid \mathbf{Y}_M) > \Pr(B_M = 1) = f.$$

Homozygosity at multiple linked marker loci reinforces the inference that the affected individual is IBD in this segment of the genome. Data on multiple affected individuals, all homozygous in the same genome region, together provide evidence that the hypothesized trait locus is also located in this region.

3. Exact computation of probabilities on pedigrees. In (2.1) the terms $\Pr(B_M \mid \mathbf{Y}_M)$ are the conditional probabilities of marker loci IBD status given marker data. In the current paper we shall not consider explicit trait

models, but focus on the IBD information conveyed by marker loci, and the estimation of $\Pr(B_M | \mathbf{Y}_M)$. Now

$$\Pr(B_M | \mathbf{Y}_M) = \Pr(\mathbf{Y}_M | B_M) \Pr(B_M) / \Pr(\mathbf{Y}_M)$$

and thus exact computation of the conditional probability requires the computation of $\Pr(\mathbf{Y}_M)$ the overall probability of the marker data observed on the pedigree. We consider first, therefore, the evaluation of such probabilities. Since now we consider only marker loci, we drop the subscript M .

Algorithms for the computation of probabilities on pedigrees have followed one of two paradigms. The first, dating to the early days of human linkage analysis (Fisher 1934; Haldane 1934), considers the probability of phenotypic data \mathbf{Y} as the sum over underlying genotypic configurations \mathbf{G} :

$$(3.1) \quad \Pr_{\psi}(\mathbf{Y}) = \sum_{\mathbf{G}} \Pr_{\psi}(\mathbf{Y} | \mathbf{G}) \Pr_{\psi}(\mathbf{G}).$$

Algorithms for the computation of this sum rely on the conditional independence structure of genotypes on pedigrees, which permits the summation to be performed sequentially through the pedigree structure. The best-known such algorithms derive from the algorithm of Elston and Stewart (1971), and have come to be known as “(pedigree) peeling” (Cannings, Thompson and Skolnick 1978). Generally, peeling algorithms are linear in the size of the pedigree, but exponential in pedigree complexity as measured by the number of interlocking loops. More seriously, they are exponential in the number of alternative (multilocus) genotypes an individual can have. Hence computation rapidly becomes infeasible as the number of loci increases, especially if the loci are multi-allelic.

An alternative approach also dates back to the earliest days of linkage analysis (Sturtevant 1913; Fisher 1922). This method involves direct observation or inference of the segregation events in an experimental cross, and hence scoring of the recombination events. The segregation events can be specified by “segregation indicators” $\mathbf{S} = \{S_{il}, i = 1, \dots, m, l = 1, \dots, L\}$ where

$$\begin{aligned} S_{il} &= 0 && \text{if copied gene at segregation } i \text{ locus } l \text{ is parent's maternal gene} \\ &= 1 && \text{if copied gene at segregation } i \text{ locus } l \text{ is parent's paternal gene.} \end{aligned}$$

Here $i = 1, \dots, m$ indexes the segregations, and $l = 1, \dots, L$ indexes the genetic loci. Where not all segregation events can be precisely inferred, the probability of observed data \mathbf{Y} may again be considered as a sum:

$$(3.2) \quad \Pr_{\psi}(\mathbf{Y}) = \sum_{\mathbf{S}} \Pr_{\psi}(\mathbf{Y} | \mathbf{S}) \Pr_{\psi}(\mathbf{S}).$$

Algorithms based on (3.2) rely on the conditional independence structure of the segregation indicators S_{il} , which permits the summation to be performed sequentially along the chromosome (“chromosomal peeling”): such an algorithm was developed by Lander and Green (1987). This approach is ideal for data on experimental crosses, since, in the absence of missing data, computation is linear in the number of loci. However, computation is exponential in the number of meioses which cannot be directly observed.

4. MCMC estimation of probabilities on pedigrees. For multilocus computations, on complex extended pedigrees, with many of the individuals unobserved, exact computation is infeasible with either approach. Therefore, in recent years, alternative Monte Carlo procedures for the summations in (3.1) and (3.2) have been proposed. Most of these proposals have been of Markov chain Monte Carlo (MCMC) algorithms, which rely on the same conditional independence structures as do the exact algorithms. Most of the proposals to date have considered (3.1), the objective being therefore to sample genotypes \mathbf{G} from their conditional distribution given the data \mathbf{Y} . The simplest algorithms involve single-site updating via a Metropolis (Lange and Matthyssse 1989) or Gibbs (Sheehan *et al.* (1989)] sampler. That is, the update proposal is of the genotype of a single individual at a single locus.

Such algorithms work well on small examples, but do not mix adequately on large pedigrees, especially where there are many unobserved individuals, and/or data at many loci at which multilocus phase is not easily determined. Moreover, for multi-allelic loci, the partial constraints imposed by data may make the single-site updating MCMC methods reducible. There have been numerous proposals to ensure irreducibility of samplers, and to improve mixing. Some examples are the “heating” methods of Sheehan and Thomas (1993) and of Lin *et al.* (1994), the “tunneling” method proposed by Sobel and Lange (1993), the “mode-jumping” method proposed by Lin (1995), and the simulated tempering approach developed by Geyer and Thompson (1995).

Thompson (1994a) proposed use of the alternative paradigm (3.2), in which MCMC sampling is of the segregation indicators \mathbf{S} conditional upon data \mathbf{Y} . Where there are many unobserved individuals on a pedigree, especially for multi-allelic loci, the space of segregation indicators is much smaller than the space of genotypes. It is generally much less constrained by data, except where components are fully determined (see section 7). For the estimation of the posterior probabilities of gene IBD patterns at marker loci, it has the added advantage that the IBD pattern B is fully determined by the segregation indicators \mathbf{S} . Consider, for example, the segregation pattern on the pedigree shown in Figure 2. The founder genes are labeled $1, \dots, 2n$ where n is the number of founders, and

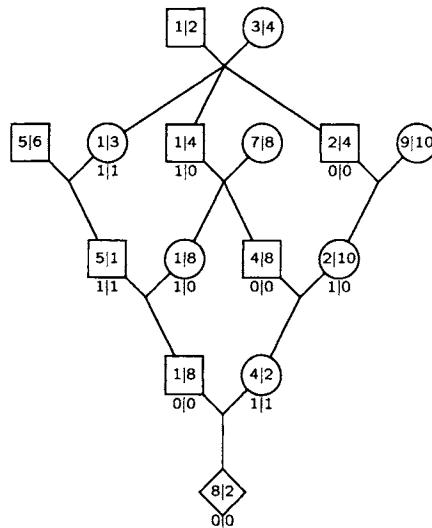


FIG. 2. *The same pedigree structure as Figure 1, showing a particular single-locus realization of the segregation indicators, and the implied gene identities.*

the genes of non-founders are then determined successively by the segregation indicators. In particular, we see than for this realization the final individual does not have two IBD genes. We see also that, although the individual receives his mother's maternal gene (gene "2"), he shares the founder gene "8" IBD with his maternal grandfather.

Thompson (1994b) developed a single-site Metropolis algorithm for sampling the S_{il} , and implemented it in the context of homozygosity mapping where normally the only data are on a single inbred individual in each pedigree. In this case, the segregation-indicator sampler performs much better than the genotypic sampler. However, the single-site updating scheme does not work well when the loci are very tightly linked, since the proposal to update a single locus then often involves the formation of a double-recombinant. Sobel and Lange (1996) also implemented a single-site Metropolis algorithm for \mathbf{S} in a variety of pedigree analysis situations, with similar conclusions as to performance. In this paper, we propose a meiosis-by-meiosis sampler which updates S_{il} jointly for all loci l in a given meiosis i .

5. Implementing the whole-meiosis Gibbs sampler. For notational convenience we define $S_{*l} = (S_{il}, i = 1, \dots, m)$ the vector of segregation indicators at locus l , and $S_{i*} = (S_{il}, l = 1, \dots, L)$ the vector of indicators at segregation i .

In order to implement a whole-meiosis Gibbs sampler for S_{i*} we must compute

$$\Pr(S_{i*} \mid \{S_{k*}, k \neq i\}, \mathbf{Y})$$

We suppose that the marker data \mathbf{Y} can be partitioned into data relating to each locus $l = 1, 2, \dots, L$, and that the loci are numbered in order along the chromosome. Then $\mathbf{Y} = (Y_1, \dots, Y_L)$. Let $Y^{(l)} = (Y_1, \dots, Y_l)$, so $\mathbf{Y} = Y^{(L)}$. We suppose also that $\Pr(Y_l \mid S_{*l})$ can be easily computed: we show below how this computation may be done. Now define

$$(5.1) \quad Q_l(s) = \Pr(S_{il} = s \mid \{S_{k*}, k \neq i\}, Y^{(l)}) \text{ for } s = 0, 1.$$

That is, $Q_l(s)$ is the cumulative probability for the segregation indicator S_{il} , given the data at loci up to and including locus l . Then

$$(5.2) \quad Q_1(s) \propto \Pr(Y_1 \mid S_{*1}) \text{ and}$$

$$Q_l(s) \propto \Pr(Y_l \mid S_{*l}) (Q_{l-1}(s)(1 - \theta_{l-1}) + Q_{l-1}(1 - s)\theta_{l-1}) \text{ for } l = 2, \dots, L,$$

where S_{*l} takes the current value at meioses other than i , and the value s for meiosis i , and where θ_{l-1} is the recombination frequency between locus $l-1$ and locus l . Thus we may compute (5.1) for each l in turn, working forwards sequentially along the chromosome.

Finally we have computed

$$Q_L(s) = \Pr(S_{iL} = s \mid \{S_{k*}, k \neq i\}, \mathbf{Y} = Y^{(L)})$$

and thus S_{iL} may be sampled from this desired conditional distribution. Suppose S_{ij} has been similarly sampled for $j = l, \dots, L$. Then

$$(5.3) \quad \begin{aligned} & \Pr(S_{i,l-1} = s \mid \{S_{k*}, k \neq i\}, \{S_{ij}, j = l, \dots, L\}, \mathbf{Y}) \\ & \propto Q_{l-1}(s) (|S_{il} - s|\theta_{l-1} + (1 - |S_{il} - s|)(1 - \theta_{l-1})) \end{aligned}$$

Thus we may work backwards down the chromosome, sampling each S_{il} in turn ($l = L, \dots, 1$), obtaining overall a joint realization of S_{il} , $l = 1, \dots, L$ from its conditional distribution given $\{S_{k*}, k \neq i\}$ and \mathbf{Y} . We note the similarity of this forwards-backwards algorithm (equations (5.2) and (5.3)) along a chromosome to the method of Ploughman and Boehnke (1988), which samples genotypes jointly at a locus by peeling up the pedigree, saving the partial probabilities computed en route, and then sampling down using these partial probabilities. The same method is used in a method to determine feasible genotypic configurations on a pedigree (Heath 1997a), as well as in MCMC genotypic samplers that sample jointly all genotypes at a locus (Kong 1991; Heath 1997b).

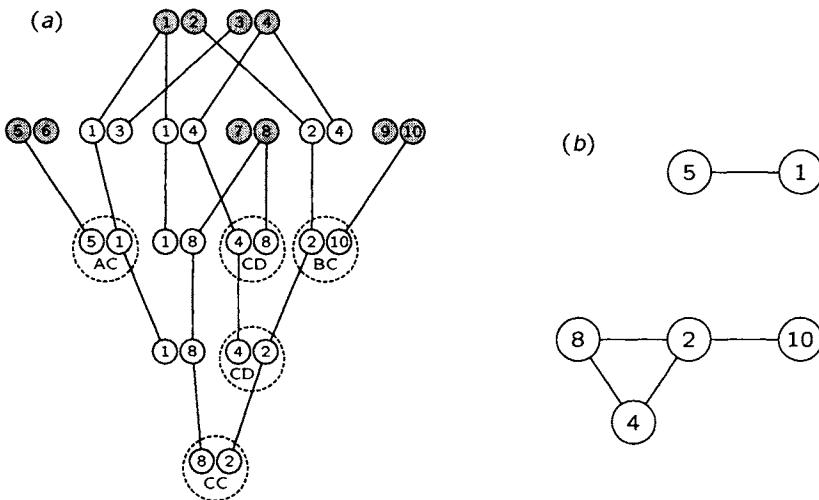


FIG. 3. (a) The gene pedigree implied by the segregation pattern on the pedigree of figure 2, showing marker data on five individuals, and (b) the gene dependency graph resulting from the segregation pattern of and marker data of (a).

For completeness we outline briefly the method for efficient computation of the single-locus probability $\Pr(Y_l | S_{*l})$, where $S_{*l} = (S_{il}, i = 1, \dots, m)$. We illustrate the calculation with the small but complex pedigree shown in Figure 1. As before, the founder genes in the pedigree are labeled from $1 \dots 2n$, where n is the number of founders in the pedigree (five in this case). For non-founder individuals, the genes they carry at locus l is determined by S_{*l} , the segregation pattern for that locus (Figure 2). Figure 3(a) shows the same pedigree, but with the individual genes rather than individuals drawn. The ten founder genes (shaded in the figure) have been labeled, and the figure shows the descent of genes to the non-founders for a particular realization of S_{*l} . For this example, five of the individuals are observed; these individuals are marked by the dotted circles on the figure.

Only genes that appear in observed individuals contribute non-trivially to the single-locus probability $\Pr(Y_l | S_{*l})$ so in this example we only have to consider the six genes 1, 2, 4, 5, 8 and 10. However, calculation requires summing over all possible assignments of allelic identities to these six distinct genes (Thompson 1974). A naïve approach for a locus with k alleles would require summing over all k^6 possible combinations of allelic identities. We can improve on this by exploiting the dependence between founder genes. This dependence structure can be shown by a graph whose nodes are the genes that appear in observed individuals. An edge connects two genes if they appear together in an observed individual (Figure 3(b)). If, as in this example, the graph has several

components, then each can be considered separately for the purposes of calculating the probability. For codominant markers each component can have at most two possible joint assignments of allelic identities for the genes in the cluster, so the probability calculation becomes trivial (Sobel and Lange 1996). For this example the gene cluster (1, 5) has 2 possible allelic assignments, (A, C) or (C, A), and the cluster (2, 4, 8, 10) has 1 possible assignment, (C, D, C, B). For general loci, the graph of Figure 3(b) defines a conditional independence structure. The desired probability can thus be calculated efficiently by “peeling” the allelic assignment of types to founder genes, in a method analogous to pedigree peeling (Elston and Stewart 1971; Cannings *et al.* 1978) or chromosomal peeling (Lander and Green 1987). Additional details of the calculation are given in Heath and Thompson (in preparation).

Any MCMC sampler needs an initial configuration for the latent variables. In the small examples considered here, values of \mathbf{S} that are consistent with the data were found by hand. For larger or more complex examples, we may use existing methods for obtaining genotypic configurations \mathbf{G} that are consistent with observed data \mathbf{Y} even for highly polymorphic loci on large and complex pedigrees (Heath 1997a). Where the pedigree can be peeled for single-locus data, the initial configuration is from the required equilibrium distribution marginally for each locus. The method produces an ordered genotype for each individual, and this genotypic configuration then provides the implied segregation indicators. These indicators are then necessarily also consistent with the data \mathbf{Y} . This is not necessarily the best way to obtain a starting configuration \mathbf{S} , but it is a possible and practical way for which the programs already exist.

6. Performance of the sampler: two examples. For examples, we use the small but complex pedigree (Figure 1), considered in the previous sections. We consider first the case of homozygosity mapping which was the objective of the original study (Goddard *et al.* 1997). Only marker data on the final affected individual were available. This pedigree provides a useful example, since while it is easily analyzed by MCMC methods, exact computation by pedigree peeling is infeasible for more than about four loci, due to the three interlocking loops. Due to the 20 meioses in the pedigree, this example is also close to the limits of feasibility for exact computation using chromosomal peeling.

For homozygosity mapping, the question is of the extent that patches of marker homozygosity imply gene identity by descent in the sampled individual. As output from our sampler, we therefore score the IBD pattern at a set of homozygous marker loci. The single-site Metropolis sampler was previously implemented for this case (Thompson 1994a,b), and we now compare this with the whole-meiosis Gibbs sampler for the same situation. As an example, we con-

TABLE 1

Autozygosity probabilities conditional on homozygosity at five tightly linked loci (recombination = 0.02), as a function of the frequency of the homozygous marker allele. There are 2 states not involving any switches between N (non-IBD) and I (IBD), 8 states involving one switch (e.g. N N N I I), 12 involving two switches (e.g. I I N N I), 8 involving three switches (e.g. N I I N I) and two involving four switches (e.g. N I N I N), for a total of $2^5 = 32$ possible patterns.

pattern of IBD	$q = 1.0$	$q = 0.5$	$q = 0.1$
N N N N N	0.8389	0.2362	0.00013
I I I I I	0.0644	0.5819	0.9673
1-switch (8)	≈ 0.08	≈ 0.15	≈ 0.03
2-switch (12)	≈ 0.01	≈ 0.03	≈ 0.001
other (10)	≈ 0	≈ 0	≈ 0
single-locus Pr(I)	0.1094	0.1972	0.5512

sider five equally-spaced marker loci (L1 to L5), with recombination frequency 0.02 between adjacent loci. Table 1 shows the results as a function of q , the frequency of the allele for which the observed individual is homozygous.

Figure 4 compares the cumulative IBD probabilities at each locus for the single-site and whole-meiosis samplers in the case when the marker allele frequency is 0.5. To provide a fair comparison, there are five times as many single-site updates as whole meiosis updates. (Each total run is 10,000,000 whole-meiosis updates, or 50,000,000 single-site updates.) Clearly the whole-meiosis sampler has much better mixing, and provides more reliable results. Furthermore, the CPU time for the run using the whole-meiosis sampler was only about 2/3 of that for the single-site sampler: 10 million whole-meiosis Gibbs updates took 328 secs CPU, while 50 million single-site updates took 467 seconds on a DEC Alpha 400M workstation. (The efficiencies of the two programs are quite comparable. Each could be further optimized.)

Over the five loci, as expected, the central locus (L3) of the five homozygous markers has the highest IBD probability, followed by the next two loci (L2, L4), with the end loci (L1, L5) having the lowest IBD probability. For these very tightly linked loci the differences in IBD probability are not large, but they are non-negligible. For both samplers we see very strong correlations among the five loci for the cumulative IBD probability; the five paths track each other closely. This correlation is due to the tight linkage. For unlinked loci, there are no such correlations, even when the sampler is run jointly on the loci from a common starting configuration (results not shown).

We now consider also the case where other individuals are observed on this same pedigree structure; for example, the data of Figure 3, shown again in Figure 5. We see the C allele labeled C_2 must descend from grandparent to parent, to the final individual, while the two D alleles must also be IBD. However, the

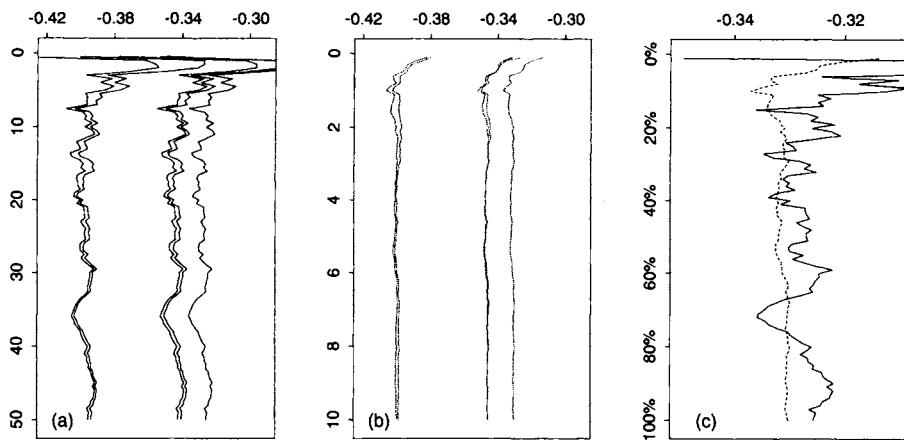


FIG. 4. Comparison of the single-site Metropolis sampler and the whole-meiosis Gibbs sampler for the example of homozygosity mapping. In each case there are five linked marker loci, with recombination frequency 0.02 between adjacent loci, and the allele frequency of the marker allele for which the final individual is assumed homozygous is 0.5 at each locus. On the horizontal axis of each graph is the \log_e of the probability of gene identity by descent between the two haplotypes of the final individual. Plotted are the cumulative estimates over each run, for each locus. (a) Plot for the five linked loci, over a total of 50 million single-site updates. (b) Plot for the five linked loci, over a total of 10 million whole-meiosis Gibbs updates. (c) For easier comparison, the curve for the central one of the five loci, for each of (a) and (b)

other three C alleles, labeled C_1 , C_3 and C_4 in Figure 5, may or may not be IBD to each other or to C_2 . In fact, each of the 15 partitions of these four genes is possible, given the data and the pedigree structure. A question of interest might be the posterior probabilities of the patterns of gene identity among these four potentially distinct C alleles.

As above we consider five linked marker loci. To assess the effect of tight linkage, we present results for two different recombination frequencies between adjacent loci; tight linkage ($\theta = 0.02$), and loose linkage ($\theta = 0.1$). We assume the same marker data (Figure 5) at each of the five loci. At each locus, the same allele frequencies are assumed; the allele C has frequency 0.4, and each of the other alleles (A , B , and D) has frequency 0.2. Table 2 gives the results, again as a function of the frequency of allele C . For comparison, we give also the single-locus probability, and also the prior probability of identity patterns among the four genes, given only the pedigree structure.

Each of the probabilities in Table 2 is estimated from 10 million whole-meiosis Gibbs steps. For a problem with 5 loci on this size of pedigree, such a run takes just under 45 minutes CPU on a DEC Alpha 400M workstation. As for the homozygosity probabilities, examination of the cumulative state probabilities

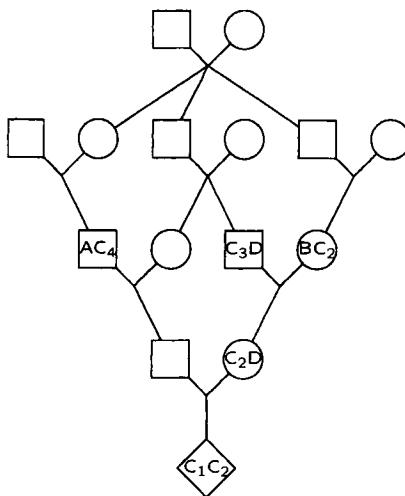


FIG. 5. The pedigree structure of Figure 1, with the data of Figure 3, and the four possibly distinct C alleles labeled.

over the run show that the sampler mixes well.

7. Irreducibility: neither necessary nor sufficient. There are many ways to make an MCMC sampler irreducible. Note first that, provided recombination frequencies are strictly positive, irreducibility is a single-locus question. Thus, irreducibility (or otherwise) is the same for a single-segregation-indicator updating sampler as for the whole-meiosis Gibbs sampler. As we have seen, the mixing properties of these two samplers can differ greatly. Irreducibility is not a sufficient criterion for a sampler; in practice, a more important question concerns its mixing properties.

However, there are some interesting features of the irreducibility properties of a sampler based on \leqslant , which illuminate the structure of the problem. Note also that, whereas \emptyset constrains the allelic types of genes, \leqslant constrains only which genes must be of like allelic type. Thus there are many examples in which a genotypic sampler is reducible, but in which a segregation-indicator sampler is irreducible. Generally, irreducibility of the MCMC sampler of \leqslant can fail when the allelic types of founder genes are constrained, either directly through founders of observed genotype or through constraints on the number of distinct founder genes.

Reducibility is not a problem for homozygosity mapping when only a single inbred individual is observed in a pedigree (Thompson 1994b). At a given locus l , the two genes are either IBD or not, and the latter state is necessarily

TABLE 2

Probabilities of the fifteen gene identity patterns among the four potentially distinct C alleles. The frequency of marker allele C is assumed to be 0.4 at each locus, and the recombination frequency between adjacent loci is either 0.02 or 0.1 as indicated.

pattern $C_1 C_2 C_3 C_4$	$\theta = 0.02$			$\theta = 0.1$			single locus	
	L3	L2,L4	L1,L5	L3	L2,L4	L1,L5	post.	prior
g g g g	0.598	0.590	0.561	0.382	0.359	0.287	0.049	0.029
g g g h	0.216	0.214	0.207	0.191	0.183	0.155	0.039	0.060
g g h g	0.038	0.039	0.046	0.076	0.079	0.091	0.091	0.137
g h g g	0.049	0.051	0.059	0.102	0.107	0.121	0.118	0.088
g h h h	0.006	0.007	0.010	0.014	0.016	0.021	0.013	0.070
g g h h	0.003	0.003	0.004	0.007	0.007	0.008	0.006	0.002
g h g h	0.016	0.017	0.019	0.030	0.032	0.035	0.039	0.012
g h h g	0.037	0.039	0.046	0.081	0.086	0.100	0.084	0.229
g g h u	0.004	0.004	0.005	0.011	0.012	0.015	0.026	0.004
g h g u	0.022	0.022	0.026	0.047	0.050	0.059	0.125	0.123
g h u g	0.004	0.005	0.008	0.031	0.036	0.057	0.270	0.164
g h u h	0.001	0.001	0.001	0.004	0.005	0.009	0.027	0.055
g h h u	0.005	0.006	0.008	0.015	0.017	0.023	0.026	0.088
g h u u	0.001	0.001	0.001	0.004	0.005	0.009	0.026	0.052
g h u v	0.001	0.001	0.001	0.004	0.005	0.010	0.066	0.055

consistent with the data. Sobel and Lange (1996) give an example where the single-site Metropolis sampler of $\{S_{il}\}$ is reducible, due to severe constraints on the types of founder genes. In practice, we are unlikely to have fully observed homozygous founders, except perhaps in crosses among inbred lines. Note that data on descendants can never force an unobserved ancestor to be homozygous, and nor can data on descendants alone force the maternal/paternal origins of genes in an unobserved ancestor. Thus the example of Sobel and Lange (1996) is unlikely to be a practical concern in human genetics.

However, reducibility can also arise from restrictions on the number of founder genes, or number of genes available to segregate to observed descendants. The segregation indicators S_{*l} define a partition of the ordered genes at locus l in observed individuals determining which are identical by descent, and hence must be of like allelic type. Note that if any given partition is consistent with the data, then any finer partition must be so also.

Figure 6(a) shows an example in which there are three full sibs with unobserved parents. The sibs have genotypes AB , AC and BC as shown. Note that any one of the three alleles must be present in each parent, and the other two must be represented once only among the four founder genes. In each case, one pair of sibs share no genes IBD with each of the pair sharing one gene IBD with the third sib. The set of six observed genes at each locus are partitioned into four sets of IBD genes, two partitions size two, and two partitions size one.

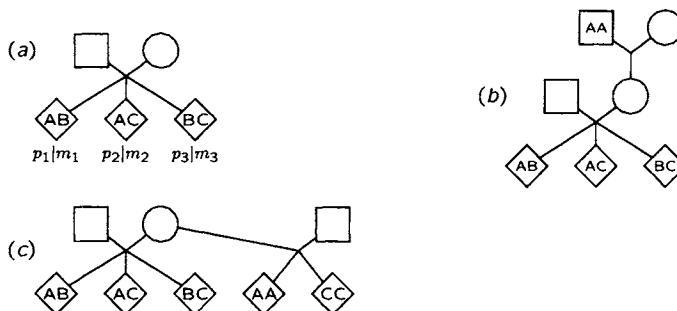


FIG. 6. Examples of failure of irreducibility. (a) Data on a sibship of size three. (b) An added maternal grandfather (case (b) in Table 3). (c) Added information on maternal genotype (case (c) in Table 3).

States are compatible with the data if two of the sibs have indicators $(0, 1)$ and $(1, 0)$, and the third is either $(1, 1)$ or $(0, 0)$, or if two of the sibs have indicators $(0, 0)$ and $(1, 1)$, and the third is either $(1, 0)$ or $(0, 1)$. Of the 64 potential values of the 6 binary segregation indicators, 24 are consistent with the data (Table 3).

If only a single meiosis is updated in a given step, the 24 feasible states fall into two disjoint cycles of length 12. The sampler is reducible. In Table 3, the states are listed in order so that each is obtainable from the previous one by updating a single meiosis. Each feasible state implies an ordered pair of parental genotypes. In Table 3, the parental genotypes are given as implied by the first column of 12 states. For the second column of states, the father's genotype is reversed. Thus, although the space of \leq values that are consistent with the phenotypic data is divided into two non-communicating classes, this does not affect MCMC estimation of probabilities. The founder (i.e. parental) genotypes, whose probabilities determine the contribution of a state to any overall probability, are identical on each of the two communicating classes. Thus in this example we see that irreducibility is not necessary in order to obtain correct MCMC probability estimates at a single locus.

We can extend this example. Suppose first the mother's father is known to be of type AA , forcing the mother's paternal gene to be of type A (Figure 6(b)). Then only the 8 states labeled "b" in Table 3 are consistent with the data, and these fall into two non-communicating classes, each of size 4. However, again the same set of parental genotypes, and hence the same probability contributions, are implied by each class of states.

Or, we could suppose the existence of other maternal half sibs (e.g. AA and CC) forcing the mother to be of type AC (Figure 6(c)). Then only the 8 states labeled "c" are consistent with the data. Now we have four non-communicating classes, each of size two. Again, the same parental genotype combinations are

TABLE 3

The feasible states for the three-sibs example. States are specified as $(p_1, m_1, p_2, m_2, p_3, m_3)$ the paternal and maternal indicators for each of the three sibs. The data are that the sibs have genotypes AB, AC and BC. For further details, see text.

sib 1 p_1, m_1	sib 2 p_2, m_2	sib 3 p_3, m_3	sib 1 p_1, m_1	sib 2 p_2, m_2	sib 3 p_3, m_3		parental types father mother $p\ m\ p\ m$
0 0	0 1	1 0	1 0	1 1	0 0		CA CB
0 0	0 1	1 1	1 0	1 1	0 1		BA CB
1 0	0 1	1 1	0 0	1 1	0 1	c	BA CA
1 0	0 0	1 1	0 0	1 0	0 1	c	BC CA
1 0	0 0	0 1	0 0	1 0	1 1		BC BA
1 1	0 0	0 1	0 1	1 0	1 1		AC BA
1 1	1 0	0 1	0 1	0 0	1 1		AC BC
1 1	1 0	0 0	0 1	0 0	1 0	b c	AB BC
0 1	1 0	0 0	1 1	0 0	1 0	b c	AB AC
0 1	1 1	0 0	1 1	0 1	1 0	b	CB AC
0 1	1 1	1 0	1 1	0 1	0 0	b	CB AB
0 0	1 1	1 0	1 0	0 1	0 0	b	CA AB

implied by each of the four classes. We could further require both the above conditions to be met, reducing the feasible states space to only four states in two classes (those labeled both "b" and "c"). Yet again the parental genotypes are the same in each class of communicating feasible states. Thus the symmetries of this example mean that irreducibility is unnecessary in obtaining valid MCMC single-locus probability estimates.

However, although single-locus irreducibility implies multi-locus irreducibility, validity of single-locus probabilities does not imply validity of multilocus probabilities in a reducible sampler. (We are indebted to Ken Lange for drawing our attention to this fact.) A simple example will suffice; consider the phenotypic data of Figure 6(b) at each of two very tightly linked loci. If both loci are initialized in the same class of four feasible states, approximately correct probabilities are obtained. If one locus is initialized in each of the two different sets of four states, at least two recombination events are required in the six meioses. For very tight linkage, absolute probabilities are almost negligible, but the relative probabilities bias the sampler towards the states where only these two recombination events are required. In this example, the sampler is biased towards states where *C* is non-IBD at one locus and *B* is non-IBD at the other, and away from the states where each parent carries an *A* allele. In general, it is not easily determined whether irreducibility is necessary.

In examples we have considered, irreducibility fails due to constraints of equality or inequality of segregation indicators from a given parent. There is a

strong constraint on the number of available genes, since the parent individual can have at most two distinct genes at a locus. Simultaneous updating of all the meioses from a given individual, or even parental couple, will often be possible, using the computational algorithm of Kruglyak *et al.* (1995), and is a practical way to obtain irreducibility in many cases. However, it is not a universal solution, as is shown by the example of Sobel and Lange (1996) in which two indicators in different sibships are constrained to be unequal.

8. Discussion. As maps become both denser and more precise, there is an increasing demand for multipoint linkage analysis on large and complex pedigrees. Exact computations become infeasible, necessitating the use of Monte Carlo or other approximation methods. The whole-meiosis Gibbs sampler presented here is just one of many possible Monte Carlo algorithms. It is easily implemented, and mixes much better than any single-site MCMC method, particularly when there is tight linkage. In some cases, multi-site genotypic samplers can be implemented. In particular, where pedigree peeling is feasible for each locus separately, a whole-locus Gibbs sampler is possible (Kong 1991; Heath 1997b), and is likewise a great improvement over a single-site updating genotypic scheme. Genotypic samplers work well if there are few missing data on the pedigree, but where there are many unobserved individuals we expect the whole-meiosis sampler to have better performance. For very complex pedigrees, even single-locus (pedigree) peeling is computationally intensive, whereas implementation of the whole-meiosis sampler is almost unaffected by pedigree complexity. For very tight linkage, any sampler will have decreased mixing. However, whereas the performance of a genotypic sampler can be severely adversely affected by tight linkage, the whole-meiosis Gibbs sampler performed well, even for multiple loci at recombination frequencies as low as 0.02.

Exact computational and genotypic sampling methods preclude the inclusion of interference for more than three loci, except on very small pedigrees with few missing data (Lin and Speed 1996). However, a meiosis sampler can include interference in the computation of linkage likelihoods or conditional probabilities of genome sharing. The recombination events within a meiosis may be jointly sampled, or a recombination location may be resampled conditionally upon the locations of other recombinations in the same meiosis. Equations 5.2 and 5.3 become more complex, since dependence in S_{il} extends beyond the adjacent marker loci, but the same approach is computationally feasible.

For almost any sampler, irreducibility is a non-trivial question. Generally, segregation indicators provide fewer constraints on genes. Except where founder genes are constrained in either type or number the sampler will often be irreducible. However, in pedigrees with few founders and for loci with many alleles,

it is possible for irreducibility of the whole-meiosis Gibbs sampler to fail. In the examples of section 6, irreducibility is easily established. For homozygosity mapping, irreducibility is trivially satisfied. For the case of the five observed individuals on the same pedigree, one segregation indicator is completely determined; the final individual receives the C_2 allele from his mother's mother. However, parental origins of genes cannot be determined from data on descendants alone; other segregation indicators are freely varying. Where an indicator is completely determined, this reduces the size of the space to be sampled, but of course does not affect the irreducibility of the sampler. (The same is true of determined genotypes in a genotypic sampler.)

This fact could be used to improve the efficiency of the sampler, and to simplify consideration of irreducibility, by conditioning on the segregation from any founder having only one offspring. Such a founder provides no information for linkage, and the conditioning simply determines the ordering of the genes in the founder. This method was used by Thompson (1994b) to improve efficiency of the single-site Metropolis sampler for homozygosity mapping, and has been extended to the whole-meiosis sampler. Similar considerations have been used by Kruglyak et al. (1995) to extend feasibility of exact computational methods using the approach of Lander and Green (1987). Even where a founder has multiple offspring, at a single locus the segregation to a given offspring (say the eldest) may be constrained. However, where a founder has multiple offspring and data at multiple loci are available on descendants, such data can provide partial information on which offspring of the founder are recombinant. It is necessary for the computation to allow for alternate recombination patterns in the meiosis from founder to each offspring, within each linkage group. Similarly to the situation with a reducible MCMC sampler, caution is necessary in ensuring that correct multilocus probabilities are computed.

Where recombination frequencies are strictly positive, irreducibility can be assessed on a single-locus basis. Thus, in particular, irreducibility of the single-site segregation-indicator sampler is the same as for the whole-meiosis Gibbs sampler. More importantly, where a whole-locus genotypic Gibbs sampler is feasible, it is necessarily irreducible. This guarantee of irreducibility, balanced against the greater computational burden and possibly poorer mixing of the genotypic updates, raises the attractive possibility of combining the two samplers. Since a given segregation configuration can be used easily to obtain a genotypic realization, and a genotypic configuration supplies a segregation configuration, the combination of the two samplers is quite practical, wherever single-locus peeling is feasible. In some situations, a sampler interleaving whole-meiosis and whole-locus updates has better mixing properties than either alone (Heath and Thompson 1997). We intend a more detailed study of the whole-

meiosis Gibbs sampler, the whole-locus Gibbs sampler, and of samplers which combine the two approaches, to determine the preferred sampler under a variety of pedigree structures and linkage patterns (Heath and Thompson [in preparation]).

Acknowledgments. We are very grateful to Dr. Bernt Guldbrandtsen and to a referee for careful reading of the details of this paper.

REFERENCES

- CANNINGS, C., THOMPSON, E. A. and SKOLNICK, M. H. (1978). Probability functions on complex pedigrees. *Adv. Appl. Prob.* **10** 26–61.
- ELSTON, R. C. and STEWART, J. (1971). A general model for the genetic analysis of pedigree data. *Human Heredity* **21** 523–542.
- FISHER, R. A. (1922b). On the systematic location of genes by means of crossover observations. *American Naturalist* **56** 406–411.
- FISHER, R. A. (1934). The amount of information supplied by records of families as a function of the linkage in the population sampled. *Ann. Eugen.* **6** 66–70.
- GEYER, C. J. and THOMPSON, E. A. (1995). Annealing Markov chain Monte Carlo with applications to ancestral inference. *Journal of the American Statistical Association* **90** 909–920.
- GODDARD, K. A. B., YU, C-E., OSHIMA, J., MIKI, T., NAKURA, J., PIUSSAN, C., MARTIN, G. M., SCHELLENBERG, G. D., WIJSMAN, E. M. (1996). Towards localization of the Werner syndrome gene by linkage disequilibrium and ancestral haplotyping: lessons learned from analysis of 35 chromosome 8p11.1-21.1 markers. *American Journal of Human Genetics* **58** 1286–1302.
- HALDANE, J. B. S. (1934). Methods for the detection of autosomal linkage in man. *Annals of Eugenics* **6** 26–65.
- HEATH, S. C. (1997a). Generating consistent genotypic configurations for multi-allelic loci and large complex pedigrees. *Human Heredity* **48** 1–11.
- HEATH, S. C. (1997b). Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. *American Journal of Human Genetics* **61**: 748–760.
- HEATH, S. C. and THOMPSON, E. A. (1997). MCMC Samplers for multilocus analyses on complex pedigrees. *American Journal of Human Genetics* **61**: A278.
- KONG, A. (1991). Analysis of pedigree data using methods combining peeling and Gibbs sampling. *Computer Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, (E.M. Keramidas and S. M. Kaufman, eds.) Pp 379–385. Interface Foundation of North America, Fairfax Station, VA.
- KRUGLYAK, L., DALY, M. J., and LANDER, E. S. (1995). Rapid multipoint linkage analysis of recessive traits in nuclear families including homozygosity mapping. *American Journal of Human Genetics* **56** 519–527.
- LANDER, E. S. and BOTSTEIN, D. (1987). Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. *Science* **236**: 1567–1570.
- LANDER, E. S. and GREEN, P. (1987). Construction of multilocus linkage maps in humans. *Proceedings of the National Academy of Sciences (USA)* **84** 2363–2367.
- LANGE, K. and MATTHYSSE, S. (1989). Simulation of pedigree genotypes by random walks. *American Journal of Human Genetics* **45** 959–970.
- LIN, S. (1995). A scheme for constructing an irreducible Markov chain for pedigree data. *Biometrics* **55** 318–322.
- LIN, S. and SPEED, T. P. (1996) Incorporating crossover interference into pedigree analysis using the χ^2 model. *Human Heredity* **46**: 315–322.

- LIN, S., THOMPSON, E. A. and WIJSMAN, E. M. (1994). A faster mixing algorithm for Hastings-Metropolis updates on complex pedigrees. *Annals of Human Genetics* **58** 343-357.
- PLoughman, L. M. and BOEHNKE, M. (1989). Estimating the power of a proposed linkage study for a complex genetic trait. *American Journal of Human Genetics* **44** 543-551.
- SOBEL, E. and LANGE, K. (1993). Metropolis sampling in pedigree analysis. *Statistical Methods in Medical Research* **2** 263-282.
- SOBEL, E. and LANGE, K. (1996). Descent graphs in pedigree analysis: Applications to haplotyping, location scores and marker-sharing statistics. *American Journal of Human Genetics* **58** 1323-1337.
- SHEEHAN, N. A., POSSOLO, A. and THOMPSON, E. A. (1989). Image processing procedures applied to the estimation of genotypes on pedigrees. *American Journal of Human Genetics* **45** (Suppl.) A248.
- SHEEHAN, N. A. and THOMAS, A. W. (1993). On the irreducibility of a Markov chain defined on a space of genotype configurations by a sampling scheme. *Biometrics* **49** 163-175.
- STURTEVANT, A. H. (1913). The linear association of six sex-linked factors in *Drosophila*, as shown by their mode of association. *J Exp Zool* **14** 43-59.
- THOMPSON, E. A. (1974). Gene identities and multiple relationships. *Biometrics* **30** 667-680.
- THOMPSON, E. A. (1994a). Monte Carlo estimation of multilocus autozygosity probabilities. *Proceedings of the 1994 Interface Conference* J. Sall and A. Lehman (eds.) Pp. 498-506, Interface Foundation of North America: Fairfax Station, VA.
- THOMPSON, E. A. (1994b). Monte Carlo likelihood in linkage analysis. *Statistical Science* **9** 355-366.
- WIJSMAN, E. M. and AMOS, C. I. (1997). Genetic analysis of simulated oligogenic traits in nuclear and extended pedigrees: Summary of GAW10 contributions. *Genetic Epidemiology* **14** 719-735.

DEPARTMENT OF STATISTICS
UNIVERSITY OF WASHINGTON
BOX 354322, SEATTLE WA 98195-4322
THOMPSON@STAT.WASHINGTON.EDU

LABORATORY OF STATISTICAL GENETICS
THE ROCKEFELLER UNIVERSITY
NEW YORK, NY 10021
HEATH@LINKAGE.ROCKEFELLER.EDU

A REVIEW OF METHODS FOR IDENTIFYING QTLS IN EXPERIMENTAL CROSSES

BY KARL W. BROMAN¹ AND T. P. SPEED

Marshfield Medical Research Foundation and University of California, Berkeley

Identifying the genetic loci contributing to variation in traits which are quantitative in nature (such as the yield from an agricultural crop or the number of abdominal bristles on a fruit fly) is a problem of great importance to biologists. The number and effects of such loci (called quantitative trait loci or QTLs) help us to understand the biochemical basis of these traits, and of their evolution in populations over time. Moreover, knowledge of these loci can aid in designing selection experiments to improve the traits.

There are a large number of different methods for identifying the QTLs segregating in an experimental cross. Little has been written critically comparing the methods, and there have been few studies comparing their performance; we make an attempt at this.

1. Introduction. Identifying the genetic loci contributing to variation in traits which are quantitative in nature (such as the yield from an agricultural crop or the number of abdominal bristles on a fruit fly) is a problem of great importance to biologists. The number and effects of such loci (called quantitative trait loci or QTLs) help us to understand the biochemical basis of these traits, and of their evolution in populations over time. Moreover, knowledge of these loci can aid in designing selection experiments to improve the traits.

There are a large number of different methods for identifying the QTLs segregating in an experimental cross. Little has been written critically comparing the methods, and there have been few studies comparing their performance; we make an attempt at this.

There are three major points which we would like to make in this paper: first, identifying QTLs is best seen as a model selection problem; most current methods do not view the problem in this way. Second, it is important to critically compare the different approaches to any problem; more refined analyses and more complicated algorithms do not necessarily lead to improved results. Finally, different situations may require different methods; with each new experiment, one must reconsider the available approaches, as each new problem may require a new method.

We will focus on a backcross experiment, and will assume that the QTLs

¹This work was part of the author's Ph.D. dissertation at the University of California, Berkeley.
AMS 1991 subject classifications. Primary 92D10; secondary 62J05.

Key words and phrases. QTL, interval mapping, model selection.

act additively. Identifying interactions between loci is a much more difficult problem; considering only the simple additive case will lead to greater clarity. We concentrate almost exclusively on detecting QTLs, considering the estimation of the QTLs' effects and precise locations of secondary importance.

In the remaining part of this section, we will describe the backcross experiment, the resulting data, some statistical models and the goals of the QTL experiment and its analysis. In Section 2, we will review the major approaches to identifying QTLs. In Section 3, we will describe the results of some simulations to compare the performance of a number of the most important methods.

1.1. Experiments. Most experiments aimed at identifying quantitative trait loci (QTLs) begin with two pure-breeding lines which differ in the trait of interest. We will call these the low (L) and high (H) parental lines. The lines are the result of intensive inbreeding, so that each is essentially homozygous at all loci (meaning that, at each locus, they received the same allele from each of their two parents). Crossing these two parental lines gives the first filial (or F_1) generation. The F_1 individuals receive a copy of each chromosome from each of the two parental lines, and so, wherever the parental lines differ, they are heterozygous. All F_1 individuals will be genetically identical, just as the individuals in each of the parental lines were.

In a backcross, the F_1 individuals are crossed to one of the two parental lines, for example, the low line. The backcross progeny, which may number from 100 to over 1000, receive one chromosome from the low parental line, and one from the F_1 . Thus, at each locus, they have genotype either LL or HL. As a result of crossing over during meiosis (the process during which gametes or sex cells are formed), the chromosome received from the F_1 parent is a mosaic of the two parental chromosomes. At each locus, there is half a chance of receiving the allele from the low parental line (L) and half a chance of receiving the allele from the high parental line (H). The chromosome received will alternate between stretches of L's and H's.

The goal is to look for an association between the phenotype of an individual and whether it received the L or H allele from the F_1 parent at various marker loci.

We use the backcross as our chief example, because of its simplicity. At each locus in the genome, the backcross progeny have one of only two possible genotypes. The strategies developed for analyzing backcross experiments will generally work for other experimental crosses as well.

1.2. Data. In an experiment like a backcross, each of the progeny is scored for one or more traits. (We will consider only one trait.) In addition, the progeny

are typed at a number of genetic markers: at each marker, it is determined whether the allele an individual received from the F_1 parent was that from the low or high parental line. Thus, at each of these marker loci, we determine, for each of the progeny, whether its genotype is LL or HL.

A genetic map specifying the relative locations of the markers may be known, or will be estimated using the data from the current experiment. Such a map gives the linear order of the markers on the various chromosomes. The distance between markers in a genetic map is given by genetic distance, in the units centiMorgans (cM). Two markers are separated by d cM, if d is the expected number of crossovers between the markers in 100 meiotic products.

Generally, we will write y_i for the phenotype (trait value) of individual i , and $x_{ij} = 1$ or 0 according to whether individual i has genotype HL or LL at the j th marker.

Typical experiments involve 100 to 1000 progeny, and use between 100 and 300 genetic markers.

1.3. Models.

1.3.1. *Model for recombination.* A diploid organism has two copies of each chromosome, one from its mother and one from its father. During the formation of gametes (sex cells), in the process of meiosis, the two homologous copies of a chromosome may undergo exchanges, called crossovers. Each of the gametes formed contains one copy of each chromosome, and each of these will be a mosaic of the two original homologs.

The locations of the crossovers along a chromosome are often modelled as a Poisson process (the assumption of “no interference”), with the processes in different individuals and on different chromosomes in one individual being independent. Moreover, at each locus, there is an equal chance that the allele is either paternally or maternally derived.

Consider a chromosome with k markers, and let $x_{ij} = 1$ or 0 if the i th individual has genotype HL or LL, respectively, at the j th marker. Then $\Pr(x_{ij} = 1) = \Pr(x_{ij} = 0)1/2$, for all i, j , and letting $x_j = (x_{ij})$, the x_j form a Markov chain.

Consider markers j_1 and j_2 , separated by a distance of d cM (so that d is the expected number of crossovers between these two markers in 100 meioses). If an odd number of exchanges occur between these markers, then $x_{ij_1} \neq x_{ij_2}$. This event is called a recombination. Let $r \Pr(x_{ij_1} \neq x_{ij_2})$. Then $r = \frac{1}{2}(1 - e^{-2d/100})$. This is called the Haldane map function (Haldane 1919).

1.3.2. *Model connecting genotype and phenotype.* Let y denote the phenotype for an individual derived from a backcross experiment. Let g be a vector giving its genotype at all loci. Let $\mu_g = E(y|g)$, the average phenotype for individuals

with genotype g , and $\sigma_g^2 = \text{Var}(y|g)$, the variance of the phenotypes of individuals with genotype g . In principle, these could be arbitrary functions of g . But imagine that there are a small number, p , of loci which affect the trait. Let (g_1, \dots, g_p) denote the genotypes of the individual at these loci. Then

$$\begin{aligned} E(y|g) &= \mu_{g_1 \dots g_p} \\ \text{and } \text{Var}(y|g) &= \sigma_{g_1 \dots g_p}^2. \end{aligned}$$

Often, we assume that the trait is homoscedastic - that the variance is constant within the genotype groups:

$$\text{Var}(y|g) = \sigma^2.$$

There are 2^p different possible genotypes at the p QTLs. Each genotype could have a distinct trait mean. But it is often assumed that the loci act additively. Let $z_j = 1$ or 0 , according to whether $g_j = \text{HL}$ or LL . We imagine that

$$E(y|g) = \mu + \sum_{j=1}^p \beta_j z_j.$$

Deviation from additivity (i.e. interactions between the QTLs) is called epistasis.

Most current methods use this assumption of additivity. Pairwise interactions are occasionally included, but few studies have found significant effects when using such an approach (Tanksley 1993), possibly because of the enormous number of pairwise interactions which must be considered. Strong evidence for epistasis has been demonstrated in one of the most studied quantitative traits, the number of abdominal bristles in *Drosophila* (Shrimpton and Robertson 1988; Long et al. 1995). Thus one should not discount the importance of epistasis.

It is important to note that additional cofactors, such as sex and treatment effects, may also be included in the above model, though we do not discuss that issue here.

A further often used assumption is that, given the genotypes at the QTLs, the trait y follows a normal distribution. Thus, if we group the backcross progeny according to their genotypes at the p QTLs, the phenotypes within each group will be normally distributed. The phenotypes for the backcross progeny, considered as a whole, will follow a mixture of normal distributions.

In this paper, we will focus on the case of strict additivity, with the further assumption of normality. This is not because we feel that it is the best approach, but rather because this simple case is still not well solved. In comparing the major approaches to identifying QTLs, the important differences will stand out most clearly if we avoid the added difficulties which accompany a search for epistasis.

1.4. Goals. Consider a backcross giving n progeny. For individual i , we obtain the phenotype, y_i , and the genotype at a set of M markers. Let $x_{ij} = 1$ or 0, according to whether individual i has genotype HL or LL at the j th marker.

We imagine that there are a set of p QTLs, and write $z_{ij} = 1$ or 0, according to whether individual i has genotype HL or LL at the j th QTL. Let

$$y_i = \mu + \sum_{j=1}^p \beta_j z_{ij} + \epsilon_i$$

where the ϵ_i are independent and identically distributed (iid) normal $(0, \sigma^2)$.

The ultimate goal is to estimate the number of QTLs, p , the locations of the QTLs, and their effects, β_j . In estimating the number and locations of the QTLs, we may make two errors: we may miss some of the QTLs, and we may include additional, extraneous loci.

In practice, a scientist may be satisfied with finding a few QTLs with large effect. In QTL experiments aimed at improving an agricultural crop, one seeks only the major QTLs, which may then be introgressed from one line into another. Furthermore, with a few major QTLs in hand, it may be possible to design experiments which identify the other QTLs segregating in a cross.

How one chooses to balance the two errors, of missing important loci and of including extraneous loci, depends on the goals of the scientists who designed the cross. In some cases, one may wish to find as many of the QTLs as possible and be undeterred by the possibility that several of the identified loci are, in fact, extraneous ones, of no effect. In other situations, one may be satisfied with identifying only a few major QTLs, in order to avoid including extraneous ones.

2. Major approaches. There are a large number of different methods for identifying the QTLs segregating in an experimental cross. In this section, we describe most of the proposed methods and discuss their advantages and disadvantages. It is best to distinguish between methods which model a single QTL at a time from those which attempt to model the effects of several QTLs at once. In Section 2.1, we review the single QTL methods, and in Section 2.2, we review the multiple QTL methods.

2.1. Single QTL methods. We will consider five basic single QTL methods: analysis of variance at a single marker, maximum likelihood using a single marker, interval mapping (i.e., maximum likelihood using flanking markers), an approximation to interval mapping called “regression mapping,” and a further method which gives results approximating interval mapping, called “marker regression.” Each of these methods includes a so-called “genome scan.” The loci

are considered one at a time, and a significance test for the presence of a single QTL is performed at each. Generally, the significance level used for the tests is adjusted to account for the multiple tests performed. Areas of the genome which give significant results are indicated to contain a QTL.

2.1.1. Analysis of variance. Analysis of variance (ANOVA) is the simplest method for identifying QTLs (see Soller et al. 1976). Consider a single marker locus, and group the progeny according to their genotypes at that marker. To test for the presence of a QTL, we look for differences between the mean phenotype for the different groups using ANOVA. If a QTL is tightly linked to the marker, then grouping the progeny according to their marker genotypes will be nearly the same as grouping them according to their (unknown) QTL genotypes, with recombinants being placed in the wrong groups.

Consider a backcross with a single segregating QTL. Suppose that the progeny with QTL genotype HL have mean phenotype μ_H , and that progeny with QTL genotype LL have mean phenotype μ_L , so the QTL has effect $\beta = \mu_H - \mu_L$. Consider a marker locus which is a recombination fraction r away from the QTL. Of the individuals with marker genotype HL, a fraction $(1 - r)$ of them have QTL genotype HL, while the remainder have QTL genotype LL, and so these individuals have mean phenotype $(1 - r)\mu_H + r\mu_L = \mu_H - \beta r$. The individuals with marker genotype LL have mean phenotype $(1 - r)\mu_L + r\mu_H = \mu_L + \beta r$. Thus the mean difference between the two marker genotype groups is $(\mu_H - \beta r) - (\mu_L + \beta r)\beta(1 - 2r)$. And so a non-zero mean difference between the marker genotype groups indicates linkage between the marker and a QTL.

There are two drawbacks to this method. First, we do not receive separate estimates of the location of the QTL relative to the marker (r) and its effect (β). QTL location is indicated only by looking at which markers give the greatest differences between genotype group means. Second, when the markers are widely spaced, the QTL may be quite far from all markers, and so the power for detection will decrease, since the difference between marker genotype means decreases linearly as the recombination fraction between the marker and the QTL increases.

2.1.2. Maximum likelihood with a single marker. To get around the problems with ANOVA, several authors have proposed to explicitly model the location of the QTL with respect to the marker, and then use maximum likelihood (ML), or an approximation to ML, to estimate the distance between the marker and the QTL as well as the QTL's effect (Weller 1986, 1987; Simpson 1989). This method makes use of the fact that the marker genotype groups are not quite the same as the QTL genotype groups.

Consider again the backcross discussed in the previous section. Suppose that the individuals who are HL at the QTL have phenotypes which are normal

(μ_H, σ^2) , and the individuals who are LL at the QTL have phenotypes which are normal (μ_L, σ^2) . Then at a marker which is a recombination fraction r away from the QTL, the phenotype distribution for individuals who are HL is a mixture of two normals, with density

$$f_1(y; \mu_H, \mu_L, \sigma, r) = (1 - r)\phi\left(\frac{y - \mu_H}{\sigma}\right) + r\phi\left(\frac{y - \mu_L}{\sigma}\right),$$

where ϕ is the density of the standard normal distribution. The phenotype distribution for individuals who are LL at the marker has density

$$f_2(y; \mu_H, \mu_L, \sigma, r) = (1 - r)\phi\left(\frac{y - \mu_L}{\sigma}\right) + r\phi\left(\frac{y - \mu_H}{\sigma}\right).$$

Let $x_i = 1$ or 0, according to whether individual i has marker genotype HL or LL. Let y_i denote the phenotype for individual i . Then the likelihood under this model, letting θ denote the vector of parameters (μ_H, μ_L, σ) , is

$$L(\theta, r; y, x) = \prod_i [f_1(y_i; \theta, r)]^{x_i} [f_2(y_i; \theta, r)]^{1-x_i}$$

Maximizing this function over θ , using, for example, the EM algorithm (Dempster et al. 1977), gives the maximum likelihood estimates. This is done for a particular value of the recombination fraction r . We then maximize the likelihood over r to obtain \hat{r} .

Linkage between the marker and the QTL is tested by performing a likelihood ratio test, comparing the above model, with a single QTL linked to the marker, to the null hypothesis of no segregating QTLs, where the individuals are assumed to have phenotypes which are normal (μ, σ^2) .

The likelihood under the null hypothesis, letting $\theta_0 = (\mu, \sigma)$ is

$$L_0(\theta_0; y) = \prod_i \phi\left(\frac{y_i - \mu}{\sigma}\right).$$

The likelihood ratio test is performed by calculating the likelihood ratio, or, as seems to be preferred by geneticists, the LOD score, which is the log (base 10) likelihood ratio

$$\text{LOD}(r) = \log_{10} \left[\frac{\max_\theta L(\theta, r; y, x)}{\max_{\theta_0} L_0(\theta_0; y)} \right]$$

and comparing it to the distribution of the maximum LOD score under the null hypothesis (that is, under the assumption that no QTLs are segregating).

This method has the advantage of giving separate estimates of the QTL's location with respect to a marker and its effect. One disadvantage is the great increase in computation associated with maximizing the likelihood function to obtain parameter estimates. But a bigger problem involves combining the information for different markers to give a single estimate of the QTL location; it is not at all clear how this can be done.

2.1.3. Interval mapping. Lander and Botstein (1989) improved on the previous single marker approaches by considering flanking markers. Their method has been called "interval mapping," and is currently one of the most commonly used methods for identifying QTLs in experimental crosses. (Note that Mather and Jinks (1977) proposed a similar approach, using the method of moments with flanking markers.)

Again, they assume that there is a single segregating QTL, and that back-cross individuals have phenotypes which are normally distributed with mean μ_H or μ_L , according to whether their QTL genotype is HL or LL, and common variance σ^2 . Further, they use the assumption of no crossover interference, and require a genetic map specifying the locations of the markers.

Consider two markers which are separated by d cM, corresponding to a recombination fraction of $r = \frac{1}{2}(1 - e^{-2d/100})$, and a putative QTL located d_L cM from the left marker, corresponding to a recombination fraction of $r_L = \frac{1}{2}(1 - e^{-2d_L/100})$. The recombination fraction between the QTL and the right marker is then $r_R = (r - r_L)/(1 - 2r_L)$. There are four possible sets of genotypes at the two marker loci; for each, we can calculate the conditional probability for each of the two QTL genotypes, given the marker genotypes. These are displayed in Table 1. Note that, with fully informative markers, the flanking markers provide all of the information about the QTL genotypes.

TABLE 1
Conditional probabilities for the QTL genotype given the genotypes at two flanking markers

marker genotype		QTL genotype	
left	right	HL	LL
HL	HL	$(1 - r_L)(1 - r_R)/(1 - r)$	$r_L r_R/(1 - r)$
HL	LL	$(1 - r_L)r_R/r$	$r_L(1 - r_R)/r$
LL	HL	$r_L(1 - r_R)/r$	$(1 - r_L)r_R/r$
LL	LL	$r_L r_R/(1 - r)$	$(1 - r_L)(1 - r_R)/(1 - r)$

For each of the four sets of marker genotypes, we can now write down the conditional phenotype density, which has the form of a mixture of two normal distributions, similar to those seen in Section 2.1.2. Thus we can obtain the

likelihood for our four parameters, $(\mu_H, \mu_L, \sigma, r_L)$.

Lander and Botstein (1989) proposed to maximize this likelihood, for fixed r_L , using the EM algorithm (Dempster et al. 1977). They then calculated the LOD score, which is the log (base 10) likelihood ratio comparing the hypothesis of a single QTL at the current locus (i.e., the current value of r_L) to the null hypothesis of no segregating QTLs (meaning that the individuals' phenotypes follow a normal (μ, σ^2) distribution). The two likelihoods in this ratio must be maximized over their respective parameters.

The procedure outlined above is performed for each locus in the genome. The likelihood under the null hypothesis is calculated just once. The likelihood for the hypothesis of a single QTL must be calculated at each locus in the genome (or, really, just every 1 cM or so), and so the EM algorithm must be performed at each locus.

The LOD score is then plotted against genome location, and is compared to a genome-wide threshold. Whenever the LOD curve exceeds the threshold, we infer the presence of a QTL. The point at which the LOD is maximized is used as the estimate of the QTL location. A one- or two-LOD support interval, the region around the inferred QTL in which the LOD score is within one or two of its maximum, is used as an interval estimate for QTL location.

The genome-wide threshold, used to indicate the significance of a peak in the LOD curve, is obtained by finding the 95th percentile of the maximum LOD score, across the entire genome, under the null hypothesis of no segregating QTLs.

Figure 1 gives an example of a LOD curve calculated using interval mapping. We simulated 200 backcross progeny, having a single chromosome of length 100 cM with 11 equally spaced markers, using a model with a single QTL located 35 cM from the left of the chromosome. The effect of the QTL (the difference between the means for HL versus LL individuals) was 0.75σ , giving a heritability, the proportion of the total phenotypic variance due to the QTL, of 0.36. The dots plotted on the curve point out the locations of the marker loci. Using a LOD threshold of 2.5, the observed peak is significant. The inferred QTL is estimated to be at 37 cM, with a maximum LOD score of 3.4. The one-LOD support interval covers the region from 27 cM to 47 cM, which does indeed include the actual location of the simulated QTL.

A great deal of effort has been expended in trying to understand the appropriate LOD threshold to use. Lander and Botstein (1989) performed simulations to estimate the threshold for various different genome sizes and marker densities. They gave analytical calculations for the case of a very dense marker map. These guidelines should suffice for most uses. If one is concerned, additional simulations, conforming to the particular case under study, can be performed

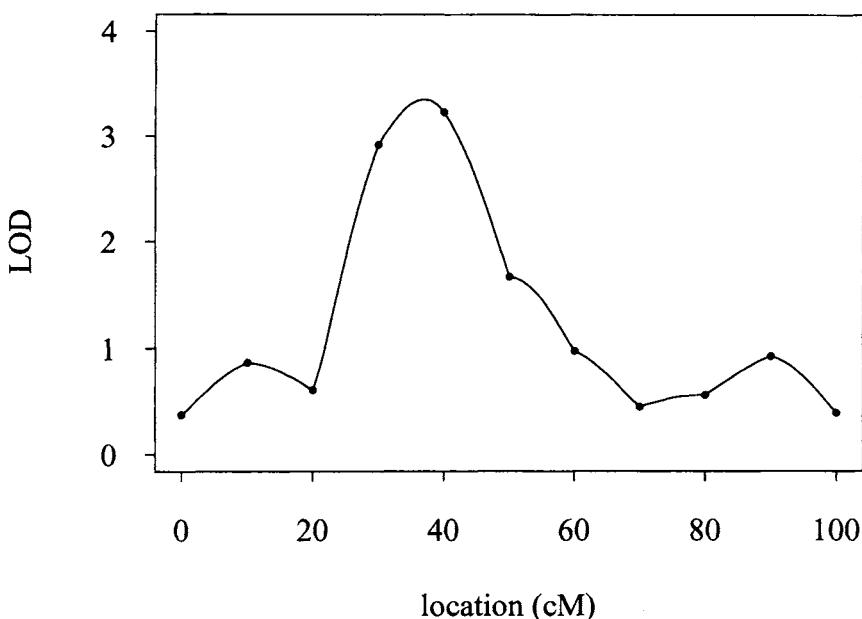


FIG. 1. *An example LOD curve, calculated using interval mapping, for some simulated data.*

quite easily, or one can use a permutation test (Churchill and Doerge 1994), which has the advantage of avoiding the assumption of normally distributed environmental variation.

A number of studies have assessed the performance of interval mapping in comparison to ANOVA (van Ooijen 1992; Knott and Haley 1992; Darvasi et al. 1993; Rebaï et al. 1995; Hyne et al. 1995). The chief benefit of interval mapping is that it gives more precise estimates of the location and effect of a QTL. It does not give an appreciable increase in the power for detecting QTLs, and it requires a great deal more computational effort than does single marker ANOVA.

Hyne et al. (1995) stated that when a QTL is located very near one end of a linkage group, its estimated location, as given by interval mapping, will be biased, since if one looks for QTLs only within the two extreme markers on the linkage group, its estimated location will never be outside of the last marker. It is possible to extend the LOD curves beyond the most extreme markers, however; outside of these markers, we can use the single marker maximum likelihood method, described in the previous section. Doing this should eliminate the bias problem. (Of course, a slight increase in variance, and a slight decrease in power, will accompany this approach.)

Look again at Figure 1. The dots on the LOD curve are at the marker loci. At these points, interval mapping is really just ANOVA, since the genotypes

there are known exactly. If we performed only ANOVA, we would get exactly those points on the LOD curve. Interval mapping links these points together, and indicates that the best estimate for the QTL position is at 37 cM. But the markers at both 30 and 40 cM are within the one-LOD support interval.

2.1.4. Regression mapping. Knapp et al. (1990), Haley and Knott (1992), and Martínez and Curnow (1992) independently developed a method which approximates interval mapping very well, but requires much less computation. The method has come to be called “regression mapping.” The presentation in Haley and Knott (1992) is by far the best.

Consider again the model of the previous section, with two markers separated by a recombination fraction r , and a putative QTL located between them, at a recombination fraction r_L from the left marker. The conditional expected value of the phenotype for an individual, given its genotypes at the flanking markers, is

$$E(y|\text{marker gen.}) = \mu_L + (\mu_H - \mu_L) \Pr(\text{QTL gen. is HL}|\text{marker gen.}),$$

where $\Pr(\text{QTL gen. is HL}|\text{marker gen.})$ is as shown in Table 1.

In regression mapping, we regress the individuals’ phenotypes on their conditional probabilities for having the genotype HL at the putative QTL, given their marker genotypes. The log likelihood is calculated assuming that

$$y|\text{marker gen.} \sim \text{normal}(\tilde{y}, \sigma^2)$$

where $\tilde{y} = E(y|\text{marker gen.})$. This gives the LOD score

$$\text{LOD} = \frac{n}{2} \log_{10} \left(\frac{\text{RSS}_0}{\text{RSS}} \right)$$

where n is the number of progeny, RSS is the residual sum of squares from the above regression, $\sum_i (y_i - \hat{y}_i)^2$, and RSS_0 is the residual sum of squares under the null hypothesis of no segregating QTLs, $\sum_i (y_i - \bar{y})^2$.

Like interval mapping, the LOD score is calculated at each locus in the genome, but here, we need only calculate a single regression at each locus, rather than perform the EM algorithm at each locus, which requires a number of iterations, each containing a regression. Thus, there is a great savings in computation time. Also, because regression mapping requires only simple regression calculations, it is much easier to include additional effects into the analysis, such as sex or treatment effects. This may translate into large increases in performance.

Figure 2 displays the difference between the LOD curves calculated by regression mapping and interval mapping, for the data used in the previous section. The difference between the two curves is very subtle, being less than 0.1 in

absolute value. Regression mapping gives results every bit as good as interval mapping, with a great deal less computation.

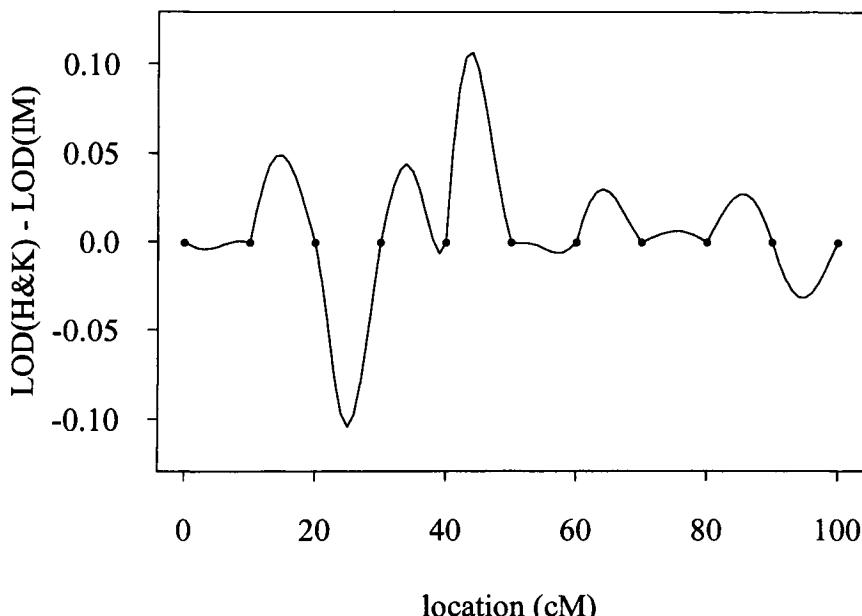


FIG. 2. *The difference between the LOD curves calculated using regression mapping and interval mapping for some simulated data.*

2.1.5. *Marker regression.* Kearsey and Hyne (1994) and Wu and Li (1994) independently developed a further method, which seems to approximate interval mapping quite well, with less intensive computation. But this method, which Kearsey and Hyne call “marker regression,” seems more awkward and less adaptable than Haley and Knott’s “regression mapping,” and has not been shown to provide any further benefits.

Consider a linkage group with M markers, and fix the location for a putative QTL. Let r_j be the recombination fraction between the QTL and the j th marker. Group the individuals according to whether they have genotype HL or LL at marker j . Let $\hat{\beta}_j$ be the difference between the phenotype means for these two groups. As shown in Section 2.1.1,

$$E(\hat{\beta}_j) = \beta(1 - 2r_j),$$

where $\beta = \mu_H - \mu_L$, the effect of the QTL.

Kearsey and Hyne (1994) suggest regressing the $\hat{\beta}_j$ for the M markers on the values $(1 - 2r_j)$, without an intercept. This is performed for each locus on the

linkage group; we seek the locus giving the minimum residual sum of squares in this regression.

Wu and Li (1994) point out that the $\hat{\beta}_j$ do not have constant variance. The variance of $\hat{\beta}_j$ is approximately $4[\sigma^2 + r_j(1 - r_j)\beta^2]/n$, where n is the number of progeny, and σ^2 is the environmental variance. They suggest using weighted least squares, using weights inversely proportional to the variances of the $\hat{\beta}_j$. But since σ and β are not known, it is not clear how to do this, unless one were to use a form of iteratively re-weighted least squares.

Wu and Li (1996) further point out that the $\hat{\beta}_j$ are correlated, and recommend using general least squares using an estimate of the covariance matrix.

We applied the method of Kearsey and Hyne (1994) to the simulated data analyzed in Sections 2.1.3 and 2.1.4. Figure 3 displays the residual sum of squares curve. The minimum is realized at 42 cM.

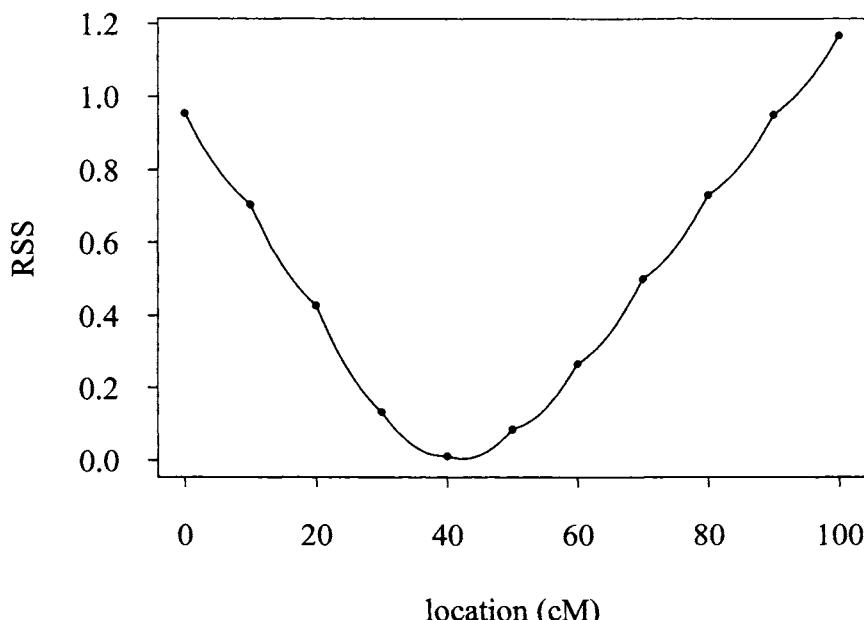


FIG. 3. *The residual sum of squares curve calculated using the marker regression method for some simulated data.*

Kearsey and Hyne (1994) gave a small number of simulations which suggested that marker regression performs as well as interval mapping. But they have not made a case for real improvements, aside from ease of computation. The method seems to have no advantages over regression mapping.

2.2. Multiple QTL methods. Recent efforts in developing methods to identify QTLs have focused on multiple QTL methods. There are three principal reasons for modelling multiple QTLs: to increase sensitivity, to separate linked QTLs, and to estimate epistatic effects (i.e., interactions between alleles at different QTLs).

When several QTLs are modelled, one can control for much of the genetic variation in a cross, and thus individual QTLs can be more clearly seen. In contrast, when one models a single QTL at a time, the genetic variation due to other segregating QTLs is incorporated into the “environmental” variation. When two QTLs are linked, single QTL methods, such as interval mapping, often view them as a single QTL. Searches which allow multiple QTLs do a better job of separating the two loci, and identifying them as distinct. The presence of epistasis can only be detected and estimated using models which include multiple QTLs. Incorporating epistatic effects into multiple QTL models will be very difficult, however. If one were to include all possible pairwise interactions, the number of parameters in the model would quickly explode. The methods discussed here all neglect the possibility of epistasis.

In this section, we discuss four important methods which explicitly consider multiple QTLs: multiple regression, interval mapping type methods using either forward selection or multi-dimensional searches, composite interval mapping (also called MQM mapping), and Markov chain Monte Carlo using a full Bayesian model.

2.2.1. Multiple regression. The obvious extension of analysis of variance is multiple regression. We attempt to form a model which includes a number of different marker loci, rather than looking at the markers one at a time. Let M be the number of markers, let $x_{ij} = 1$ or 0 , according to whether individual i had genotype HL or LL at the j th marker, and let y_i be the phenotype for individual i . We write

$$\mathbb{E}(y_i|x_i) = \mu + \sum_{j=1}^M \beta_j x_{ij}$$

where $x_i = (x_{ij})$. We presume that most of the markers have $\beta_j 0$. We seek the set of markers, S , with non-zero coefficients, β_j , so that

$$\mathbb{E}(y_i|x_i) = \mu + \sum_{j \in S} \beta_j x_{ij}.$$

The markers in S are indicated to be near QTLs.

There are two problems associated with this method. First, we must find a way to search through the set of possible models, in order to seek good ones. In

an experiment with 100 genetic markers, there are $2^{100} \approx 10^{30}$ possible models to consider; it will be impossible to look at each of them. Second, we must form a criterion for choosing from these models. For models that include the same number of markers, one generally picks the one with the smallest residual sum of squares. The difficulty is in choosing between models of different sizes: what change in the residual sum of squares must we see before we will accept an additional marker into the model?

Cowen (1989) discussed using stepwise selection and backward deletion, and using Mallows' C_p and the adjusted- R^2 criteria, when using multiple regression to identify QTLs. More recently, Doerge and Churchill (1996) described using forward selection, with permutation tests to determine the appropriate size of the model. Wright and Mowers (1994) and Whittaker et al. (1996) described the relationship between the partial regression coefficients, obtained by regressing a trait on a set of marker loci, and the locations and effects of a set of QTLs, but they did not provide a procedure for using this information to identify QTLs.

Broman (1997) discussed the use of model selection procedures in regression to identify QTLs. A number of different methods of searching through the space of models were compared: forward selection, backward selection, and Markov chain Monte Carlo (MCMC). Forward selection was found to perform as well as the other search methods, while it requires much less extensive calculations. Further discussion focussed on the criteria for choosing a model. The usual approaches to model selection focus on minimizing prediction error, and, as a result, standard criteria for choosing models, such as Mallows' C_p and adjusted- R^2 , tend to choose models with a large number of extraneous variables. With some modification, the Schwartz's BIC (Schwartz 1978) performs much better. This criterion has the form $\text{BIC}-\delta = \log \text{RSS} + q\delta \log n/n$, where RSS is the residual sum of squares for the model, q is the number of markers in the model, and n is the number of progeny. With this method, one attempts to find the model which minimizes the above criterion. The parameter δ must be chosen to balance the error of missing important QTLs with the error of including too many extraneous markers; a value between 2 and 3 may be appropriate in many situations.

2.2.2. Interval mapping revisited. Lander and Botstein (1989) briefly mentioned a method for distinguishing linked loci. If, when performing interval mapping, the LOD curve for a linkage group shows two peaks, or a single very broad peak, Lander and Botstein recommended to fix the position of one QTL at the location of the maximum LOD, and then search for a second QTL on that linkage group. In the model selection literature, this method is generally called forward selection (Miller 1990). Though some authors (Haley and Knott 1992; Satagopan et al. 1996) have interpreted this method as applying interval

mapping to the residuals from the best fit of one QTL, it is best to estimate the effects of both QTLs simultaneously, using the original data (cf Dupuis et al. 1995).

We fix the location of the first QTL, and vary the location of the second QTL along the linkage group. At each location for the second QTL, we calculate a LOD score, comparing the maximum likelihood under the hypothesis of two QTLs at these locations, to that with a single QTL, located where the first QTL was placed. Each individual's contribution to the likelihood has the form of a mixture of four normal distributions, the four components corresponding to the four possible QTL genotypes. The EM algorithm can again be used to obtain the maximum likelihood estimates and the corresponding LOD score. (One could also apply the "regression mapping" method.)

Several authors have criticized this method (Haley and Knott 1992; Martínez and Curnow 1992), pointing to the phenomenon of "ghost QTLs." When two or more QTLs are linked in coupling (meaning that their effects have the same sign), interval mapping often gives a maximum LOD score at a location in between the two QTLs.

Consider, for example, a 60 cM segment of a chromosome, with four equally spaced markers (20 cM spacing). Consider a backcross with QTLs located at 15 and 45 cM, acting additively and having equal additive effect 0.5σ . The solid line in Figure 4 gives the expected LOD (ELOD) curve for this situation, when using 200 progeny. (Since there is no closed-form expression for the ELOD curve, it was estimated by performing 1000 simulations of the above situation and averaging the LOD curves obtained. We also used the fact that the ELOD curve is symmetric about the 30 cM point, and so averaged the pairs of points on the curve which are symmetric about 30 cM.) Note that the ELOD curve is maximized at 30 cM, even though the simulated QTLs were at 15 and 45 cM. This gives rise to the term "ghost QTL." Forward selection here would give bad results. We would generally pinpoint the first QTL at around 30 cM, and then search for a second QTL, and so would be completely mistaken.

But this "ghost QTL" problem turns out to be an artifact of interval mapping. The dashed and dotted lines in Figure 4 are the ELOD curves for the above example, using marker spacings of 10 and 5 cM, respectively. When the markers are more tightly spaced, the ghost QTL disappears. The ELOD curves are not maximized exactly at the true QTL locations, but things do get better as marker density increases. Note that if one considered only the marker loci, one would not be so misled. The marker loci at which the LOD is maximized are those closest to the true QTLs. Similar observations were made by Whittaker et al. (1996) and Wright and Kong (1997).

As an alternative to forward selection, several authors have recommended

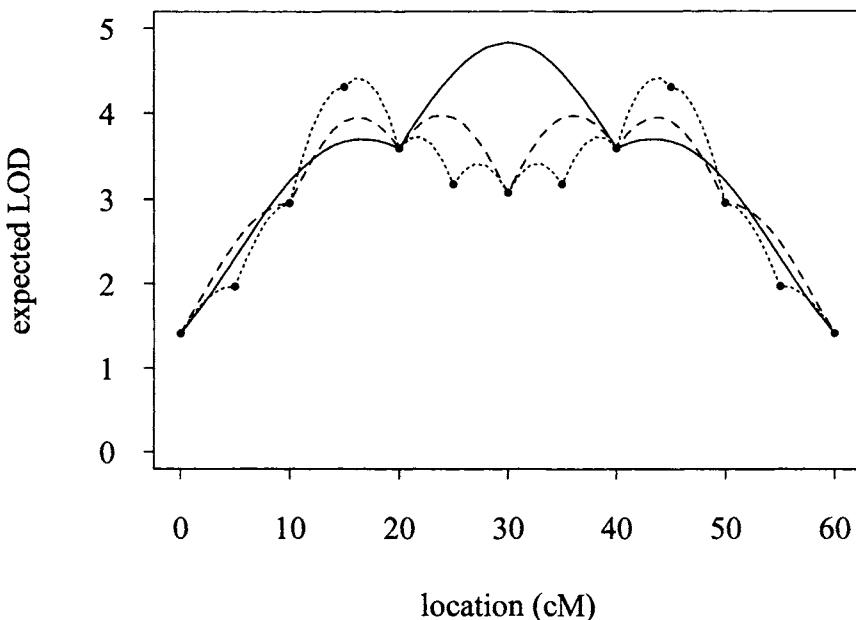


FIG. 4. *Expected LOD curves, with two QTLs located at 15 and 45 cM. The solid line, dashed line and dotted line correspond to using equally-spaced markers at spacings of 20, 10 and 5 cM, respectively.*

performing a full two-dimensional search for QTLs (Haley and Knott 1992; Martínez and Curnow 1992; Hyne and Kearsey 1995; Whittaker et al. 1996; Wu and Li 1994, 1996). Instead of fixing the location of one QTL and then searching for an additional one, the locations of both QTLs are allowed to vary simultaneously. A great deal more computation must be performed. Extending this method to more than two QTLs, as recommended by Wu and Li (1996), is possible in principle, but the computation requirements would very quickly become prohibitive.

One problem that these authors have not discussed carefully is the question of when to add an additional QTL: how large an increase in LOD should we require before allowing an additional QTL? Such guidelines are necessary, if one is to use these methods in practice.

2.2.3. *Composite interval mapping and MQM mapping.* Jansen and Zeng independently developed a method which attempts to reduce the multi-dimensional search for identifying multiple QTLs to a one-dimensional search (Jansen 1993; Jansen and Stam 1994; Zeng 1993, 1994). This is done using a hybrid of interval mapping and multiple regression on marker genotypes. By including other markers (on the same chromosome and on different chromosomes) as regressors while doing interval mapping, one hopes to control for the effects

of QTLs in other intervals, so that there will be greater power in detecting a QTL, and so that the effects of the QTLs will be estimated more precisely. Jansen called the method MQM mapping (short for “marker-QTL-marker” or “multiple QTL models”); Zeng called it composite interval mapping.

The method is performed as follows. We choose a set of markers, S , to control for background genetic variation. Then, we perform a genome scan, as in interval mapping. At each locus in the genome, we hypothesize the presence of a QTL, and we write

$$y \sim \text{normal}(\mu + \beta z + \sum_{j \in S^*} \beta_j x_j, \sigma^2),$$

where y is the phenotype, $z = 1$ or 0 , according to whether the genotype at the putative QTL is HL or LL, $x_j = 1$ or 0 , according to whether the genotype at the j th marker is HL or LL, and S^* is a subset of our set of markers, S , where we exclude any markers that are within, say, 10 cM of the putative QTL. Under this model, the contribution of each individual to the likelihood has the form of a mixture of two normal distributions with means $\mu + \beta + \sum_{j \in S^*} \beta_j x_j$ and $\mu + \sum_{j \in S^*} \beta_j x_j$, with mixing proportions equal to the conditional probabilities of the individual having QTL genotype HL and LL, given its marker genotypes. The EM algorithm, or a variant called the ECM algorithm (Meng and Rubin 1993), can be used to maximize the likelihood function.

As in interval mapping, at each locus, a likelihood ratio or LOD score is calculated, comparing the likelihood assuming that there is a QTL at that locus, to the likelihood assuming that there is not a QTL there, in which case we imagine that all progeny have phenotypes which are normally distributed with mean $\mu + \sum_{j \in S^*} \beta_j x_j$ and variance σ^2 . The LOD score is plotted as a function of genome position, and is compared to a genome-wide threshold. As in interval mapping, areas of the genome for which the LOD curve exceeds the threshold are said to contain a QTL.

The genome-wide threshold is obtained by considering the distribution of the maximum LOD score under the hypothesis of no segregating QTLs anywhere in the genome. This distribution should take into account the selection of the set of marker regressors, S . The distribution can be estimated by simulating a set of data under the hypothesis of no segregating QTLs, performing the entire procedure, and calculating the maximum LOD curve obtained, and then repeating the process a number of times. The 95th percentile of these maximum LOD scores is used as the threshold.

The key problem in this method is the choice of which markers to use as regressors: using too many markers will increase the variance of the LOD score, and thus will decrease the power for detecting QTLs. Jansen (1993) and Jansen

and Stam (1994) used backward deletion, with Akaike's Information Criterion (AIC) (Akaike 1969) or a slight variant, to pick the subset of markers. Zeng (1994) recommended using either all markers, dropping those within 10 cM of the putative QTL, or using all markers that are not linked to the putative QTL. Basten et al. (1996), in a manual for the program QTL Cartographer, recommended using forward selection up to a fixed number of markers, say five, and then dropping any markers that are within 10 cM of the putative QTL.

We have found that the methods that Zeng (1994) originally recommended, using all markers or all markers not linked to the putative QTL, work very badly. Including so many markers increases the corresponding LOD threshold to such a large value that power is reduced to almost zero. Only QTLs with extremely large effect will be found by this method.

The performance of the other methods for choosing the set of marker regressors depends on how many markers are chosen. And once we have found a way to choose this set, the task of identifying QTLs is essentially done: the best set of markers to use is exactly the set of markers which are closest to the underlying QTLs. In Section 3, we present some simulation studies which assess the performance of these methods.

2.2.4. Markov chain Monte Carlo. Satagopan et al. (1996) have applied the Markov chain Monte Carlo (MCMC) method to the problem of identifying QTLs. MCMC is a very popular approach to solving very complex statistical problems, especially those which include a large amount of missing information. Gelman et al. (1995) gives a very good introduction to the subject.

Consider again a backcross. Satagopan et al. (1996) consider a single linkage group. (The method can be extended to several linkage groups in a straightforward way.) Consider n progeny. Let y_i be the phenotype for individual i . Suppose there are M markers, at locations $D = (D_1, D_2, \dots, D_M)$, in cM, from the left end of the linkage group. Let $x_{ij} = 1$ or 0, according to whether individual i has genotype HL or LL at the j th marker.

Let S be the number of segregating QTLs, and let $\lambda(\lambda_1, \dots, \lambda_S)$ be their locations, in cM, from the left end of the linkage group. Let $z_{ij} = 1$ or 0, according to whether individual i has genotype HL or LL at the j th QTL. Let β_j be the effect of the j th QTL, and assume that the environmental variation is normally distributed, with variance σ^2 . Let μ be the mean of individuals for whom $z_{ij} = 0$ for all j .

As shorthand, we will write $y = (y_1, \dots, y_n)$, $x_i = (x_{i1}, \dots, x_{iM})$, $x = (x_1, \dots, x_n)$, and similarly for z_i , z and β . Also, let $\theta = (\mu, \beta, \sigma)$.

We have

$$y_i|z_i, \theta \sim \text{normal}(\mu + \sum_{j=1}^S \beta_j z_{ij}, \sigma^2).$$

This gives the likelihood

$$L(\lambda, \theta|y, x, D) = \prod_{i=1}^n \sum_q f(y_i|z_i = q, \theta) \Pr(z_i = q|\lambda, x_i, D)$$

where the sum over q is over the 2^S possible QTL genotypes for individual i and where f is the conditional (normal) density for y .

Satagopan et al. (1996) use a full Bayesian framework, in that they assign a prior probability distribution to the unknown parameters (λ, θ) , say $p(\lambda, \theta)$, and then look at their posterior distribution, given the data, $p(\lambda, \theta|y, x, D)$.

The goal of the MCMC method is thus to estimate the posterior distribution of the unknown parameters. This is done by creating a Markov chain whose stationary distribution is the desired posterior distribution.

Simulating from this chain gives a sequence $(\lambda_0, \theta_0), (\lambda_1, \theta_1), \dots, (\lambda_N, \theta_N)$. Estimates of the desired parameters, such as the QTL effects, β_j , are obtained by averaging over these samples. Interval estimates for the QTL locations can be obtained by looking at the smallest intervals which contain, say, 95% of the samples.

In order to determine the number of QTLs, S , Satagopan et al. (1996) run separate chains for different values of S , and use Bayes factors. In brief, for each value of S , they use their samples to estimate the probability of the data given the model, $p(y, x|S)$. They estimate the number of QTLs to be the value of S for which this estimated probability is large. If one were willing to give a prior on the number of QTLs, say $\Pr(S = s)$, the posterior distribution for S could be calculated

$$\Pr(S = s|y, x) = \frac{p(y, x|S = s) \Pr(S = s)}{\sum_s p(y, x|S = s) \Pr(S = s)}$$

The estimated number of QTLs would then simply be the value of S with the largest posterior probability.

A later report (Satagopan and Yandell 1996), using an idea developed by Green (1995), describes how to allow the unknown number of QTLs, S , to be included as an unknown parameter, so that a single Markov chain can be used to estimate S along with the other parameters. Doing this requires placing a prior distribution on the number of QTLs.

We have skipped all of the details of the MCMC method. The difficulties in applying this approach are entirely in those details. First, you need to create a Markov chain which has your posterior distribution as its stationary distribution. There are a number of standard ways to do this, such as the Gibbs sampler (Geman and Geman 1984) and the Metropolis-Hastings algorithm (Hastings 1970). The most important characteristic in the chain is that it mixes well: that it moves around the parameter space rather easily, and that it very quickly reaches its stationary distribution. Forming good Markov chains, and monitoring their behavior, is a delicate and sophisticated art.

The other important problem is in the determination of the number of QTLs. Whether we assign a prior to the unknown number of QTLs or use Bayes factors, we must make choices which balance the problem of missing real QTLs with that of including extraneous loci.

3. Simulations. In this section, we present the results of a small simulation study aimed at comparing several different methods for identifying QTLs. Our focus is on identifying QTLs, and so we look only at whether the methods detect the simulated QTLs, and not at the estimated effects and the precision with which the location is estimated. Simulations are necessary, because the methods for identifying QTLs are too complex to be assessed by analytical means, at least in the situations in which they would be used in practice.

Most authors have used simulations to demonstrate their methods for finding QTLs. Many have presented the results of applying their method to a single data set (Jansen 1993; Knapp 1991; Lander and Botstein 1989; Zeng 1994), a practice which precludes a true assessment of the method's performance. Others consider only very simple situations, such as simulating only one or two chromosomes with one or two segregating QTLs (Haley and Knott 1992; Kearsey and Hyne 1994). In practice, most QTL studies involve a search over ten or more chromosomes, and very often there is evidence for at least a moderate number of segregating QTLs (from three or four to as many as a dozen). A method's ability to detect QTLs in simulation studies which use very limited searches and in which only a small number of QTLs are allowed will say little about its performance in the more complex situations where the method is anticipated to be used.

Also missing from the literature is a careful comparison of the performance of the many methods available for identifying QTLs. It is surprising that such comparisons are not a routine part of the presentation of a new method. Before dropping a simple approach in favor of a more complex one, we should have evidence that the complexities of the new approach will be accompanied by a real improvement in performance.

We compared four different methods for identifying QTLs: analysis of variance (ANOVA) at the marker loci, the method of Zeng (1994), forward selection using a BIC-type criterion, and forward selection using a permutation test at each stage (Doerge and Churchill 1994). These methods were described in Section 2.

Interval mapping (IM) was ignored, because it provides no improvement over simple ANOVA when using a relatively dense marker map (10 cM spacing or less) and a small or moderate number of progeny (500 or less), at least when it comes to identifying QTLs. This can easily be seen when inspecting the one- or two-LOD support intervals which accompany any application of IM: they invariably span several markers. The benefit of IM is in providing more precise estimates of QTL location and effects.

For Zeng's method, we used forward selection up to either 3, 5, 7 or 9 markers to obtain the set of regressors, and limited the search for QTLs to marker loci. With ANOVA and Zeng's method, we obtained genome-wide thresholds by performing 1000 simulations under the hypothesis of no segregating QTLs: the estimated threshold was the 95th percentile of the maximum LOD score across all markers. In addition, for these two methods, we required that the LOD dropped by at least 2.2 in base 10 (corresponding to 5 in base e) between "peaks" before we declared that two QTLs were identified. This value was obtained empirically (in other words, by trial and error).

The BIC-type criterion used is $\log \text{RSS} + \delta q \log n / n$, where RSS is the residual sum of squares, n is the number of progeny, q is the number of markers in the model, and δ is either 2, 2.5 or 3. We use BIC-2, BIC-2.5 and BIC-3 to identify these criteria. For the permutation method, at each stage we used the 95th percentile of 500 permutations to determine whether to add another marker.

In the study described in this section, we simulated 250 backcross progeny, obtained from inbred lines, with nine chromosomes, each of length 100 cM and having 11 equally spaced markers per chromosome (thus at a 10 cM spacing). The recombination process was assumed to exhibit no interference. The environmental variation followed a normal distribution with standard deviation $\sigma = 1$.

We modelled three QTLs with equal additive effect 0.5. One QTL was located at the center of chromosome 1, and two QTLs were located on chromosome 2 at 30 and 70 cM. The linked QTLs were either in coupling (effects of equal sign) or repulsion (effects of opposite sign). The QTLs were assumed to act additively. The heritability for these models (defined as the ratio of the genetic variance to the total phenotypic variance) was 0.20 and 0.12 when the linked QTLs were in coupling and repulsion, respectively. Note that all QTLs were located exactly at marker loci.

For each QTL model we performed 1000 simulations. The result of the application of each method was a set of marker loci indicated to be at or near QTLs. In assessing the results, we defined a chosen marker to be correctly identifying a QTL if it was within 20 cM of a QTL; otherwise it was deemed incorrect. If more than one chosen marker were within 20 cM of the same QTL, one was called correct and the others were called incorrect.

The estimated genome-wide LOD (base 10) thresholds for ANOVA and Zeng's method (using forward selection up to 3, 5, 7 and 9 markers) are displayed in Table 2. The estimated standard errors for the thresholds, obtained using a bootstrap (Venables and Ripley 1994), are approximately 0.1. For ANOVA, the threshold corresponded closely to the threshold in Figure 4 of Lander and Botstein (1989). For Zeng's method, the threshold increased with the number of regressors used.

TABLE 2

Estimated genome-wide LOD thresholds for a backcross with 250 progeny and nine 100 cM chromosomes each containing 11 equally-spaced markers

ANOVA	Zeng			
	3	5	7	9
2.5	3.3	3.6	3.8	4.0

In Table 3, we display the joint distribution, across the 1000 simulations, of the numbers of correctly and incorrectly chosen markers for the case of three QTLs with two QTLs linked in coupling, and using 250 progeny. The four columns labelled "Zeng" correspond to Zeng's method using forward selection up to either 3, 5, 7 or 9 markers. The three columns labelled "BIC" correspond to forward selection using the BIC-2, BIC-2.5 and BIC-3 criteria. The column "permu" gives the results for using forward selection with a permutation test at each stage. The second-to-last row in the table includes all simulations with two or more incorrectly chosen markers. The last row in the table gives the number of simulations in which at least one incorrect marker was chosen.

ANOVA nearly always found at least one QTL, and often found two, but it had difficulty in separating the two linked QTLs. ANOVA added incorrect markers about 7% of the time. Zeng's method did worse than ANOVA in this situation. It suffered from low power for detection, and the power decreased sharply as the number of markers used as regressors increased; using three markers as regressors worked best in this case. Forward selection using BIC-2 did a better job of detecting the QTLs, but included incorrect markers 11% of the time - much more often than the other methods. The use of a larger

TABLE 3

Distribution of the numbers of correctly and incorrectly chosen markers in 1000 simulations of a model containing three QTLs with two QTLs linked in coupling, and using 250 progeny

# cor	# incor	ANOVA	Zeng				BIC			permu
			3	5	7	9	2	2.5	3	
3	0	69	31	25	19	13	180	65	19	133
2	0	526	412	315	240	199	509	496	395	539
1	0	332	429	443	430	421	199	395	554	246
0	0	1	97	184	281	334	0	2	9	0
3	1	4	0	0	0	0	7	0	0	6
2	1	26	6	7	5	6	59	13	5	37
1	1	40	18	12	18	13	35	28	17	34
0	1	0	6	11	5	12	0	0	0	0
other		2	1	3	2	2	11	1	1	5
≥ 1 wrong		72	31	33	30	33	112	42	23	82

multiplier helped to avoid this problem, but at the expense of a lower power for detection. Forward selection using a permutation test did well: it detected more QTLs than ANOVA and Zeng's method, while including incorrect markers only 8% of the time.

Table 4 shows which of the QTLs were correctly identified by the different methods. The first three columns, labelled "model," correspond to the three QTLs: first the QTL on chromosome 1, and then the two linked QTLs on chromosome 2. A one in these columns indicates that the QTL was correctly identified; a zero indicates that it was not found. Note that in this table, we ignore the markers which were incorrectly identified. For example, in the column labelled "ANOVA," the model "1 1 1" was identified 73 times out of 1000 simulations; this includes 69 times in which no extraneous markers were included, and 4 times in which one extraneous marker was included (see Table 3).

TABLE 4

Models identified in 1000 simulations of the model containing three QTLs with two QTLs (represented in the second and third columns) linked in coupling, and using 250 progeny

model	ANOVA	Zeng				BIC			permu
		3	5	7	9	2	2.5	3	
111	73	31	25	19	13	187	65	19	139
110	254	189	140	102	83	258	239	189	260
101	257	191	144	112	92	264	241	192	274
011	41	39	40	33	31	52	30	20	45
100	3	115	173	182	180	1	3	8	1
010	200	161	136	131	120	131	218	293	152
001	171	171	147	135	135	107	202	270	129
000	1	103	195	286	346	0	2	9	0

When forward selection and ANOVA identified just one QTL, it was almost always one of the two linked QTLs, but Zeng's method often picked only the QTL on chromosome 1. When two QTLs were identified, all of the methods tended to pick the QTL on chromosome 1 and one of the two linked QTLs. Note that the two linked QTLs on chromosome 2 were chosen at approximately equal frequencies by all of the methods, as expected by symmetry: the models "1 1 0" and "1 0 1" were chosen nearly the same number of times, as were the models "0 1 0" and "0 0 1."

Table 5 displays the joint distribution, across the 1000 simulations, of the numbers of correctly and incorrectly chosen markers when the linked QTLs are in repulsion.

The methods did not perform as well when the linked QTLs were in repulsion; ANOVA and forward selection suffered much more than Zeng's method. The number of incorrectly chosen markers showed little change from the case of coupling, for all of the methods. But the number of correctly identified QTLs, in comparison to coupling, was halved for ANOVA and forward selection. Zeng's method, on the other hand, showed very little change in its ability to identify QTLs, with the result that here his method worked better than ANOVA.

TABLE 5

Distribution of the numbers of correctly and incorrectly chosen markers in 1000 simulations of a model containing three QTLs with two QTLs linked in repulsion, and using 250 progeny

# cor	# incor	ANOVA	Zeng				BIC			permu
			3	5	7	9	2	2.5	3	
3	0	4	102	83	60	46	174	80	27	78
2	0	123	222	225	203	168	135	79	46	99
1	0	572	402	412	385	381	426	458	398	524
0	0	231	230	242	314	372	156	338	507	219
3	1	0	0	1	2	2	25	5	0	6
2	1	7	7	6	10	10	21	6	2	12
1	1	30	19	20	19	13	29	12	4	28
0	1	32	18	10	4	7	24	22	16	31
other		1	0	1	3	1	10	0	0	3
≥ 1 wrong		70	44	38	38	33	109	45	22	80

It is interesting to see that whereas Zeng's approach performed quite poorly when the QTLs were linked in coupling, even in comparison to ANOVA, it performed somewhat better than all of the other methods when the QTLs were in repulsion. The reason that Zeng's method is more successful in teasing out a pair of QTLs linked in repulsion, may be that such QTLs look more important when both are included in the model. Zeng's method forces the fit of the larger model, whereas forward selection considers the markers one at a time. This

difference is best illustrated in Table 4. When identifying just one QTL, ANOVA and forward selection generally pick one of the two linked QTLs, whereas Zeng's method picks from the three QTLs at nearly equal proportions.

Whereas the simulations presented here considered only cases with three QTLs, Broman (1997) also performed simulations with five QTLs; the results were similar.

4. Conclusions and discussion. Current methods for identifying QTLs focus on interval mapping: inferring the location of a QTL between marker loci. Yet interval mapping and its approximations have been shown to provide little improvement in power over simple ANOVA at the marker loci. When we dispense with interval mapping, we are left only with ANOVA and multiple regression; the use of these more simple methods for identifying QTLs has been neglected.

In addition, most current methods use multiple tests of hypotheses. The problem of identifying QTLs is best viewed as a problem in model selection. Having discarded interval mapping, we then seek to choose a set of marker loci which are at or near QTLs. The problem is not the standard one in model selection, where attention has been on minimizing prediction error. Still, the model selection literature has much to say about our current problem. Clearly, the single-QTL methods, such as ANOVA and interval mapping, will perform poorly when multiple linked QTLs are segregating in a cross. An appropriate approach is difficult to prescribe. In the simulations in Section 3, one method (forward selection) performed best in the case of QTLs linked in coupling, while another (Zeng's approach) performed best in the case of QTLs linked in repulsion. A more refined method, such as Markov chain Monte Carlo, does not necessarily lead to improved results. For example, for the data in Satagopan et al. (1996), interval mapping seemed to give nearly identical results to MCMC.

A number of decisions must be made when performing any model selection procedure. First, one must choose a criterion. For Zeng's method, one must choose how many variables to use as initial regressors; for the BIC- δ criterion, one must choose the value of the parameter δ . Second, one must decide how to search through the space of models: will forward selection suffice, or would a more extensive search, as provided by MCMC, give improved results? The choices that one makes will depend upon the experiment being performed: whether it is a backcross or an intercross, the number of progeny, the density of marker loci, the underlying genetic structure of the trait, and the ultimate goal of the experiment. When making these choices, one will need to perform multiple simulation experiments, using scenarios that seem reasonable, and using criteria for determining the performance of an approach which correspond

to the goals of the study.

Simulation studies of the kind we mention can be helpful for designing a strategy, but, after the data are obtained, further analysis must be carried out. We see the need for research on the use of resampling and bootstrap methods of analysis, to complement the randomization approach of Churchill and Doerge (1994), which focuses on null models.

REFERENCES

- AKAIKE, H. (1969). Fitting autoregressive models for prediction. *Annals of the Institute of Statistical Mathematics* **21** 243–247.
- BASTEN, C. J., WEIR, B. S. and ZENG, Z.-B. (1996). *QTL Cartographer: A reference manual and tutorial for QTL mapping*. Program in Statistical Genetics, Department of Statistics, North Carolina State University.
- BROMAN, K. W. (1997). Identifying quantitative trait loci in experimental crosses. PhD dissertation, Department of Statistics, University of California, Berkeley.
- CHURCHILL, G. A. and DOERGE, R. W. (1994). Empirical threshold values for quantitative trait mapping. *Genetics* **138** 963–971.
- COWEN, N. M. (1989). Multiple linear regression analysis of RFLP data sets used in mapping QTLs. Pages 113–116 in *Development and application of molecular markers to problems in plant genetics*, edited by HELEN TJARIS, T. and BURR, B. Cold Spring Harbor Laboratory, Cold Spring Harbor, New York.
- DARVASI, A., WEIREB, A., MINKE, V., WELLER, J. I. and SOLLER, M. (1993). Detecting marker-QTL linkage and estimating QTL gene effect and map location using a saturated genetic map. *Genetics* **134** 943–951.
- DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* **39** 1–38.
- DOERGE, R. W. and CHURCHILL, G. A. (1996). Permutation tests for multiple loci affecting a quantitative character. *Genetics* **142** 285–294.
- DUPUIS, J., BROWN, P. O. and SIEGMUND, D. (1995). Statistical methods for linkage analysis of complex traits from high-resolution maps of identity by descent. *Genetics* **140** 843–856.
- GELMAN, A., CARLIN, J. B., STERN, H. S. and RUBIN, D. B. (1995). *Bayesian data analysis*. Chapman and Hall, New York.
- GEMAN, S. and GEMAN, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6** 721–741.
- GREEN, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82** 711–732.
- HALDANE, J. B. S. (1919). The combination of linkage values and the calculation of distances between the loci of linked factors. *Journal of Genetics* **8** 299–309.
- HALEY, C. S. and KNOTT, S. A. (1992). A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity* **69** 315–324.
- HASTINGS, W. F. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57** 97–109.
- HYNE, V., KEARSEY, M. J., PIKE, D. J. and SNAPE, J. W. (1995). QTL analysis: unreliability and bias in estimation procedures. *Molecular Breeding* **1** 273–282.
- JANSEN, R. C. (1993). Interval mapping of multiple quantitative trait loci. *Genetics* **135** 205–211.
- JANSEN, R. C. (1994). Controlling the type I and type II errors in mapping quantitative trait loci. *Genetics* **138** 871–881.
- JANSEN, R. C. and STAM, P. (1994). High resolution of quantitative traits into multiple loci via interval mapping. *Genetics* **136** 1447–1455.

- KEARSEY, M. J. and HYNE, V. (1994). QTL analysis: a simple 'marker-regression' approach. *Theoretical and Applied Genetics* **89** 698–702.
- KNAPP, S. J. (1991). Using molecular markers to map multiple quantitative trait loci: models for backcross, recombinant inbred, and doubled haploid progeny. *Theoretical and Applied Genetics* **81** 333–338.
- KNAPP, S. J., BRIDGES, W. C., JR. and BIRKES, D. (1990). Mapping quantitative trait loci using molecular marker linkage maps. *Theoretical and Applied Genetics* **79** 583–592.
- KNOTT, S. A. and HALEY, C. S. (1992). Aspects of maximum likelihood methods for the mapping of quantitative trait loci in line crosses. *Genetics Research* **60** 139–151.
- LANDER, E. S. and BOTSTEIN, D. (1989). Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121** 185–199.
- LONG, A. D., MULLANEY, S. L., REID, L. A., FRY, J. D., LANGLEY, C. H. and MACKAY, T. F. C. (1995). High resolution mapping of genetic factors affecting abdominal bristle number in *Drosophila melanogaster*. *Genetics* **139** 1273–1291.
- MARTÍNEZ, O. and CURNOW, R. N. (1992). Estimating the locations and the sizes of the effects of quantitative trait loci using flanking markers. *Theoretical and Applied Genetics* **85** 480–488.
- MATHER K. and JINKS, J. L. (1977). *Introduction to biometrical genetics*. Chapman and Hall, London.
- MENG, X.-L. and RUBIN, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika* **80** 267–278.
- MILLER, A. J. (1990). *Subset selection in regression*. Chapman and Hall, New York.
- REBAÏ, A., GOFFINET, B. and MANGIN, B. (1995). Comparing power of different methods for QTL detection. *Biometrics* **51** 87–99.
- SATAGOPAN, J. M. and YANDELL, B. S. (1996). Estimating the number of quantitative trait loci via Bayesian model determination. Special contributed paper session on genetic analysis of quantitative traits and complex diseases, Biometrics section, Joint Statistical Meetings, Chicago, Illinois.
- SATAGOPAN, J. M., YANDELL, B. S., NEWTON, M. A. and OSBORN, T. C. (1996). A Bayesian approach to detect quantitative trait loci using Markov chain Monte Carlo. *Genetics* **144** 805–816.
- SCHWARZ, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* **6** 461–464.
- SHRIMPTON, A. E. and ROBERTSON, A. (1988). The isolation of polygenic factors controlling bristle score in *Drosophila melanogaster*. I. Allocation of third chromosome sternopleural bristle effects to chromosome sections. *Genetics* **118** 437–443.
- SIMPSON, S. P. (1989). Detection of linkage between quantitative trait loci and restriction fragment length polymorphisms using inbred lines. *Theoretical and Applied Genetics* **77** 815–819.
- SOLLER, M., BRODY, T. and GENIZI, A. (1976). On the power of experimental designs for the detection of linkage between marker loci and quantitative loci in crosses between inbred lines. *Theoretical and Applied Genetics* **47** 35–39.
- TANKSLEY, S. D. (1993). Mapping polygenes. *Annual Review of Genetics* **27** 205–233.
- VAN OOIJEN, J. W. (1992). Accuracy of mapping quantitative trait loci in autogamous species. *Theoretical and Applied Genetics* **84** 803–811.
- VENABLES, W. N. and RIPLEY, B. D. (1994). *Modern applied statistics with S-Plus*. Springer-Verlag, New York.
- WELLER, J. I. (1986). Maximum likelihood techniques for the mapping and analysis of quantitative trait loci with the aid of genetic markers. *Biometrics* **42** 627–640.
- WELLER, J. I. (1987). Mapping and analysis of quantitative trait loci in *Lycopersicon* (tomato) with the aid of genetic markers using approximate maximum likelihood methods. *Heredity* **59** 413–421.
- WHITTAKER, J. C., THOMPSON, R. and VISSCHER, P. M. (1996). On the mapping of QTL by regression of phenotype on marker-type. *Heredity* **77** 23–32.
- WRIGHT, A. J. and MOWERS, R. P. (1994). Multiple regression for molecular-marker, quantitative trait data from large F₂ populations. *Theoretical and Applied Genetics* **89** 305–312.

- WRIGHT, F. A. and KONG, A. (1997). Linkage mapping in experimental crosses: the robustness of single-gene models. *Genetics* **146** 417-425.
- WU, W.-R. and LI, W.-M. (1994). A new approach for mapping quantitative trait loci using complete genetic marker linkage maps. *Theoretical and Applied Genetics* **89** 535-539.
- WU, W.-R. and LI, W.-M. (1996). Model fitting and model testing in the method of joint mapping of quantitative trait loci. *Theoretical and Applied Genetics* **92** 477-482.
- ZENG, Z.-B. (1993). Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci. *Proc. Natl. Acad. Sci. USA* **90** 10972-10976.
- ZENG, Z.-B. (1994). Precision mapping of quantitative trait loci. *Genetics* **136** 1457-1468.

CENTER FOR MEDICAL GENETICS
MARSHFIELD MEDICAL RESEARCH FOUNDATION
1000 N OAK AVENUE
MARSHFIELD WI 54449
BROMANK@CMG.MFLDCLIN.EDU

UNIVERSITY OF CALIFORNIA
DEPARTMENT OF STATISTICS
367 EVANS HALL # 3860
BERKELEY, CALIFORNIA 94720-3860
SPEED@STAT.BERKELEY.EDU

MARKOV CHAIN MONTE CARLO FOR THE BAYESIAN ANALYSIS OF EVOLUTIONARY TREES FROM ALIGNED MOLECULAR SEQUENCES

BY MICHAEL A. NEWTON, BOB MAU AND BRET LARGET¹

University of Wisconsin-Madison and Duquesne University

We show how to quantify the uncertainty in a phylogenetic tree inferred from molecular sequence information. Given a stochastic model of evolution, the Bayesian solution is simply to form a posterior probability distribution over the space of phylogenies. All inferences are derived from this posterior, including tree reconstructions, credible sets of good trees, and conclusions about monophyletic groups, for example. The challenging part is to approximate the posterior, and we do this by constructing a Markov chain having the posterior as its invariant distribution, following the approach of Mau, Newton, and Larget (1998). Our Markov chain Monte Carlo algorithm is based on small but global changes in the phylogeny, and exhibits good mixing properties empirically. We illustrate the methodology on DNA encoding mitochondrial cytochrome oxidase 1 gathered by Hafner *et al.* (1994) for a set of parasites and their hosts.

1. Introduction. Stochastic models have long been considered useful for describing variation in the molecular sequences of extant populations (e.g., Jukes and Cantor, 1969; Felsenstein, 1973; Kimura, 1980). Parameters in such models include the phylogeny, which encodes the pattern of evolutionary relationships among populations, and substitution rates, which describe how molecules change over time within populations. It seems quite natural to infer these parameters using the induced likelihood function in some way, but such inference has been difficult in practice because computations can be prohibitively expensive. Owing to the Markovian nature of the standard models, evaluation of the likelihood function follows straightforward recursive equations, and so evaluation is not the difficult part. The difficulty arises with optimization, since the likelihood resides over a complicated parameter space, and seems to admit no simple representation (Felsenstein, 1981, 1983; Goldman, 1990; Yang, Goldman, Friday, 1995). Nevertheless, computer code is available for approximate maximum likelihood calculation (Olson, *et al.*, 1994; Felsenstein, 1995; Swofford, 1996).

Beyond estimation, practitioners have demanded some way to assess uncertainty in aspects of the estimated phylogeny, just as error bars accompany simpler kinds of point estimates. A standard and appealingly simple calculation

¹B. Larget acknowledges the support of the National Science Foundation.

AMS 1991 subject classifications. Primary 62F99; secondary 92D20.

Key words and phrases. Cospeciation, Metropolis-Hastings algorithm, phylogeny.

is to apply Efron's bootstrap (Felsenstein, 1985), and although this method may accurately approximate sampling distributions, its role for statistical inference about the phylogeny has been a matter of some debate (e.g., Felsenstein and Kishino, 1993; Newton, 1996; Efron, Halloran, and Holmes, 1996; Chernoff, 1997). Of course, bootstrapping a complicated estimator serves to compound the computational problem. Thus, bootstrapping a full-blown maximum likelihood estimator is practically impossible with today's implementations. A common practice is to bootstrap a much simpler estimator.

An alternative, model-based, assessment of uncertainty was postulated some time ago by J.F.C. Kingman, in the discussion of Joe Felsenstein's 1983 paper on statistical issues in evolutionary biology:

In view of the difficulties of the maximum likelihood approach, it seems worth asking what a Bayesian analysis would look like. The author has shown us how to write down the likelihood function, and this has only to be multiplied by a suitable prior. . . . The result is a set of posterior probabilities for collections of possible phylogenies, not just a single estimate, and it may well be that there are tractable approximations of the probabilities of some compound events. Has this approach been explored?

Until recently, this Bayesian approach had not been explored. Sinsheimer et al. (1996) developed exact Bayesian calculations for the four-species problem. Several groups have been pursuing Markov chain Monte Carlo (MCMC) approximations. Mau and Newton (1997) described an MCMC method for models satisfying a molecular clock, and presented calculations for binary, restriction-sites data. Mau, Newton, and Larget (1998) have extended these calculations to problems with more taxa and nucleotide sequence data. Yang and Rannala (1997), and Li, Pearl and Doss (1996) have developed different Markov chain Monte Carlo strategies for the same general problem. In fact, the MCMC method of Kuhner, Yamato, and Felsenstein (1995) can be modified to produce approximate Bayesian phylogenetic inferences, even though their model considers within population sampling of sequences. The purpose of the present article is to review the Mau, Newton, Larget approach and to illustrate the calculations in an example.

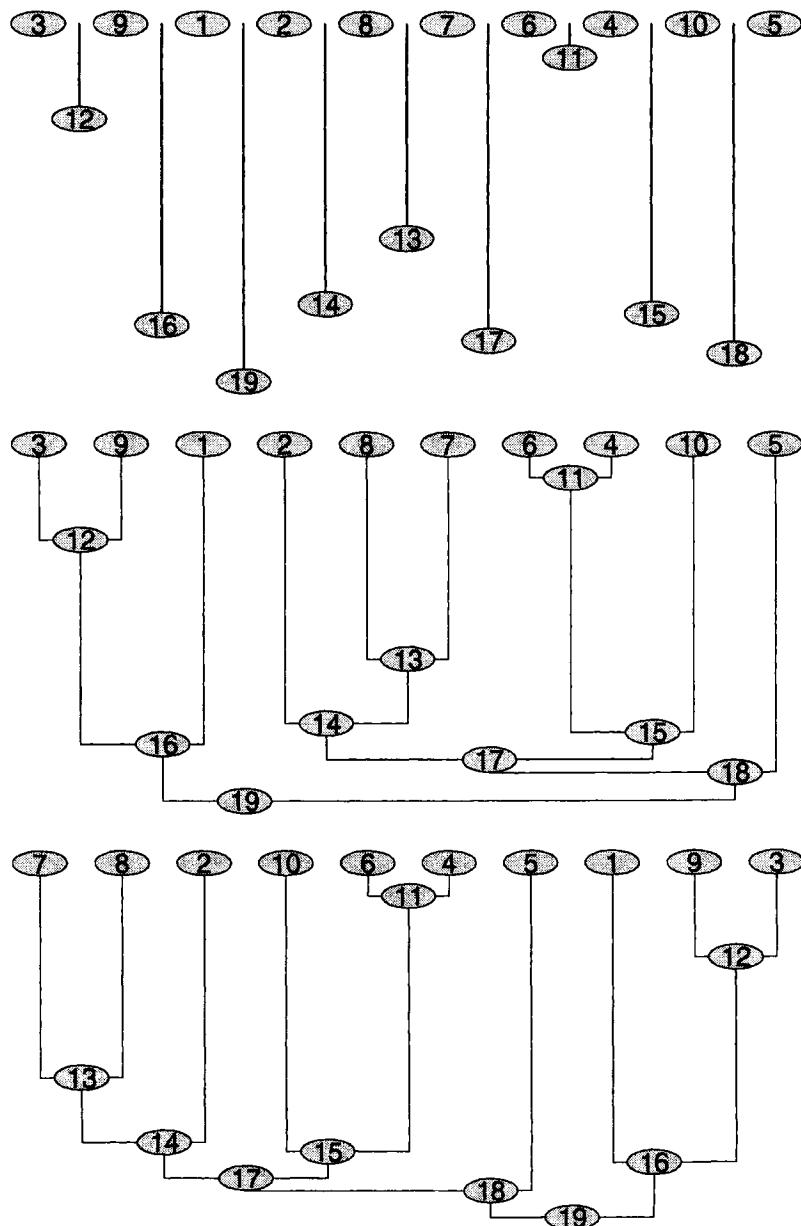
Of course Bayesian analysis provides more than assessments of uncertainty about phylogenies, the focus of this work. The array of inference problems presented in Huelsenbeck and Rannala (1997), for example, all may be approached from a Bayesian perspective. We anticipate that future research will clarify the role of Bayesian analysis for evolutionary biology, but first some essential computational problems must be addressed.

2. Phylogeny. A phylogeny or evolutionary tree, τ , admits various representations. For the present discussion, it will be convenient to treat τ as a pair (t, σ) where $t = (t_1, \dots, t_{s-1}) \in \mathbb{R}^{s-1}$ is a vector of positive speciation times, s is the number of species under consideration, and σ is a permutation of $\{1, 2, \dots, s\}$. The path of evolution corresponding to τ can be envisioned by processing t and σ in a manner as illustrated in Figure 1. Species labels $\{1, 2, \dots, s\}$ become leaf nodes of a tree upon being arranged horizontally in the order determined by σ . Moving from left to right, we drop a vertical line of length t_i from a point in between the i th and $(i+1)$ th leaf node, and we call the lower endpoint an internal node. In Figure 1, internal nodes are labeled by increasing speciation time. We draw a tree by moving downwards from the species labels, establishing in turn each internal node as the parent of two descendant nodes. Acting on a given internal node j , say, an edge is formed between node j and the parentless node $k < j$ having horizontal position closest to node j from the left, and a second edge is formed to the nearest such parentless node from the right. Eventually, all nodes are connected, and a tree results. The node corresponding to the largest t_i is called the root.

Several remarks are in order regarding this construction. The horizontal axis serves only to organize the information, and has no intrinsic scale. On the other hand, the vertical axis records time into the past. As drawn, our trees are rooted and have contemporaneous tips, and will be considered parameters of models which satisfy the molecular clock hypothesis. In work to extend our methods to models where evolutionary rates vary among branches of the tree, a somewhat different tree representation is more appropriate than the one just described. It is noteworthy, however, that the essential elements of the Monte Carlo algorithm to be described carry over readily to this more general case. Note that the drawing algorithm has not been defined when two times are equal, and so we omit this case (i.e., assume $t_i \neq t_j$ for all i, j) and thus consider only binary trees. This is not a serious restriction because the likelihood function (Section 3) is continuous in t , and hence the likelihood of a polytomy may be arbitrarily close to the likelihood of a binary tree produced by resolving the polytomy with tiny branch lengths.

The phylogeny $\tau = (t, \sigma)$ records the path of evolution from a single ancestral population to the present array of s populations under study. Any point on the tree drawn according to the rules above thus represents a population at some time in the past. Evidently, different (t, σ) pairs determine the same path of evolution, noting again that the horizontal axis in Figure 1 has no scale. For example, rearranging species 3 and 9 does not change the path of evolution. A given unordered set of times $\{t_1, t_2, \dots, t_{s-1}\}$ induces the same path of evolution in 2^{s-1} different ways. To see this, take a phylogeny and rotate the graph by

FIG. 1. *Phylogeny:* The top panel shows the raw ingredients in the (t, σ) representation of a phylogeny. Internal nodes appear at the lower end of the vertical line dropped from the horizontal. The middle panel shows the tree formed by processing the first panel, that is by moving down from the leaves towards the root, establishing connections each time an internal node is reached. The bottom panel shows a second version of the same phylogeny.



180 degrees above any of the $s - 1$ internal nodes. Strictly speaking, therefore, the phylogeny τ is an equivalence class containing 2^{s-1} different versions (t, σ) . The third panel in Figure 1 shows a second version of the preceding tree.

The representation of a phylogeny as 2^{s-1} points (t, σ) is particularly conducive to Markov chain Monte Carlo (MCMC), as we discuss in Section 4. A key feature is that the tree is part of a continuum, and the branching pattern of the tree is induced by the permutation σ and the relative ordering of the times t_1, \dots, t_{s-1} . Indeed, the branching pattern inherent in τ may be of interest, but we do not represent that pattern directly, choosing instead to work with more elementary objects which combine to produce the pattern. This somewhat indirect approach leads to very simple MCMC steps (Section 4) and may be associated with the efficiency of the algorithm.

Different summaries of τ may be of interest to the biologist. The *labeled history* describes the branching pattern of τ obtained by ignoring the magnitude of the times t_1, t_2, \dots, t_{s-1} , but respecting their ordering. Incidentally, counting labeled histories is quite simple given our construction, as there are $s!$ ways to arrange the s species labels, $(s - 1)!$ orderings of the times, but we have overcounted by 2^{s-1} , leaving $s!(s - 1)!/2^{s-1}$ distinct labeled histories. The *tree topology* corresponding to τ is another property of its branching pattern, where we record only the sequence of connections, but ignore details of their time ordering. The tree topology may be characterized by nested parentheses, such as

$$(2.1) \quad \text{top}(\tau) = ((1, (3, 9)), (((2, (7, 8)), ((4, 6), 10)), 5))$$

for the phylogeny shown in Figure 1. Here we have taken the convention that when two groups of organisms are merged, we place on the left that group containing the smallest species label. Assessing uncertainty in the tree topology is often of interest and will be the focus of our application in Section 5.

3. Modeling substitutions. The probability of data given a tree τ is derived from a model of DNA evolution along the branches, and many such models have been studied. We follow convention here and take the same general assumptions as those characterizing many standard models. That is, we consider the extant DNA sequences to be aligned into n sites, and we suppose that the evolution of different sites is independent. At any given site, a stochastic process associates a DNA base to each point on the branches of τ , and observed data are the s DNA bases at the leaf nodes. The standard models assume that base substitutions occur at points of Poisson processes, with independent evolution among branches. These restrictions still leave us some flexibility in the modeling of substitutions (Yang, Goldman, and Friday, 1994). It is noteworthy that the

MCMC algorithm discussed in Section 4 is not linked to the particular model of evolution. As long as likelihood evaluation is a feasible calculation, we can readily implement a posterior simulation. This is in contrast to Gibbs sampler algorithms, for example, whose very structure is determined by the likelihood function under consideration.

In Section 5 we report calculations for the model of Hasegawa, Kishino, and Yano (1985), which, being richer in parameters, subsumes the earlier models of Jukes and Cantor (1969), Kimura (1980), and Felsenstein (1981). The generator matrix for a process governed by HKY85 contains the following infinitesimal rates of change (the diagonal is determined because rows sum to 0):

$$\begin{array}{cccc} & A & G & C & T \\ \begin{matrix} A \\ G \\ C \\ T \end{matrix} & \left(\begin{array}{cccc} \cdot & \kappa\pi_g & \pi_c & \pi_t \\ \kappa\pi_a & \cdot & \pi_c & \pi_t \\ \pi_a & \pi_g & \cdot & \kappa\pi_t \\ \pi_a & \pi_g & \kappa\pi_c & \cdot \end{array} \right) \end{array}$$

The π 's indicate long-run probabilities of each base along one very long branch and κ allows different substitution rates for transitions (changes between A and G or between C and T) and transversions (any other changes). The infinitesimal rates determine transition probabilities from one base to another over any extended time period, and these further involve a mutation rate parameter θ . We omit details here.

By the independent-sites assumption, the likelihood is a product of n factors, one from each site in the aligned sequences. This naturally collapses to a product over $m \leq n$ unique observed patterns of s bases, and so fixing s , the likelihood evaluation takes $O(m)$ operations. Furthermore, the Poisson process assumption implies that evolution is Markovian, and thus that the probability of a given pattern can be calculated recursively in $O(s)$ steps. This *pruning* algorithm is a critical component of our procedure, and so we review it briefly (see also, Felsenstein, 1983). Let u_i denote the unknown base at the site of interest in the ancestral sequence associated with internal node i of the tree τ . Note that $s + 1 \leq i \leq 2s - 1$ and $u_i \in \{A, G, T, C\}$. Each internal node partitions the descendant species into two distinct groups, whose observed DNA data we label $A(i)$ and $B(i)$. By the assumed independence of substitutions among branches, the conditional probability of all data descending from i , given u_i , is

$$(3.2) \quad p\{A(i), B(i) | u_i, \tau\} = p\{A(i)|u_i, \tau\} \times p\{B(i)|u_i, \tau\}.$$

These probabilities are important because the likelihood contribution from a

site with the given pattern is

$$(3.3) \quad \sum_{u_{\text{root}}} p\{\mathbf{A}(\text{root}), \mathbf{B}(\text{root}) | u_{\text{root}}, \tau\} p(u_{\text{root}}).$$

That is, it is a mixture of transition probabilities against the distribution of the unknown base at the root node. By taking these initial base probabilities equal to the stationary base probabilities π_a , π_c , π_t , or π_g , the Markov process becomes reversible. To implement the pruning algorithm, one observes that by the Markov property, probabilities in (3.2) may be obtained recursively, moving from the leaves of the tree to the root. In our labeling system, the recursion moves successively through internal nodes $i = s + 1$ to $i = 2s - 1$.

On the phylogeny in Figure 1, for example, $p\{\mathbf{A}(12), \mathbf{B}(12) | u_{12}, \tau\}$ is the product of the conditional probability of data for species 3 and species 9 given u_{12} , the base at internal node 12. These four conditional probabilities are used subsequently to evaluate the conditional probability of data descending from node 16, given u_{16} , which finally enters the likelihood calculation (3.3).

We note that the (t, σ) representation of τ is not the one most conducive to the pruning calculation which relies directly on relationship information in τ . Our software uses a second representation in which every internal node is associated with its descendant nodes.

In summary, likelihood evaluation is a straightforward calculation when we fix the data, the tree, and parameters governing the substitution model.

4. The posterior and MCMC. In contrast to other forms of statistical inference, Bayesian inference centers on the extent to which opinion about an unknown is affected by data. Furthermore, probability is the sole medium for transmitting uncertainty and opinion (e.g., Bernardo and Smith, 1994). To implement an analysis, a Bayesian evolutionary biologist must therefore begin with a probability distribution over the set of possible phylogenies. This might be derived from a model of speciation, or from the analysis of existing data. To our knowledge, little work has been done on the assessment of *prior* probabilities for trees, but this certainly represents an important problem if Bayesian analysis is to be helpful in evolutionary biology. In the present work, we illustrate calculations with a particularly simple *flat* prior, and note that the algorithm proceeds easily with any user-supplied prior distribution. The flat prior we assume is relative to the (t, σ) representation of the phylogeny, in suggestive notation:

$$(4.4) \quad p(\tau) = p(t) p(\sigma) = \left(\prod_{i=1}^{s-1} p(t_i) \right) \frac{1}{s!} \propto \left(\prod_{i=1}^{s-1} 1[0 \leq t_i \leq t_{\max}] \right) \propto 1[0 \leq t_i \leq t_{\max}, \text{ for all } i].$$

Here t_{\max} bounds the time to the root node. One consequence of this prior is that, like the Kingman coalescent (Kingman, 1982), we induce a uniform probability distribution over labeled histories, and thus a non-uniform distribution over topologies. This prior favors balanced tree topologies because they have many more labeled histories than unbalanced ones (e.g., Lapointe and Legendre, 1991; Brown, 1994). Checking the sensitivity of our calculations to the choice of prior will be critical in applications, but we do not pursue such sensitivity analysis here.

Parameters of the substitution model also are unknown and a full Bayesian analysis requires a prior distribution for them. In this section we focus on the phylogeny, and thus we consider all other parameters to be known. For example, the stationary base probabilities $\pi_a, \pi_c, \pi_t, \pi_g$ can be estimated by the relative frequency of the different bases in the observed sequences.

In light of the data, and taking as reasonable the stochastic model of evolution, inference about the phylogeny τ must be based on the posterior distribution, having density

$$(4.5) \quad p(\tau | \text{data}) \propto p(\text{data} | \tau) \times p(\tau).$$

Monte Carlo appears to be the only effective method for summarizing this distribution, even though the pruning algorithm enables evaluation of the posterior up to a constant. Inference about monophyletic groups, most probable topologies, and the uncertainty in certain branch points, for example, all are based on expectations with respect to this posterior. Within the class Monte Carlo algorithms, Markov chain methods present the most promising integration methods, and we review here the proposal of Mau, Newton, and Larget (1998).

An MCMC algorithm realizes a Markov chain $\tau^1, \tau^2, \dots, \tau^B$ that has (4.5) as its stationary distribution (e.g., Tierney, 1994). Empirical averages in the chain converge as B grows to posterior expectations by the law of large numbers for Markov chains. We construct our Markov chain using the Metropolis-Hastings approach. That is, we move from $\tau^i = \tau$ to the next state τ^{i+1} by first proposing a candidate phylogeny τ^* generated according to a proposal distribution that has transition density $q(\tau, \tau^*)$. Next we compute the Metropolis-Hastings ratio

$$(4.6) \quad r = \frac{p(\tau^* | \text{data}) q(\tau^*, \tau)}{p(\tau | \text{data}) q(\tau, \tau^*)}.$$

If $r \geq 1$, then $\tau^{i+1} = \tau^*$. Otherwise, we move to τ^* with probability r and stay put with probability $1 - r$. The power of this approach resides both in its simplicity and in its great flexibility, because the choice of q , which affects the Monte Carlo efficiency of the algorithm, is almost arbitrary.

Monte Carlo approximations of posterior probabilities can be biased if the distribution of τ^1 is far from the target posterior distribution, and so it is common practice to let the chain run for a burn-in period before using any of the sampled states. Determining the length of the burn-in and the total chain length B to ensure accurate approximations is difficult in advance, and is typically based on realizations of the chain that are monitored using a range of diagnostic checks (e.g., Cowles and Carlin, 1996).

In most implementations of the Metropolis-Hastings algorithm, a collection of proposal distributions determine the complete algorithm (e.g., Besag *et al.*, 1995). We have found that a single proposal distribution works for the phylogeny problems considered so far. This proposal distribution is global in that τ^* can differ from τ in *all* respects, and so, in a sense, we have attempted to design efficiency into the algorithm. Inefficient algorithms are ones which traverse the parameter space slowly and thus exhibit significant positive correlations on one-dimensional summaries. Local, single-site updating proposals change parts of the parameter at a time, and are at risk for low efficiency. One risk of a global proposal distribution, on the other hand, is that we may reject candidates too frequently, and thus produce an inefficient algorithm. We avoid this in two ways: by making our global changes small in magnitude, and by basing changes on distance within the tree, so that proposed trees are close in posterior density to the current tree.

More specifically, our proposal distribution works like this. We obtain at random from the equivalence class defining the current tree τ one of its 2^{s-1} versions, thus identifying a pair (t, σ) . Fixing the leaf label permutation σ , we generate a new vector t^* of times by

$$t_i^* = t_i \oplus \epsilon_i, \quad \text{for } i = 1, 2, \dots, s-1$$

where ϵ_i are independent and identically distributed $\text{Uniform}(-\delta, \delta)$ random variables for some tuning parameter $\delta > 0$, and \oplus indicates addition reflected into the interval $(0, t_{\max})$. For example, \oplus returns $|t_i + \epsilon|$ if $t_i + \epsilon < 0$. Thus the proposal is to perturb the speciation times of a version of the current tree.

When the tuning parameter δ is small, the candidate tree is close to the current tree in terms of pairwise distance between species, and so we expect the likelihood of the candidate tree to be close to that of the current tree. Similarity in likelihood is derived by the similar distance structure, and not by a direct appeal to the model, making the proposal method independent of the model form. Interestingly, the candidate tree can be quite different from the current tree in terms of branching structure.

In Figure 2, a version of τ from Figure 1 has had its species times perturbed,

FIG. 2. *Proposal:* This graph shows how a candidate phylogeny τ^* is obtained from one version of the current tree by perturbing speciation times in the (t, σ) representation. The shaded boxes indicate the range of the uniform perturbations. The dark circles indicate times within the current tree, and crosses indicate times in a particular candidate τ^* .

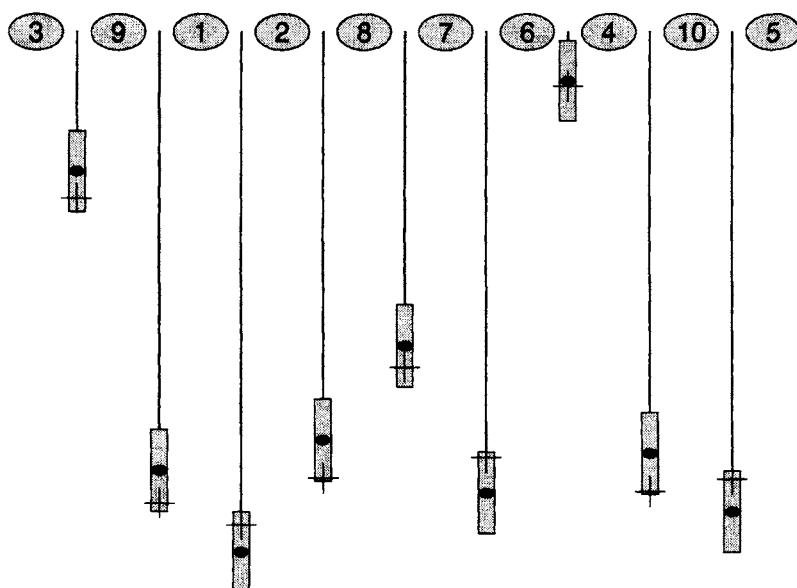


TABLE 1
Gopher/Lice Species Labels

Label	Louse Species	Label	Gopher Species
1	G. texanus	1	G. personatus
2	G. ewingi	2	G. breviceps
3	G. oklahomensis	3	G. bursarius (a)
4	G. geomydis	4	G. bursarius (b)
5	G. nadleri	5	P. bulleri
6	G. chapini	6	O. hispidus
7	G. panamensis	7	O. cavator
8	G. setzeri	8	O. underwoodi
9	G. cherriei	9	O. cherriei
10	G. costaricensis	10	O. heterodus
11	G. expansus	11	C. castanops
12	G. perotensis	12	C. merriami
13	G. trichopi	13	Z. trichopus

leading to a tree τ^* with a different topology:

$$\text{top}(\tau^*) = ((1, (3, 9)), ((2, ((4, 6), (7, 8))), (5, 10)))$$

Compare with (2.1). Certainly movements induced by this proposal mechanism are restricted, unless δ is very large. Mau, Newton, and Larget (1998) established irreducibility; i.e., that starting at any phylogeny τ , and given any other phylogeny τ_0 , there exists $K < \infty$ such that there is positive probability density of moving to τ_0 after K applications of the proposal. Mau, Newton, and Larget (1998) also established symmetry which means that the Metropolis-Hastings ratio (4.6) reduces to a ratio of posterior densities, and thus, under a flat prior, to a ratio of likelihoods.

5. An example: Host parasite evolution.

5.1. Data and model structure. We illustrate the MCMC calculations with data reported by Hafner *et al.* (1994) regarding a study of molecular evolution in hosts and their parasites. To facilitate a comparison, we focus on a subset that was analyzed by Huelsenbeck, Rannala, and Yang (1997) (hereafter HRY97). The data consist of 26 aligned DNA sequences, $n = 379$ bases long, encoding mitochondrial cytochrome oxidase I (COI) for 13 lice species and the corresponding 13 species of their gopher hosts (Table 1). There are $m = 156$ and $m = 130$ unique site patterns in the lice and gopher data, respectively. Table 2 shows summary frequencies of the different nucleotide bases, as well as the numbers of sites at which all bases are the same and sites exhibiting base variation.

Because the life cycle of the parasitic lice occurs exclusively in the fur of the host gophers, a natural hypothesis is that the organisms have coevolved, and thus have a common branching structure in their phylogenies. On the other hand, factors such as interaction among gopher species could produce differences between louse and gopher trees. Using the HKY85 model of DNA substitution discussed in Section 3, we compare the topological structure of host and parasite phylogenies. Our approach is to integrate results from separate analyses of the louse and gopher data.

The overall amount of DNA variation differs between gophers and lice, but within each group a molecular clock assumption is reasonable (HRY97, p. 414). By splitting the 26 taxa into two groups, we overcome the need to specifically model violations of a molecular clock.

We allow the mutation rate θ to vary over codon position because of the significant rate variation among sites (Table 2). It turns out that a model with codon-specific mutation rates fits somewhat better than the one used by HRY97

TABLE 2

Summary Statistics: Middle four columns show the observed base frequencies broken down by codon position for both louse and gopher data sets. The total number of sites is decomposed into n_c sites that have constant base value among all species, and n_v sites that exhibit variation.

data	codon	$\hat{\pi}_a$	$\hat{\pi}_g$	$\hat{\pi}_c$	$\hat{\pi}_t$	n_c	n_v
lice	1	0.281	0.362	0.101	0.254	101	25
	2	0.136	0.196	0.202	0.464	120	6
	3	0.321	0.185	0.112	0.388	2	125
	all	0.246	0.248	0.138	0.369	223	156
gopher	1	0.314	0.300	0.121	0.262	113	13
	2	0.160	0.172	0.230	0.435	122	4
	3	0.386	0.051	0.232	0.337	14	113
	all	0.287	0.174	0.195	0.345	249	130

in which mutation rates follow a discretized Gamma distribution. Observed base frequencies also vary over codon position, and so we similarly allow codon-specific relative frequency parameters. In the calculations reported below, we consider the base frequency parameters (Table 2) and the mutation rates to be fixed at their estimated values. Actually, we use some preliminary MCMC runs to estimate the mutation rate parameters, (0.136, 0.026, 2.838) for the lice and (0.154, 0.021, 2.825) for the gophers. These mean 1 vectors are empirical averages taken from long preliminary MCMC runs in which both the phylogeny and the mutation rate parameters were updated. Small posterior variance lead us to fix these rates at their estimated values.

HRY97 found evidence of transition/transversion bias, and so we follow suit, allowing a free parameter κ for each data set. Rather than fixing an estimated value, we augment the MCMC algorithm, including κ as an additional unknown, having a flat prior on the positive line. We considered a model with codon-specific κ , but this did not significantly improve fits.

5.2. MCMC implementation. For each data set, our Monte Carlo estimate of the posterior distribution over phylogenies is based on realizing four independent Markov chains of length 1,020,000. In the first 20,000 cycles of each run, only the phylogeny is updated, starting from a random tree drawn from the uniform prior distribution, and κ is fixed at a rough estimate obtained from preliminary runs (9.87 for lice, 11.45 for gophers). Subsequently, each cycle alternates between an update of τ given κ and an update of κ given τ , the latter based on a simple uniform window proposal distribution. During the initial cycles, we adaptively change the window size δ for the τ update, ultimately fixing values of

$\delta = 0.00625$ and $\delta = 0.003125$ for lice and gophers, respectively. Our automated method increases δ if the recent acceptance rate is high, and decreases it otherwise, but only adapts during the early burn-in phase. We routinely monitor the loglikelihood of sampled trees throughout this burn-in period, noting a typical pattern of dramatic increase followed by stability at some plateau. The pattern is consistent across independent runs.

After burn in, each production run is subsampled every 20 cycles to reduce the size of output used in estimating posterior probabilities. We calculated the tree topology and loglikelihood of all subsampled phylogenies. Figures 3 and 4 show some diagnostic plots from a further one out of 20 subsampling of these 50,000 phylogenies from one of the four runs. At this level of subsampling, there is very little within chain dependence both for the loglikelihood series and for the binary series indicating whether or not $\text{top}(\tau)$ equals the most frequently observed topology. Rather than show a simple time-series plot of this binary variable, we use the cusum diagnostic suggested in Yu (1995). We simply plot the cumulative sum of the binary time series, centered by a cumulative sum of ones times the overall mean. Slow mixing can be diagnosed when we compare the cusum plot to a similar plot calculated on a random permutation of the binary series. Slow mixing is characterized by long excursions and a fairly smooth plot. Rapid mixing is indicated by these plots.

It is more relevant to consider dependence properties within each production run of 50,000 phylogenies than within the subsamples in Figures 3 and 4 because our Monte Carlo approximations arise from the former. Nevertheless, the plots provide some indication of sampler behavior and demonstrate adequate mixing on several summary quantities.

That four independent runs produce comparable results suggests that any posterior multimodality is not adversely affecting the sampler. Furthermore, the four independent runs provide simple Monte Carlo standard error estimates in place of the somewhat more complicated within-chain methods (Geyer, 1992).

5.3. Posterior summaries. Tables 3 and 4 summarize our Monte Carlo estimates of the posterior distribution over tree topologies separately for the lice and gophers. In each run, the posterior probability of a topology is calculated simply as its empirical relative frequency. Over the four runs, the average of these proportions is our reported Monte Carlo estimate, and the standard deviation divided by two is our reported Monte Carlo standard error. Tables 3 and 4 take advantage of clear subtopological structure and also report only topologies in an 80% credible set.

The most probable tree topologies that we find agree with those determined by approximate maximum likelihood in HRY97 (their Figure 4). The best louse

TABLE 3

Posterior Distribution over Topologies, Lice: Subtopologies are $A_1 = (((1, 2), (3, 4)), 11)$, $A_2 = ((1, 2), ((3, 4), 11))$, $B_1 = (5, 13)$, $C_1 = (7, 8)$, and $D_1 = (9, 10)$.

Rank	Topology $\text{top}(\tau)$	$p[\text{top}(\tau) \text{data}] \pm \text{se}$	cumulative
1	$((A_1, B_1), ((6, (C_1, D_1)), 12))$	0.514 ± 0.005	0.514
2	$(((A_1, B_1), (6, (C_1, D_1))), 12)$	0.101 ± 0.002	0.615
3	$(((A_1, B_1), 12), (6, (C_1, D_1)))$	0.073 ± 0.001	0.688
4	$((A_1, B_1), (((6, D_1), C_1), 12))$	0.044 ± 0.001	0.732
5	$((A_2, B_1), ((6, (C_1, D_1)), 12))$	0.043 ± 0.002	0.775
6	$((A_1, (6, (C_1, D_1))), (B_1, 12))$	0.027 ± 0.003	0.803

TABLE 4

Posterior Distribution over Topologies, Gophers: Subtopologies are $E_1 = (1, (2, (3, 4)))$, $F_1 = (6, ((7, 8), (9, 10)))$, and $G_1 = (11, 12)$.

Rank	Topology $\text{top}(\tau)$	$p[\text{top}(\tau) \text{data}] \pm \text{se}$	cumulative
1	$((E_1, F_1), ((5, G_1), 13))$	0.118 ± 0.003	0.118
2	$(((E_1, F_1), 5), (G_1, 13))$	0.084 ± 0.003	0.202
3	$(((E_1, F_1), (5, G_1)), 13)$	0.082 ± 0.001	0.284
4	$((E_1, F_1), (5, (G_1, 13)))$	0.062 ± 0.003	0.346
5	$(((E_1, F_1), 5), G_1), 13)$	0.055 ± 0.001	0.400
6	$((E_1, F_1), ((5, 13), G_1))$	0.050 ± 0.002	0.451
7	$(((E_1, F_1), 13), (5, G_1))$	0.043 ± 0.001	0.494
8	$(((E_1, F_1), 5), 13), G_1)$	0.033 ± 0.002	0.526
9	$((E_1, 13), ((5, G_1), F_1))$	0.032 ± 0.003	0.558
10	$((E_1, (5, F_1)), (G_1, 13))$	0.027 ± 0.003	0.585
11	$((E_1, 13), ((5, F_1), G_1))$	0.024 ± 0.001	0.609
12	$(E_1, ((5, G_1), (F_1, 13)))$	0.023 ± 0.001	0.632
13	$((E_1, ((5, G_1), F_1)), 13)$	0.021 ± 0.002	0.653
14	$(E_1, (((5, F_1), G_1), 13))$	0.020 ± 0.001	0.673
15	$(E_1, (((5, G_1), F_1), 13))$	0.019 ± 0.001	0.692
16	$(E_1, (((5, G_1), 13), F_1))$	0.018 ± 0.002	0.710
17	$((E_1, ((5, F_1), G_1)), 13)$	0.015 ± 0.001	0.725
18	$(((E_1, F_1), (5, 13)), G_1)$	0.015 ± 0.001	0.740
19	$(E_1, ((5, (G_1, 13)), F_1))$	0.011 ± 0.001	0.752
20	$(((E_1, (5, F_1)), G_1), 13)$	0.011 ± 0.001	0.763
21	$((E_1, (5, G_1)), (F_1, 13))$	0.010 ± 0.001	0.773
22	$((E_1, (F_1, 13)), (5, G_1))$	0.010 ± 0.001	0.783
23	$(((E_1, F_1), (G_1, 13)), 5)$	0.009 ± 0.001	0.792
24	$(((E_1, (5, F_1)), 13), G_1)$	0.009 ± 0.001	0.801

topology is well supported, with a posterior probability of 51.4%. Here, most of the uncertainty involves the placement of taxon 12, and to a lesser extent, taxon 6. Marginally, the most probable subtopology for the remaining 11 taxa has probability 74.3%, and only ten subtopologies are in a 99% credible set. Monophyletic groups, or clades, $A = \{1, 2, 3, 4, 11\}$, $B = \{5, 13\}$, $C = \{7, 8\}$, and $D = \{9, 10\}$, occur with probability exceeding 99.4%. Collapsing subtopologies within clades is another type of marginalization that leads to succinct summaries of the posterior. Ignoring taxa 6 and 12, these clades are grouped as either $((A, B), (C, D))$, $((A, (C, D)), B)$, or $(A, (B, (C, D)))$ with probabilities 88.1%, 9.5%, and 2.3% respectively (and a 0.1% probability that not all these clades appear).

Somewhat greater uncertainty is present in the gopher phylogeny (Table 4), with 24 topologies in the 80% credible set, and the most probable one of probability only 11.8%. Much of the uncertainty lies in the positioning of taxa 5 and 13. Marginally for the remaining 11 taxa, the best gopher subtopology has probability 65.6%, and only six subtopologies are necessary to form a 99% credible set. Three clades are identified with posterior probability 1: $E = \{1, 2, 3, 4\}$, $F = \{6, 7, 8, 9, 10\}$, and $G = \{11, 12\}$. Ignoring variation in subtree topology within clades and the placement of taxa 5 and 13, the clades are joined as either $((E, G), F)$, $(E, (F, G))$ or $((E, F), G)$ with probabilities 67.8%, 27.5%, and 4.7% respectively.

While there is substantial structural similarity for most of the sampled trees from both posterior distributions, there is no single tree topology which appears in both samples. The absence of posterior overlap provides evidence against strict coevolution of the hosts and parasites, similar to the conclusion in HRY97. From our posterior sample, we can infer more. In particular, we are able to isolate and quantify those species pairs where coevolution appears to fail. The unanimous placement of taxa 11 and 12 as nearest relatives with probability 1 in the gopher posterior contrasts markedly with the highly variable placement of taxa 12 in the louse posterior. Similarly, taxa 5 and 13, bound together in the louse posterior, vary wildly between clades in the gopher posterior. These gopher/louse pairs are the most likely candidates for an alternate evolutionary pathway. In contrast, we can identify the largest set of gopher/louse pairs supporting strict cospeciation. The common subtopology of seven stably attached taxa is $((1, (3, 4)), ((7, 8), (9, 10)))$ and has marginal posterior probability 99.5% for lice and 99.6% for gophers. Having a large Monte Carlo sample makes such a determination fairly straightforward.

Without pursuing it further, we note that parameter estimation and model building are effectively carried out with the help of an MCMC sampler. We settled upon codon-specific mutation rates and a single transition/transversion

FIG. 3. *Output Analysis, Lice:* Panels on the right are autocorrelation functions for two summaries of the phylogeny sequence sampled by MCMC: loglikelihood, and indicator of best topology. For the loglikelihood series, the left panel shows simple time series plots of the output. A cusum plot is given in the left panel for the binary indicator of best topology. The dotted curve is the cusum plot of a random permutation of the series.

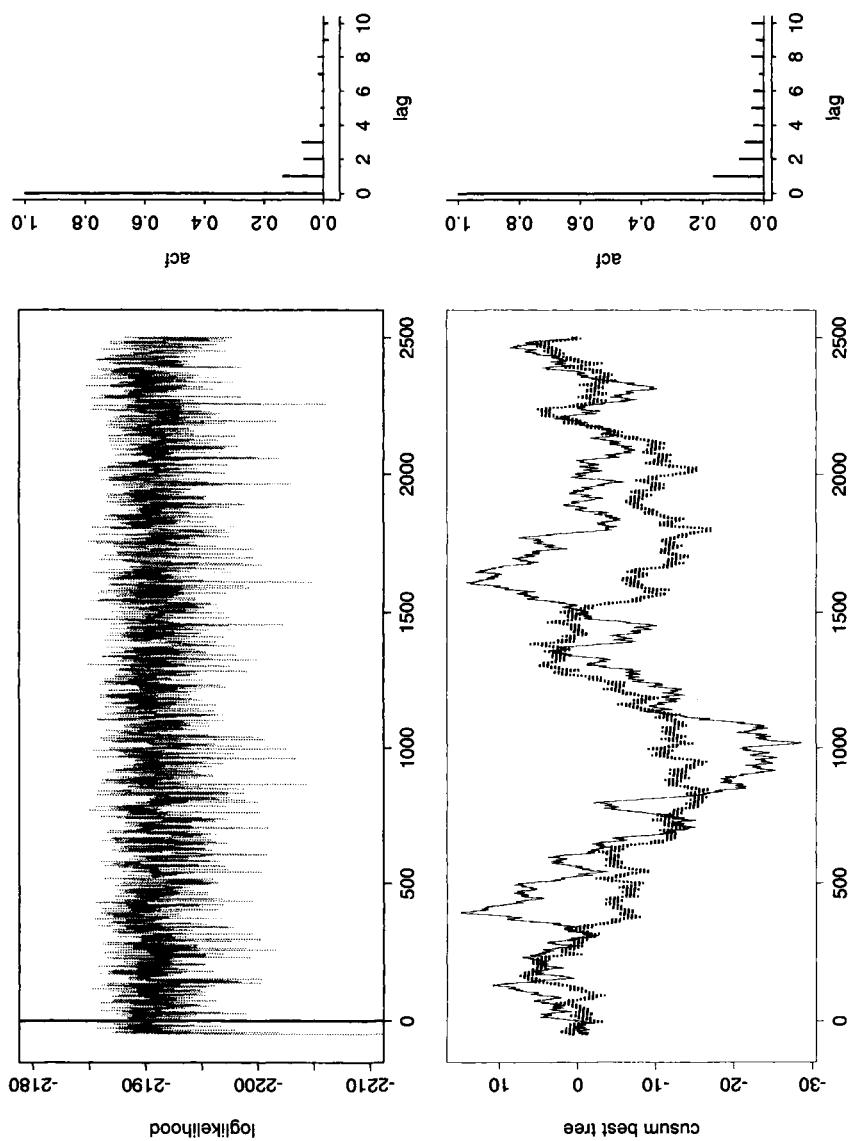
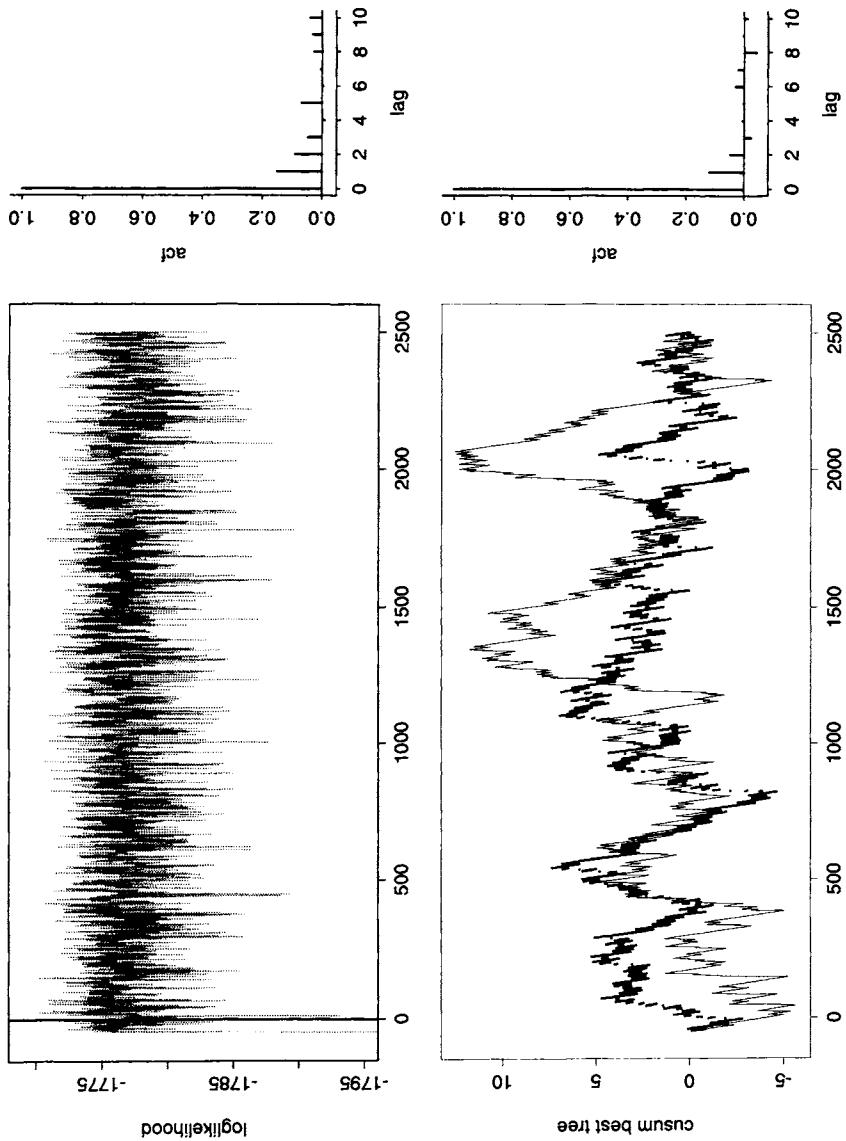


FIG. 4. *Output Analysis, Gophers: Same diagnostics as Figure 3.*

bias parameter by simply rerunning our chains using different likelihood evaluation routines, and allowing simultaneous parameter updating. Importantly, final parameter estimates do not condition on an estimated phylogeny, and no optimization methods are used.

6. Concluding remarks. Much remains to be done before we understand the utility of Bayesian methods for evolutionary biology. They may be helpful for studying cospeciation because the scientific questions of interest relate to topological structure of the phylogeny, and more classical statistical methods do not provide completely satisfactory results. Inference based on likelihood ratio tests may be effective, but frequency calibration typically requires fixing parameter estimates and phylogenies, and hence exact significance levels are unknown. Other tests ask if the similarity between host and parasite estimated phylogenies is more than can be attributed to chance, under a model of phylogenesis (e.g., Page 1988), but the reference measure here appears to have little to do with the cospeciation hypothesis. A Bayesian approach, on the other hand, allows us to make direct probabilistic statements concerning relevant aspects of phylogeny structure.

Even the most simple questions require sophisticated computation, and so we have started our investigation by trying to approximate the posterior distribution over phylogeny space within the context of a standard parametric model of evolution. Initial experimentation with these computations is cause for some optimism. The problem with Markov chain Monte Carlo is not so much in developing an algorithm as it is in developing a reasonably efficient algorithm, and we think that our simple technique of perturbing speciation times shows promise. Further research is needed to uncover the relative merits of competing algorithms. In addition, it may be helpful to clarify the relationship of Bayesian and bootstrap methodology, so that biologists will with more confidence be able to assess uncertainty in evolutionary hypotheses.

REFERENCES

- BERNARDO, J.M. and SMITH, A.F.M (1994). *Bayesian Theory*. Wiley, New York.
- BESAG, J., GREEN, P.J., HIGDON, D. and MENGERSEN, K. (1995). Bayesian computation and stochastic systems. *Statistical Science* **10** 3–66.
- BROWN, J.M.K. (1994). Probabilities of evolutionary trees. *Systematic Biology* **43** 78–91.
- COWLES, M.K. and CARLIN, B.P. (1996). MCMC convergence diagnostics: A comparative review. *Journal of the American Statistical Association*, **91** 883–904.
- CHERNOFF, H. (1997). Invited Lecture. IMA Summer Research Program on Statistics in the Health Sciences. Minneapolis, July, 1997.
- EFRON, B., HALLORAN, B. and HOLMES, S. (1996). Bootstrap confidence levels for phylogenetic trees. *Proceedings of the National Academy of Sciences of the USA* **13429–13434**.
- FELSENSTEIN, J. (1973). Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. *Systematic Zoology* **22** 240–249.

- FELSENSTEIN, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution* **17** 368–376.
- FELSENSTEIN, J. (1983). Statistical inference of phylogenies (with discussion). *Journal of the Royal Statistical Society, Series A*, **146** 246–272.
- FELSENSTEIN, J. (1985). Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39** 783–791.
- FELSENSTEIN, J. (1995). PHYLIP (phylogeny inference package) version 3.5c. Computer program distributed by the University of Washington.
- FELSENSTEIN, J. and KISHINO, H. (1993). Is there something wrong with the bootstrap? A reply to Hillis and Bull. *Systematic Biology* **42** 193–200.
- GEYER, C.J. (1992). Practical Markov chain Monte Carlo (with discussion). *Statistical Science* **7** 437–511.
- GOLDMAN, N. (1990). Maximum likelihood inference of phylogenetic trees, with special reference to a Poisson process model of DNA substitution and to parsimony analyses. *Systematic Zoology* **39** 345–361.
- HAFNER, M.S., SUDMAN, P.D., VILLABLANCA, F.X., SPRADLING, T.A., DEMASTES, J.W. and NADLER, S.A. (1994). Disparate rates of molecular evolution in cospeciating hosts and parasites. *Science* **265** 1087–1090.
- HASEGAWA, M., KISHINO, H. and YANO, T. (1985). Dating the Human-Ape Splitting by a Molecular Clock of Mitochondrial DNA. *Journal of Molecular Evolution* **22** 160–174.
- HUELSENBECK, J.P. and RANNALA, B. (1997). Phylogenetic methods come of age: Testing hypotheses in an evolutionary context. *Science* **276** 227–232.
- HUELSENBECK, J.P., RANNALA, B., and YANG, Z. (1997). Statistical tests of host-parasite cospeciation. *Evolution* **51** 410–419.
- JKUES, G.H. and CANTOR, C.R. (1969). Evolution of protein molecules. In *Mammalian Protein Metabolism*, Munroe, H.N. (ed.), pp. 21–132. Academic Press, New York.
- KIMURA, M. (1980). A simple method for estimating rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution* **16** 111–120.
- KINGMAN, J.F.C. (1982). The Coalescent. *Stochastic Processes and their Applications* **13** 235–248.
- KUHNER, M.K., YAMATO, J. and FELSENSTEIN, J. (1995). Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. *Genetics* **140** 1421–1430.
- LAPOINTE, F.-J. and LEGENDRE, P. (1991). The generation of random ultrametric matrices representing dendograms. *Journal of Classification* **8** 177–200.
- LI, S., PEARL, D.K. and DOSS, H. (1996). Phylogenetic tree construction using Markov chain Monte Carlo. Technical Report 583, Department of Statistics, Ohio State University.
- MAU, B. and NEWTON, M.A. (1997). Phylogenetic inference for binary data on dendograms using Markov chain Monte Carlo. *Journal of Computational and Graphical Statistics*, **6** 122–131.
- MAU, B., NEWTON, M.A. and LARGET, B. (1998). Bayesian phylogenetic inference via Markov chain Monte Carlo methods. *Biometrics*, to appear.
- NEWTON, M.A. (1996). Bootstrapping phylogenies: Large deviations and dispersion effects. *Biometrika* **83** 315–328.
- OLSEN, G.J., MATSUDA, H., HAGSTROM, R. and OVERBECK R. (1994). FastDNAML: a tool for the construction of phylogenetic trees of DNA sequences using maximum likelihood. *CABIOS* **10** 41–48.
- PAGE, R.D.M. (1988). Quantitative cladistic biogeography: Constructing and comparing area cladograms. *Systematic Zoology* **37** 254–270.
- SINSHEIMER, J.S., LAKE, J.A. and LITTLE, R.J.A. (1996). Bayesian hypothesis testing of four-taxon topologies using molecular sequence data. *Biometrics* **52** 193–210.
- SWOFFORD, D.L. (1996). PAUP: Phylogenetic analysis using parsimony and other methods, Sinauer, Sunderland, MA.
- TIERNEY, L. (1994). Markov chains for exploring posterior distributions (with discussion). *Annals of*

Statistics **22** 1701–1762.

- YANG, Z., GOLDMAN, N. and FRIDAY, A. (1995). Maximum likelihood trees from DNA sequences: A peculiar statistical estimation problem. *Systematic Biology* **44** 384–399.
- YANG, Z. and RANNALA, B. (1997). Bayesian Phylogenetic Inference Using DNA Sequences: A Markov Chain Monte Carlo Method. *Molecular Biology and Evolution* **14** 717–724.
- YU, B. (1995). Comment: extracting more diagnostic information from a single run using the cusum path plot. *Statistical Science* **10** 54–58.

DEPARTMENT OF STATISTICS
UNIVERSITY OF WISCONSIN-MADISON
1210 WEST DAYTON ST.
MADISON WI, 53706-1685
NEWTON@STAT.WISC.EDU

DEPARTMENT OF GENETICS
UNIVERSITY OF WISCONSIN-MADISON
445 HENRY MALL
MADISON WI, 53792
ROBERTM@GENETICS.WISC.EDU

DEPARTMENT OF MATHEMATICS AND COMPUTER SCIENCE
DUQUESNE UNIVERSITY
440 COLLEGE HALL
PITTSBURGH PA 15282
LARGET@MATHCS.DUQ.EDU

LIKELIHOODS ON COALESCENTS: A MONTE CARLO SAMPLING APPROACH TO INFERRING PARAMETERS FROM POPULATION SAMPLES OF MOLECULAR DATA

BY JOSEPH FELSENSTEIN, MARY K. KUHNER, JON YAMATO AND PETER BEERLI

Department of Genetics, University of Washington, Box, 7360, Seattle WA 98195-7360

When population samples of molecular data, such as sequences, are taken, the members of the sample are related by a gene tree whose shape is affected by the population processes, such as genetic drift, change of population size, and migration. Genetic parameters such as recombination also affect that genealogy. Likelihood inference of these parameters involves summing over all possible genealogies. There is a vast number of these, so that exact computation is not possible. Griffiths and Tavaré have proposed computing these likelihoods by Monte Carlo integration. Our group is doing this by the Metropolis-Hastings method of Markov Chain Monte Carlo integration. We now have, in our LAMARC package, programs to do this for constant-sized and growing populations, and for geographically structured populations. The bias of the estimator of population growth rate is discussed. One can also allow for samples stratified in time, as with fossil DNA or sequential samples from the population of a virus in a patient. A program for recombining sequences is in progress, and we hope to put together an object-oriented environment which can cope with a variety of evolutionary forces.

1. Introduction. Samples of genes from natural populations of organisms are related by a genealogy, which is usually unknown. At the level of the copies of the genes, such a genealogy would specify where each copy of the gene came from. Thus, a particular copy that we sample may have come from the mother of that individual, from her father, from his father, from his mother, and so on, back in time. Other copies are doing the same. As we go back, occasionally two of these lineages will coalesce, as it happens that two copies of a gene are descended from the same parental copy. Thus, my great-great-great-grandmother might happen to be the sibling of your great-great-great-grandfather, and the genes we possess might then turn out to be copied from the same copy in one of their parents. Such coalescences are inevitable in natural populations.

Figure 1 shows such a pattern of ancestry. Each circle is an individual who has two copies of the gene; we are concerned not just with the genealogy of the individuals, but with the genealogy at the gene level. In the figure, time

Supported by NIH grant R01 GM-51929 and NSF grant BIR-9527687

AMS 1991 subject classifications. Primary 62F03, 92D15; secondary 92D10, 92D20.

Key words and phrases. Coalescent, molecular evolution, genetics.

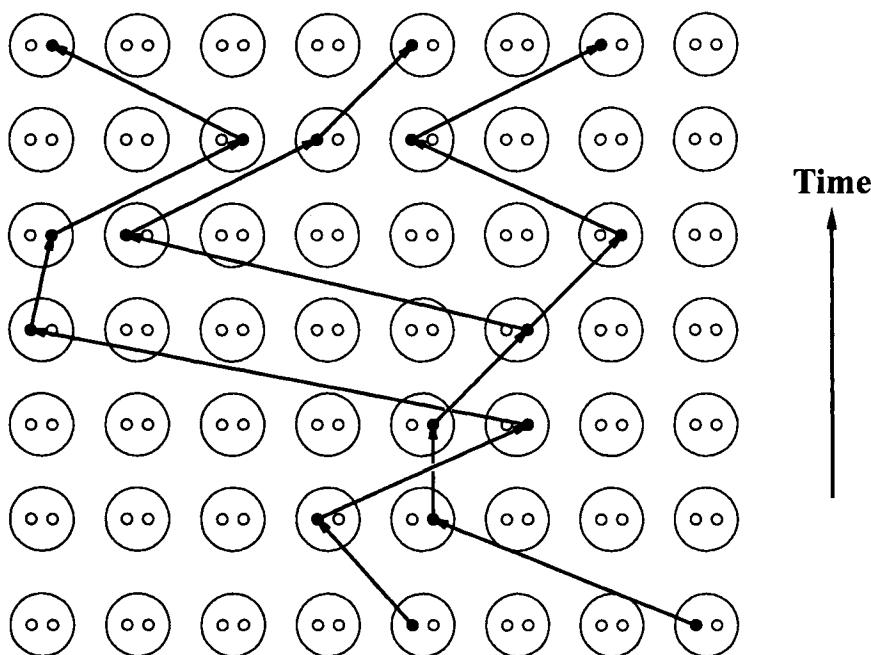


FIG. 1. *A coalescent tree of gene copies that is formed in a diagram showing from which gene in the previous generation each gene copy comes. Large circles are individuals, small circles are copies of genes. Three copies in the current generation trace back to two copies 6 generations earlier.*

flows upwards. The sample consists of three copies of the gene taken from the latest generation (at the top). Arrows show the copies of the gene transmitted from parent to offspring. When we go backwards in time along the arrows, we go downwards, and the lineages gradually coalesce. The rate of this coalescence is higher in small populations than in large ones, simply because the chance that the ancestors of two copies of the gene are the same is greater in a small population.

We have reasonably straightforward models of change in the DNA sequences of such genes, based on the neutral mutation theory of evolution. We can, for example, assume that all sites in the gene change at the same rate μ per generation, according to one of the standard Markov models for base substitution, which specify probabilities of change among the four states A, C, G, and T. If we were to know the genealogy of the copies in detail, statistical estimation of the rates of mutation would be possible, as well as testing of hypotheses about the mutational process. The genealogy is itself the result of a stochastic process, dependent on N_e , the effective population size. This would be the population size if the population reproduced according to an idealized Wright-Fisher model; as it is, it corrects for some departures from such a model. We could imagine

using the genealogy to estimate N_e and test hypotheses about it.

However, we don't know the genealogy. We must therefore integrate over our uncertainty about it. This turns out to confound N_e and μ , and create a large computational problem. In this paper, we will outline the problem, our own Markov Chain Monte Carlo approach, and relate it to the work of Griffiths and Tavaré, who have suggested another Monte Carlo sampling approach. We will also sketch how population growth, migration, recombination, and fossil DNA sequences can be accommodated in our scheme.

2. The coalescent. It has been known since the work of Sewall Wright, in the 1930's, that if we choose two copies of a gene from a random-mating population, the time since their common ancestor is geometrically distributed, with expectation $2N$ generations. (For the moment we use N , the actual population size, as we are dealing with idealized models). As $2N$ is typically reasonably large, it is also well-approximated by an exponential distribution with that expectation. In 1982 Kingman (Kingman 1982a, 1982b, 1982c) generalized this to n copies by defining the coalescent process, and proving that the distribution of the genealogies of the n copies converges to it when scaled properly. While Kingman's methods were sophisticated, the resulting distribution is easy to describe and use. This is fortunate, for Kingman's result is fundamental to the analysis of population samples of DNA sequences.

In the coalescent in a population whose size is N , one can sample from the distribution of the possible genealogies of n copies by the following procedure:

1. Set $k = n$ and $T = 0$.
2. Draw a random quantity u_k from an exponential distribution with expectation $4N/(k(k - 1))$.
3. Pick two of the k copies of the gene at random, without replacement.
4. Create a node of the genealogical tree which is the immediate common ancestor of these two copies, and which existed $T + u_k$ generations before the present.
5. Set $T = T + u_k$.
6. Replace these two copies by this common ancestor and set $k = k - 1$.
7. If $k = 1$ we are done. Otherwise return to step 2.

Thus we go back through a series of exponential time intervals, combining randomly chosen pairs of lineages, until we get a complete tree. The expected time to reach the common ancestor of all copies is $4N(1 - 1/n)$ generations. The

lineages combine rapidly at first, then more slowly as we go back, and the last two are expected to take $2N$ generations to find their common ancestor, more than half of the expected time. An interesting implication of the coalescent is that a sample of modest size has an excellent chance that its common ancestor will also be the common ancestor of all copies of the gene in the population.

Kingman's coalescent is an approximation, valid when $n^2 \ll N$, but it is in practice extraordinarily accurate. Given the departures that real populations show from any of these idealized models, inaccuracy of Kingman's approximation is the least of our worries. Kingman's coalescent defines the prior distribution of genealogies, and has given its name to the whole area: researchers studying ancestry of samples of genes from populations are said to be working on coalescents.

There are many possible departures from the idealized Wright-Fisher model that underlies Kingman's result, but the coalescent is in effect a diffusion approximation. Many different models of reproduction of single populations will have the same diffusion approximation, and hence the same coalescent process, provided we replace the actual population size N by the appropriate effective population size N_e .

3. Likelihoods. The coalescent gives us a prior distribution of the genealogy G' , which has its intervals expressed in generations. As a product of exponential densities, it is easily written down and easily computed. Its density function is

$$(3.1) \quad f(G'|N_e) = \prod_{k=2}^n \frac{2}{4N_e} \exp\left(-\frac{k(k-1)}{4N_e} u_k\right)$$

where u_k is the length of the interval during which the genealogy G' has k lineages. If we were able to observe the coalescence intervals u_k , we could estimate N_e . Note that the event that actually occurs brings in a factor of $2/(4N_e)$ rather than $k(k-1)/(4N_e)$ as we know which two lineages have coalesced. The product of these factors of $2/(k(k-1))$ represents the probability of sampling the particular "labelled history" (Edwards 1970) from among all those possible.

Of course, we do not actually observe coalescence intervals. For most kinds of contemporary data, we can observe only the differences between the members of our sample. For example, for DNA sequences, we can see the number of positions (sites) at which the molecules differ. That gives us a picture of the coalescence times, but only a clouded picture. We need to make inferences about parameters such as N_e by using a model of the change in the DNA. The notion of a molecular clock provides such a model. We assume a Markov process operating independently at each site in the DNA, with a mutation rate μ .

By equating long-term change to mutation, we are implicitly basing ourselves on the neutral mutation model of evolution made famous by Motoo Kimura (Kimura 1968, Kimura 1983). We can use a stochastic model of DNA change, and make assumptions of independence of change in different sites and in different lineages, to compute the probability of the observed sequences D given a genealogy G' . One of us (J.F.) has outlined how to do this (Felsenstein 1981) and Ziheng Yang, Gary Churchill, and he have more recently shown how to incorporate autocorrelated variation of evolutionary rates from site to site using a Hidden Markov Model approach (Yang 1993, 1994, 1995; Felsenstein and Churchill 1996).

We cannot be certain of the genealogy G' . In fact, it is the role of the data to illuminate it, however dimly. To compute the likelihood of the coalescent parameter N_e and the mutation rate μ given the data D , we must integrate over all possible genealogies (Felsenstein 1988, 1992)

$$(3.2) \quad \text{Prob}(D|N_e, \mu) = \int_{G'} f(G'|N_e) \text{Prob}(D|G', \mu).$$

We describe the integration below. The probability of D given G' and μ which appears on the right is the probability calculated by our Markov process model of evolution, the same quantity that is computed in maximum likelihood inference of phylogenies. The quantity μ is a rate of mutation per generation; in more complex cases this may be replaced by several parameters.

Neither of the terms inside the integral in equation 3.2 is hard to compute. The quantity f is given by 3.1 and the other probability requires effort proportional to the total number of DNA bases in our sample, times the square of the number of states at a site, which is 4. The computational problem comes from the vast size of the space of genealogies G' . The space of values of G' is a union of a very large number of Euclidean spaces. Edwards (Edwards 1970) enumerated these: they are his “labelled histories”. With n sequences there are $n!(n-1)!/2^{n-1}$ of them, so that with only 10 sequences there are 2.571×10^9 labelled histories. Each one of these has $n-1$ node times. The integration in 3.2 must be over all values of these, so that each of these billions of terms integrates over $n-1$ dimensions. Clearly there is a computational problem here.

All attempts to find mathematical simplifications for this integration have so far failed. Nevertheless two groups - Griffiths and Tavaré and ourselves - have attempted to use Monte Carlo integration. This can work because many of the billions of possible labelled histories make rather little contribution to the integral, because they lead to very low values of the term $\text{Prob}(D|G')$. We will describe our approach first, and then show the relationship between the two approaches, which appear at first sight to be quite different.

4. A Metropolis-Hastings approach. Our approach has been to use Markov Chain Monte Carlo sampling, in particular the Metropolis-Hastings method (Metropolis et al. 1953; Hastings 1970). We want to sample from the terms of 3.2 using importance sampling, with our importance function being as close as possible to the the that is being integrated. Our approach for the simplest case - a single population of constant size, with no recombination - is outlined by Kuhner et al. (Kuhner et al. 1995, 1997).

In that case, it turns out that we can change the time scale of the genealogies. The entities G' have their node times given in generations. Instead we can rescale them to be in units of $1/\mu$ generations, where μ is the underlying neutral mutation rate of the DNA model that we use. Thus if a node in the genealogical tree is 100,000 generations ago, and the underlying mutation rate μ is 10^{-7} , when rescaled the node is 0.01 mutations ago. These are of course expected mutations per site, not actual mutations. Informally, we can write this by saying that the genealogy is now G rather than G' , and

$$(4.1) \quad G = \mu G'.$$

The result of this change of variables is of course to alter the density f as well. The coalescence intervals u_k in 3.1 are replaced by $v_k = \mu u_k$, and a factor of $1/\mu$ comes into each term in the resulting density as well. The result is the density:

$$(4.2) \quad g(G|\Theta) = \prod_{k=2}^n \frac{2}{\Theta} \exp\left(-\frac{k(k-1)}{\Theta} v_k\right)$$

where $\Theta = 4N_e\mu$. This resembles closely the widely-used parameter θ that is frequently estimated in evolutionary genetics, except that it contains the neutral mutation rate per site rather than per locus.

The result of this change of scale is that the probability $\text{Prob}(D|G', \mu)$ can be replaced by $\text{Prob}(D|G)$, as the branch lengths of G are already multiplied by the mutation rate. In most DNA models, the elapsed time t in generations must be multiplied by a rate of mutation μ before it can be used. If we are given the product μt we can compute the transition probability directly from it. The result is that 3.2 now becomes:

$$(4.3) \quad \text{Prob}(D|\Theta) = \int_G g(G|\Theta) \text{Prob}(D|G).$$

If there were more parameters than μ , one would have to change $\text{Prob}(D|G)$ by adding ratios of parameters, such as $\text{Prob}(D|G, \mu_2/\mu_1)$. Our objective becomes computing the likelihood of the parameter Θ .

To approximate the integral we take as our importance function the quantity $g(G|\Theta) \text{Prob}(D|G)$, which immediately raises the issue of what value of Θ to use. Ideally one would want to sample at the maximum likelihood value of Θ , but we cannot know in advance what this will be. Our strategy has been to make a rough estimate of Θ , which we call Θ_0 , and use that for an initial sampling, sampling from $g(G|\Theta_0) \text{Prob}(D|G)$. We sample genealogies G_1, G_2, \dots, G_m by taking an initial genealogy and making successive alterations to it, doing acceptance/rejection sampling appropriately according to a Metropolis-Hastings algorithm. This forms a Markov chain of genealogies. We use that for an initial sampling, then find a maximum likelihood value based on the sample from that first Markov chain. This is then taken as the provisional value for a second Markov chain, and so on. We have usually run 10 of these chains, then two much longer ones at the end. The final likelihood curve is computed from the second of these long chains. In our programs the user can customize the number and lengths of the chains.

5. Importance sampling and likelihood curves. One useful property of Metropolis-Hastings sampling is that we can estimate the whole likelihood curve from a single run of a Markov chain, rather than having to compute each point on the likelihood surface from a separate run. Suppose that we sample m genealogies from a Markov chain which has its equilibrium distribution proportional to $g(G|\Theta_0) \text{Prob}(D|G)$. Call the sampled genealogies the G_i . The usual importance sampling formula for Monte Carlo integration gives:

(5.1)

$$\int_G g(G|\Theta) \text{Prob}(D|G) \simeq \frac{1}{m} \sum_{i=1}^m \frac{g(G_i|\Theta) \text{Prob}(D|G_i)}{g(G_i|\Theta_0) \text{Prob}(D|G_i)} = \frac{1}{m} \sum_{i=1}^m \frac{g(G_i|\Theta)}{g(G_i|\Theta_0)}$$

this allows us to estimate the likelihood for other values of Θ from a run of the Markov chain at Θ_0 . Note that the likelihood curve depends only on the Kingman priors of the sampled G_i at Θ and at Θ_0 . This makes it seem that the data are not involved at all; they actually affect the Markov Chain Monte Carlo sampling process and affect the final likelihood through their effect on which G_i are sampled.

6. The Markov Chain sampling. Our samples of the genealogies G must come from a distribution proportional to $g(G|\Theta_0) \text{Prob}(D|G)$. We achieve this through a sampling based on conditional coalescents. A conditional coalescent may be described as a distribution on G that has its density proportional to the coalescent density $g(G|\Theta_0)$ on some domain of G 's, and has density 0 elsewhere. In our programs the conditional coalescents are created by a process of dissolving

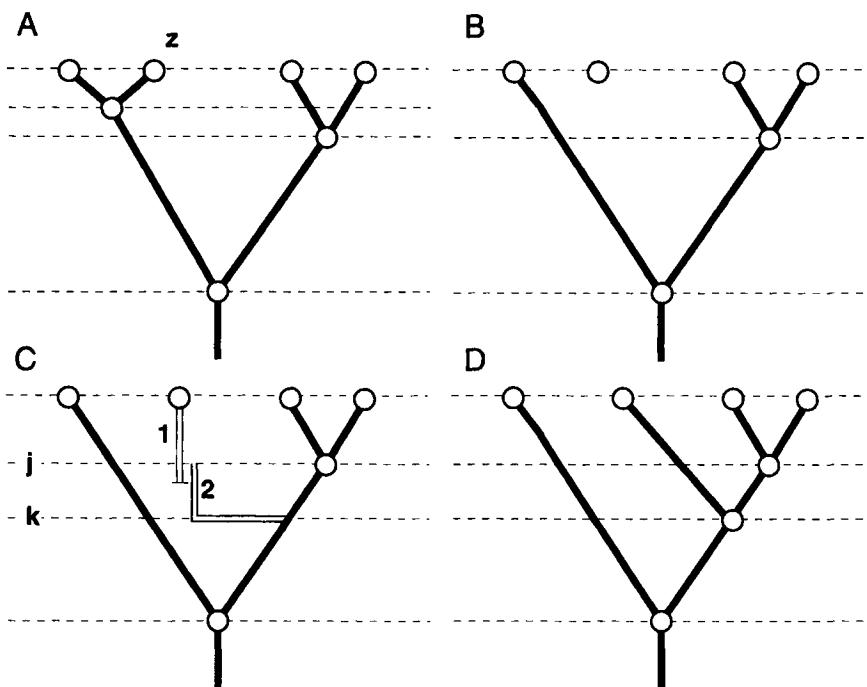


FIG. 2. *A conditional coalescent method of altering a tree. A single lineage is chosen at random to be altered (the lineage below *z* in A). It is removed from the tree (B) and then its coalescence with the remaining lineages is simulated (C). Tree D shows the result.*

part of a tree, and reforming that part by allowing lineages to sample their ancestry randomly according to a conditional coalescent. In the original paper by Kuhner et al. (1995), the region of the tree that was dissolved had a single lineage at its base and three lineages at its top. The three lineages, which were not necessarily contemporaneous, were then re-formed into a tree by allowing them to coalesce, but requiring that all three coalesce into a single lineage by the time the base of the dissolved region was reached. The details of how this was done will be found in that paper.

More recently, we have changed to a different conditional coalescent suggested by Peter Beerli. In this, a lineage is selected, and is disconnected from the genealogy, with the lineage being dissolved back up the tree to the next highest coalescent node. It is then allowed to sample its ancestry downwards (backwards in time) until it re-connects to the tree. Note that sometimes this will mean it reconnects below the previous root of the tree. Figure 2 shows this process in a single population.

A branch of the tree is chosen at random. In this case it is the one below tip *z* (tree A). Tree B shows the tree with that branch removed. In tree C we see

the process of simulating the conditional coalescence of that lineage with the remaining ones. During the topmost interval of the tree (the time down to line j), the instantaneous rate of coalescence of that lineage with each of the three others is $2/\Theta_0$, for a total of $6/\Theta_0$. We generate an exponential random variate with mean $\Theta_0/6$, which is the time until coalescence of that lineage with one of the three others. In this case (line 1 in tree C) the time is too long, and takes the lineage past line j . We then consider the lineage to have remained distinct back as far as line j . Starting at that time, we have two other lineages, for a total instantaneous rate of coalescence of $4/\Theta_0$. We then draw an exponential variate with mean $\Theta_0/4$. This time, which defines line k , turns out to be a time above the next coalescence, which is the root of the tree. So we connect our new lineage to the tree at the time of line k , choosing one of the two lineages as the one to which it will be connected. The resulting tree is D.

Note that it is possible for any of the lineages, other than the one that is below the root, to be chosen to be dissolved, and it may reconnect to the tree below the original root. The method requires one exponential variate to be generated for each coalescence interval on the remaining part of the tree. If there are m other lineages in an interval, the instantaneous rate of coalescence with them is $2m/\Theta$.

Having proposed this change, we decide whether to accept it. The method of generating the new tree is a conditional coalescent, which means that if the old tree is G_{old} and the new tree G_{new} , then

$$(6.1) \quad \text{Prob}(G_{new}|G_{old}) = K \text{Prob}(G_{new}|\Theta_0)$$

for some constant K , as the density from which G_{new} is drawn is proportional to the coalescent density. An analogous equation holds for $\text{Prob}(G_{old}|G_{new})$. In constructing the rule for acceptance and rejection, we use these in the Hastings ratio terms, accepting the new tree if a uniform random fraction r satisfies

$$\begin{aligned} r &< \frac{\text{Prob}(G_{old}|G_{new})}{\text{Prob}(G_{new}|G_{old})} \frac{\text{Prob}(G_{new}|\Theta_0)}{\text{Prob}(G_{old}|\Theta_0)} \frac{\text{Prob}(D|G_{new})}{\text{Prob}(D|G_{old})} \\ (6.2) \quad &< \frac{\text{Prob}(G_{old}|\Theta_0)}{\text{Prob}(G_{new}|\Theta_0)} \frac{\text{Prob}(G_{new}|\Theta_0)}{\text{Prob}(G_{old}|\Theta_0)} \frac{\text{Prob}(D|G_{new})}{\text{Prob}(D|G_{old})} \\ &< \frac{\text{Prob}(D|G_{new})}{\text{Prob}(D|G_{old})}. \end{aligned}$$

Thus the conditional coalescent causes cancellation of the Hastings terms and the Kingman prior term, leaving only the ratio of the likelihoods of the trees. These would be the likelihoods of these genealogies, given the data, if the genealogies were treated as parameters (which they are not). The machinery

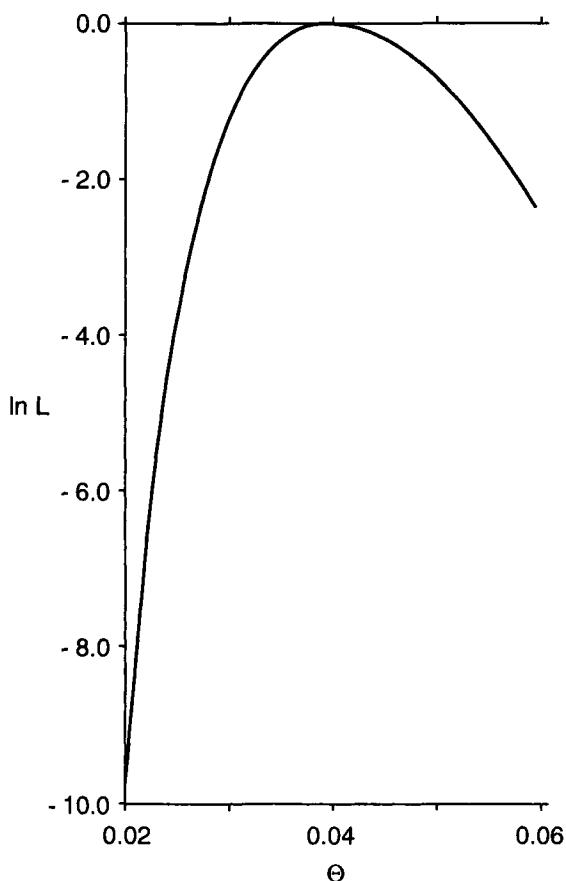


FIG. 3. The log-likelihood curve for Θ for the data of Ward et al.

to compute likelihoods on genealogies is the same as it is on phylogenies, and it is well-enough known (e.g. Felsenstein 1981) not to need to be treated here. Note that we can use any type of data for which such likelihoods are available, including DNA sequences, microsatellite copy numbers, restriction sites, and even isozyme mobilities. Note also that we have only modified part of the tree, so that we need only recalculate the likelihoods for the parts of the two trees that differ, a considerable saving. The rearrangement strategy described here has some similarity to that used by Li et al. (1998) but their strategy dissolves only branches leading to tips, and does not use the conditional coalescent for reattachment.

As an example, Figure 3 shows the likelihood curve generated by a run of on the mitochondrial DNA data set of Ward et al. (1991). The estimate of Θ is 0.0396. Taking an interval two units of log-likelihood below the maximum

suggests that the estimate lies between about 0.03 and 0.055. This curve was generated by two long chains of 12,000 steps each, sampling trees every 20 steps. Further details are given by Kuhner et al. (1995).

The method is computationally feasible on workstations or fast desktop computers. Computational effort seems to rise slowly with the number of sequences, especially since we can re-use many of the likelihood computations from one tree to the next. If only part of a tree has changed we can re-use the likelihoods from the rest of the tree. However there are no easy generalizations about how long the Markov chains must be run.

7. The method of Griffiths and Tavaré. Our Monte Carlo sampling approach was preceded by the pioneering and innovative method of Griffiths and Tavaré (1994a, 1994b, 1994c). At first sight their method appears to bear no relation to ours, and to have considerable advantages over it. A close examination shows that the two methods are related, and makes clear the advantages and disadvantages of our approach.

Griffiths and Tavaré have as their objective the same likelihood function that we compute. They form a system of recurrence equations expressing this likelihood in terms of likelihoods for data sets that have resulted from one fewer evolutionary event. In principle, recursive evaluation of these equations, as in an earlier paper by Griffiths (1989), will yield the desired likelihood. However, the recursion expands rapidly, and one must therefore use some approximate method of evaluating it. Griffiths and Tavaré (1994a, 1994b, 1994c) choose sample paths down through the recursion randomly. The great advantages of this method are that the computations are rapid, and each such sample path is independent of all the others. By contrast, our samples are autocorrelated, leading to serious problems knowing how long to continue the sampling. In each of our samples, the likelihood of a tree must be computed. Even if parts of the computation can be re-used, this is much more effort than is needed for their method.

Each step in their sampling goes back one level in the recursion, and amounts to a decision as to what the next most recent event in the genealogy is. The sequence of choices that Griffiths and Tavaré make corresponds to a sequence of events in evolution. Going backwards in time, their events are mutations and coalescences, plus choices of the ancestral nucleotides at each site. Figure 4 shows such a history of events leading to a set of four DNA sequences. It corresponds to one path through their recursion. Note the difference between such a history (H) and the genealogy (G) that we sample. Our genealogy has branch lengths; theirs does not, at least in the simplest case. They specify the place of occurrence of each mutation, while our likelihoods must sum over all

possible placements of mutations on the tree. Nevertheless, we can regard their method as Monte Carlo integration. We can make an equation analogous to our equation 3.2:

$$(7.1) \quad L = \text{Prob}(D|\Theta) = \sum_H \text{Prob}(D|H) \text{Prob}(H|\Theta),$$

where H is a history of events, corresponding to a sequence of choices in Griffiths and Tavaré's recursion. The histories that they sample have the property that they must always lead to the observed sequences. Thus $\text{Prob}(D|H)$ is, trivially, always 1. The term $\text{Prob}(H|\Theta)$ is simply the product of probabilities of the individual events in H . In the history shown in Figure 4, the most recent event could have been a mutation in any of the 11 sites in any of the four sequences, and each could have come from any of three other nucleotides. The particular event that is shown is a coalescence. There are only two sequences (1 and 3) that are identical, and thus could have coalesced at this point. The rate of coalescence for this pair will be $1/(2N_e)$. The next event is a mutation. If we use, for simplicity, a symmetric Jukes-Cantor model of evolution, the rate of occurrence of a particular mutation from C to A at a particular site will be $\mu/3$, where μ is the total mutation rate per site.

Consider all possible histories H , those that lead to the observed sequences as well as those that do not. As the most recent event, there are $4 \times 11 \times 3$ possible mutations, and $4 \times 3/2$ possible coalescences. The fraction of this probability contributed by the most recent coalescence in Figure 4 is then $(1/(2N_e))/(6/(2N_e) + 44\mu)$, which turns out to be $1/(6+22\Theta)$. Continuing in this fashion we can calculate the probability $\text{Prob}(H|\Theta)$ of the particular sequence of events in Figure 4 to be

$$\left(\frac{1}{6+22\Theta}\right) \left(\frac{\Theta}{18+99\Theta}\right) \left(\frac{\Theta}{18+99\Theta}\right) \left(\frac{2}{6+33\Theta}\right) \left(\frac{1}{1+11\Theta}\right) \left(\frac{1}{4}\right)^{11}.$$

The last term is the probability that the initial DNA sequence is as shown in Figure 4. In effect what Griffiths and Tavaré do is to sum over all such histories, adding up this quantity for all those that lead to the observed data.

Griffiths and Tavaré at each stage are considering all possible most recent events that could have led to the observed sequences. They use importance sampling, by sampling at each stage from among the possible events in proportion to their rate of occurrence. Thus at the first stage in the above calculation, they choose among the one possible coalescence and the 33 possible mutations in proportion to the contributions each would make to the numerator (in that case $1/(2N_e)$ versus $\mu/3$). This needs the usual importance sampling correction.

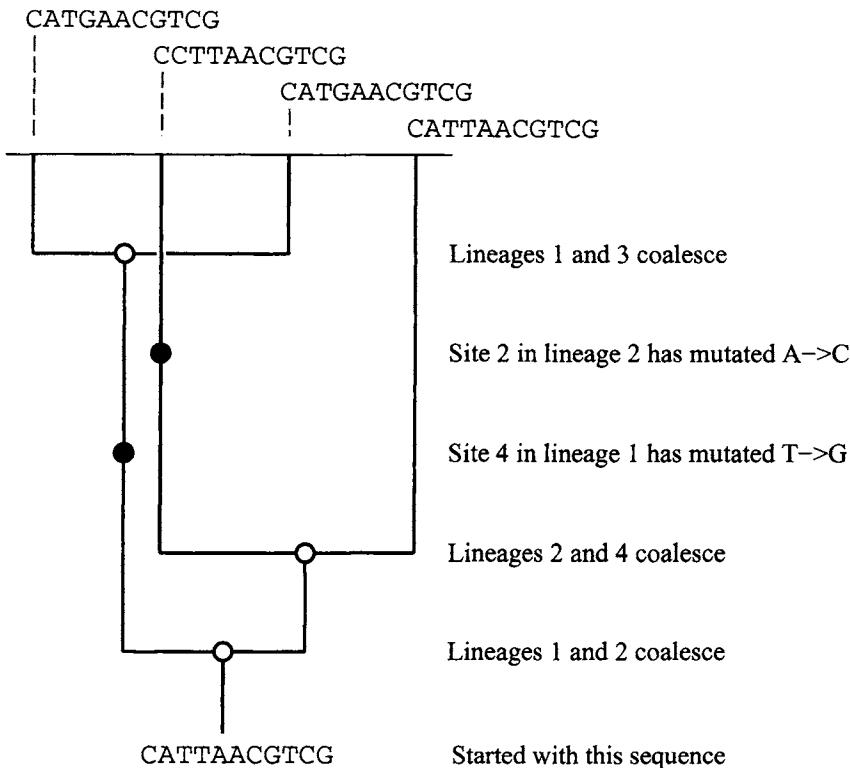


FIG. 4. *A history of mutation, coalescences, and ancestral nucleotide choices that could result in a given set of four sequences. Such histories are, in effect, what Griffiths and Tavaré's method samples. The events are described from a point of view looking backwards in time from the present.*

Their sampling is done, as ours is too, at a trial value Θ_0 . Suppose that f is the probability $\text{Prob}(H|\Theta)$, unconditioned on the data, and h is the probability for the distribution from which we sample instead. The importance sampling correction is

$$(7.2) \quad L(\Theta) = \mathcal{E}_f [\text{Prob}(D|H)] = \mathcal{E}_h \left[\frac{f}{h} \text{Prob}(D|H) \right]$$

and since for h we always have $\text{Prob}(D|H) = 1$, the likelihood is just the expectation over h of f/h .

A history H consists of a series of choices. Suppose that history H_i has at stage j a series of possibilities, with the terms of the Griffiths/Tavaré recursion being the $a_{ijk}(\Theta_0)$. Suppose that one that is actually chosen in history H_i has term $b_{ij}(\Theta_0)$. Then the probability of having taken this choice is

$$(7.3) \quad \frac{b_{ij}(\Theta_0)}{\sum_k a_{ijk}(\Theta_0)}$$

and the probability of the history is the product of this ratio over all j , the number of these depending on the number of events in history H_i . This is the expression for h . The distribution f is similar except that it has Θ in place of Θ_0 , and a wider range of possible events, including those which conflict with the data. The full set of events at stage j in this distribution we call the $c_{ijk}(\Theta)$.

We end up with

$$(7.4) \quad L(\Theta) = \mathcal{E}_h \left[\frac{\left(\frac{\Pi_j b_{ij}(\Theta)}{P_{ij}(\sum_k c_{ijk}(\Theta))} \right)}{\left(\frac{\Pi_j b_{ij}(\Theta_0)}{\Pi_j(\sum_k a_{ijk}(\Theta_0))} \right)} \right] = \mathcal{E}_h \left[\prod_j \frac{b_{ij}(\Theta)}{b_{ij}(\Theta_0)} \frac{\sum_k a_{ijk}(\Theta_0)}{\sum_k c_{ijk}(\Theta)} \right]$$

Griffiths and Tavaré's method consists of sampling from h to approximate this expectation by averaging the ratio on the right. A careful reading of their papers will show that the above expression is precisely what they compute. Thus their method too can be considered a Monte Carlo integration method with an importance function.

Given the independence of their samples, and the rapidity with which they can compute them, one might expect their method to be unequivocally superior to ours. We are, after all, burdened by more computation and autocorrelated samples. The difficulty with their method is that the distribution h from which they sample does not sample from the histories in proportion to their contribution to the likelihood. There is thus some wasted effort. By contrast our Metropolis-Hastings sampling is supposed to sample from genealogies in proportion to their contribution to the likelihood. We thus have reason to hope that our method might do better in some cases. The problem is most easily seen when considering how Griffiths and Tavaré's method will handle two DNA sequences. If those sequences happen to differ by (say) 2 bases, the mutational events that are sampled will include not only the precise changes needed to make the two sequences identical, but also all other changes in all other sites. Thus a great deal of sampling may be needed to sample from the events that contribute most of the likelihood. Griffiths and Tavaré (Griffiths and Tavaré 1994c) have worried aloud about this very issue.

8. Population growth. The model of an isolated population of constant size can be extended by allowing the population to grow exponentially. Griffiths and Tavaré (1994a) have done so, and so have we (Kuhner et al. 1998). Our program **FLUCTUATE** is currently in distribution. In a population of effective size $N_e(t)$ with k lineages, the rate of coalescence is $k(k-1)/(4N_e(t))$. If the effective population size grows exponentially at rate r , then when t is the time back from

the present (“dual time”),

$$(8.1) \quad N_e(t) = e^{-rt} N_e(0)$$

Taking this into account in the time to coalescence, that density function is (Kuhner et al. 1998)

$$(8.2) \quad f(t) = e^{[-\frac{k(k-1)}{4N_e(0)r}(e^{rt}-1)]} e^{rt} \frac{2}{4N_e(0)}.$$

This can be used to make a counterpart to equation 3.1 straightforwardly. Griffiths and Tavaré (1994a) have used this for joint likelihood inference of the current value of Θ and the growth rate. We have more recently produced a Metropolis-Hastings algorithm (Kuhner et al. 1998) for a similar model.

Once the mutation rate μ is introduced and the branch lengths of the genealogical trees rescaled in units of expected mutations per site, the parameters of the likelihood turn out to be the current value of $4N_e(0)\mu$, called Θ , and the growth rate per unit branch length, which is $g = r/\mu$. The likelihood surfaces in these parameters usually contain long, narrow ridges. At any given value of g , the estimation of Θ is reasonably accurate, but there is usually a long, narrow, slightly curving ridge whose top is nearly flat. It runs nearly parallel to the g axis, but curving gradually upwards as higher values of g are reached.

There turns out to be surprisingly little power to estimate g , except in cases where the true value of g is large. Even more surprising is the strong bias in the estimate of g . When data sets are generated from a model that has no population growth, they much more often cause us to estimate a large positive g than a negative g . The behavior is so startling as to make us wonder whether it simply be the result of a program bug.

We can verify that the bias is real by using 8.2, and considering the case of a sample of size 2 ($n = 2$). Suppose that we had very long, nonrecombining sequences. That would allow us to make a precise estimate of the rescaled time $T = \mu t$ to coalescence. The likelihood function can be written in terms of g and $\Theta = 4N_e(0)\mu$.

$$(8.3) \quad \text{Prob}(T|\Theta, g) = e^{[-\frac{2}{\Theta g}(e^{gT}-1)]} e^{gT} \frac{2}{\Theta}.$$

In the case of a sample of size 2, let us assume that Θ is known, and set in 8.3 to its true value, and that we are estimating g . There is no explicit formula solving for the maximum likelihood estimate \hat{g} in terms of T , but the likelihood can be maximized numerically. Now imagine a population whose true growth rate is zero, and whose value of Θ is known to be 1. The scaled coalescent time T for sample size 2 will be distributed exponentially with mean 0.5.

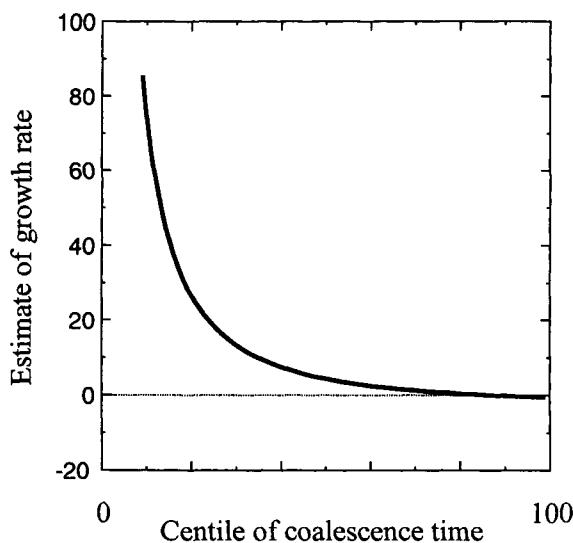


FIG. 5. *Estimates of growth rate for $n = 2$ in a data set with a large number of sites, so that coalescence time can be estimated accurately. For a case where Θ is known and the true growth rate is 0, the estimates for different quantiles of the coalescence time are shown. A large bias toward inferring growth is apparent.*

In Figure 5, the maximum likelihood estimate \hat{g} is shown for quantiles of that distribution. It is striking that 87% of the time the estimate is positive, and very strongly positive for small coalescent times (below 0.08 the curve is too high to fit onto this figure). The other 13% of the time the estimate is negative, though only moderately so. The bias in \hat{g} can be seen: it is the average height of the curve, which is strongly positive. Note that the growth rate scale means that $g = 20$ implies growth of the population by a factor of e^{10} during the expected time for two samples to coalesce. Even at the median of the coalescence times, the bias implies that we infer growth by a factor of $e^{2.166}$ during the average coalescence time. As our Metropolis-Hastings algorithm is not used here, this calculation is an independent check of the reality of the bias. This bias sounds like a serious problem for Monte Carlo integration methods. It is, but we are convinced that it is an equally serious problem for all other methods. However, although the point estimates are biased, if we make interval estimates using the usual chi-square approximation to the distribution of the likelihood ratio, accepting all values of g whose log-likelihood is within 3 units of the peak (in the more general case where two parameters, g and Θ , are being estimated), the true value of 0 is within the interval almost 95% of the time. In this case (S. Tavaré, pers. comm.) the chi-square distribution is of dubious propriety, as it has an asymptotic justification but is being used on data from a single locus.

Nevertheless, the interval based on it seems to behave appropriately. In addition, the bias becomes much smaller as we add data from more loci (Kuhner et al. 1998).

9. Migration. We can also extend the model to allow for multiple populations exchanging migrants. This has been done by Nath and Griffiths (1996), who estimate the migration rates for populations whose values of Θ are known. We have extended our Metropolis-Hastings method to a two-population case, to estimate the two values of Θ and two migration rates (Beerli and Felsenstein 1999). This seems to have advantages over methods using statistics like F_{ST} , as those cannot estimate all four parameters independently. An extension to n populations is in progress.

10. Sequential sampling. In studies of ancient DNA, we have samples that are not contemporaneous. In studies of the course of viral infection in a patient (as in HIV) one may also have sequential samples. The coalescent likelihood approach is readily adapted to such cases (Rodrigo and Felsenstein 1998). Suppose that we have a genealogical tree G^* whose branch lengths are actual times, with some tips not contemporaneous. Let the generation time of the organism be τ . As one proceeds down the tree, there are two possible events, the entry of a new sample (which has probability 1 at certain known times) or a coalescence. In place of equation 3.1, we have a product of terms, the j -th of which is either

$$(10.1) \quad \frac{2}{4N_e} \exp \left[-\frac{k_j(k_j - 1)}{4N_e} \frac{u_j}{\tau} \right]$$

or

$$(10.2) \quad \exp \left[-\frac{k_j(k_j - 1)}{4N_e} \frac{u_j}{\tau} \right]$$

depending on whether there is a coalescence or a new sample at the bottom end of interval j . Note that k_j is the number of lineages that exist in the genealogy during interval j . Note also that the chronological lengths of the interval have been divided by the generation time to convert them into generation times. The probability of the data given G^* also needs a conversion: it depends on the product of the per-generation mutation rate per site, μ and the generation time elapsed, which is t/τ .

The result is that we can restate equation 3.2 as

$$(10.3) \quad \text{Prob}(D|N_e\tau, \mu/\tau) = \int_{G^*} f(G^*|N_e\tau) \text{Prob}(D|G^*, \mu/\tau).$$

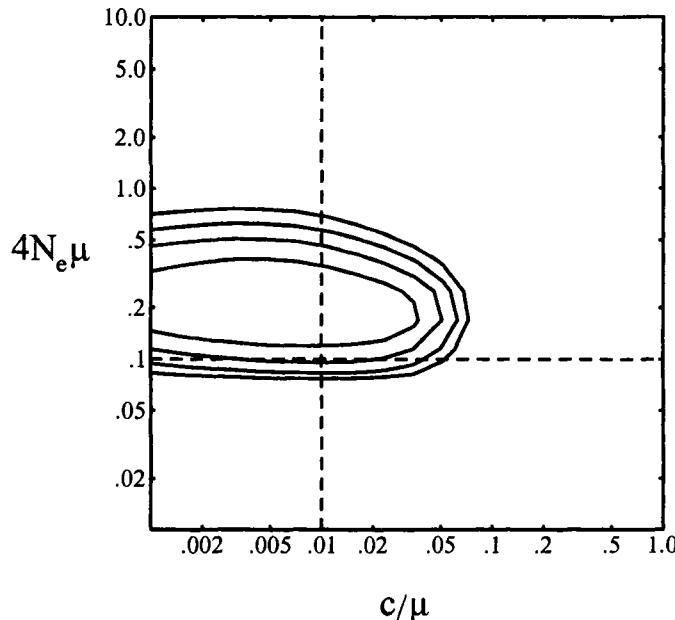


FIG. 6. *Contours of the likelihood surface from a single run of the RECOMBINE Metropolis-Hastings sampler with recombination on simulated data. The axes are $\Theta = 4N_e\mu$ and the recombination parameter c/μ . The contours shown are 1, 2, 3, ... units of log-likelihood below the peak. The true values of the parameters are shown by the dashed lines. In this run the value of Θ is a bit higher than the truth, but the true parameters lie well within the contour of 3 log-likelihood units below the peak, which defines their approximate confidence limits.*

so that the two parameters that can be estimated are $N_e\tau$ and μ/τ . This means that if we know the generation time τ we can estimate N_e and μ separately. Alternatively if (for example) we know μ , we can estimate N_e as well as the generation time τ . Note that the integration over G^* would involve all possible labelled histories and coalescent times, but would not alter the times at which the samples were taken, these being assumed known.

11. Recombination. All of the above cases involve sequences with no recombination. They are thus appropriate for mitochondrial DNA but of dubious value in the nuclear genome. For this reason it has been of great interest to everyone involved with coalescent likelihood methods to have a way of dealing with recombination. As usual, we have come in second in the race, as Griffiths and Marjoram (1996) have an algorithm that infers the likelihood of a sample with two parameters, $4N_e\mu$ and $4N_e c$, where c is the recombination fraction per

site. Their method requires substantial computation to adequately sample the histories. Their method makes use of an “ancestral recombination graph” (Griffiths and Marjoram 1997) originally described by Hudson (1990). This shows coalescences and recombination events. The latter branch as one goes rootwards, and at each such branching one needs to specify which sites take each of the two routes.

We have also produced a program for inferring these two parameters (manuscript in preparation). Although the Metropolis-Hastings approach helps concentrate the sampling on the relevant genealogies, the number of these is so large that the computation is still slow. Figure 6 shows contours of a likelihood surface produced in one of our runs.

There are serious problems ahead, as we need to know how long to run the Markov chains to get an accurate answer, and this is generally unknown. However there are also opportunities. One involves using these methods to place a firm likelihood foundation under the widely used genetic mapping method known as linkage disequilibrium mapping. A start has been made on this by Rannala and Slatkin (1998) and Graham and Thompson (1998); our methods can be used to treat the problem more generally.

12. Natural selection. Until recently it was assumed by everyone that one could not specify the coalescent for sequences that were under natural selection. Only some special cases could be solved, for cases of extreme selection (Kaplan et al. 1988, Hudson and Kaplan 1988, Takahata 1990). Recently Neuhauser and Krone (Neuhauser and Krone 1997, Krone and Nuehauser 1997) made major inroads into the problem, in what are perhaps the best papers on the coalescent since Kingman. They defined a diagram that branches both downward and upward. Unlike the similar diagrams that are produced in cases of recombination, these do not have different alleles following different loops. Instead information flowing upward on the genealogical graph can only pass through certain branches if the genotype contains one of the selected alleles. In the case of recombination, at each site the graph is a tree, although not the same tree at all sites. In the Neuhauser-Krone “ancestral selection graph” the loops are rather more serious. If one tries to compute $\text{Prob}(D|G)$ on them, likelihood must be propagated simultaneously, not independently, around both sides of a loop. However, they were able to define a recursion system that could be evaluated by Griffiths and Tavaré’s method.

Neuhauser and Krone’s work is an enormous and stimulating advance. But it seems ill adapted to our Metropolis-Hastings approach, since when the selection coefficient favoring haploid genotype A_1 over genotype A_2 is s , for moderate values of $2N_e s$ the number of loops in the ancestral selection graph can become

large. Stimulated by it, J.F. has started work on a different method, which involves carrying out Metropolis-Hastings simulation of the frequencies of the selected alleles, as well as the coalescent of other alleles within those alleles, and the "migration" between them that is caused by mutation and recombination. There are no results to report yet.

13. Software distribution. Our package LAMARC (which stands for Likelihood Analysis by Metropolis Algorithm for Random Coalescents) is available free from its Web site:

<http://evolution.genetics.washington.edu/lamarc.html>

as C source code plus PowerMac and Windows executables. It is readily compiled on workstation C compilers (except for the cc compiler on SunOS systems). As of this writing four programs were in distribution: COALESCE, which analyzes a single population of constant size, FLUCTUATE, which analyzes exponentially growing single populations, MIGRATE, which analyzes two populations exchanging migrants, and RECOMBINE, which analyzes a single population of constant size with recombination. More programs and more features will probably be available by the time you read this.

14. An object-oriented fantasy. Even if we could solve some of the problems of how long to run the Markov Chains, the sampling approach has one other serious problem. We like to call it "the 2^8 programs problem". Each one of these Markov Chain Monte Carlo programs is enormously difficult to write. It takes each of us about 2 years to write and debug one of them. And yet, the present programs are highly limited. We have programs that add one complication (population growth, migration, recombination) but do not combine these in the same program. And yet there are more complications (such as natural selection, speciation, and gene conversion) that need to be considered. Any user may want to pick some particular combination of, say, 8 complications. Do we need to resign ourselves to spending the next 500 years writing all possible programs?

There is one way out. Object-oriented programming methods (such as are embodied in C++, Objective C and Java) allow a program to self-assemble in response to a user's requirements. We therefore intend to try to create such an environment. The user would select which combination of evolutionary forces, historical events, genetic situations, and population structure were needed. The program would then use only those classes and subclasses needed to run that particular combination. Thus more like 8 programs than 2^8 need to be written. The issue of chain length remains, and we as yet have no experience with the serious issue of user interface - how do we represent the results of runs that have

many parameters, for example?

Nevertheless one may fantasize about an “evolutionary genetics black box”. The user puts in the data and the model, and out come likelihood inferences about the parameters. One still needs to know population genetic theory, of course, to comprehend the model. But a large fraction of the kind of work that has filled theoretical journals in population genetics may become obsolete if this fantasy can be realized. Many papers start with a theoretical model, pose the question of what is the expected value of some statistic (such as the probability of monomorphism, or of fixed differences between populations, or the variance of heterozygosity), and after much blood, sweat, and tears arrive at a power series, which usually remains unused by those with data. We may hope that this era can be succeeded by one where the same effort can be redirected to formulating the model and improving the computational methods.

Acknowledgments. We wish to thank Dick Hudson, Robert Griffiths, Simon Tavaré, and Kermit Ritland for helpful communications.

REFERENCES

- BEERLI, P. and FELSENSTEIN, J. (1999). Maximum likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. *Genetics*, in press.
- EDWARDS, A. W. F. (1970). Estimation of the branch points of a branching diffusion process. *Journal of the Royal Statistical Society B* **32** 155–174.
- FELSENSTEIN, J. (1981). Evolutionary trees from Dna sequences: a maximum likelihood approach. *Journal of Molecular Evolution* **17** 368–376.
- FELSENSTEIN, J. (1988). Phylogenies from molecular sequences: inference and reliability. *Annual Review of Genetics* **22** 521–565.
- FELSENSTEIN, J. (1992). Estimating effective population size from samples of sequences: inefficiency of pairwise and segregating sites as compared to phylogenetic estimates. *Genetical Research* **59** 139–147.
- FELSENSTEIN, J. and CHURCHILL, G. A. (1996). A Hidden Markov Model approach to variation among sites in rate of evolution *Molecular Biology and Evolution* **13** 93–104.
- GRAHAM, J. and THOMPSON, E. A. (1998). Disequilibrium likelihoods for fine-scale mapping of a rare allele. *American Journal of Human Genetics* **63** 1517–1530.
- GRIFFITHS, R. C. (1989). Genealogical tree probabilities in the infinitely-many-site model. *Journal of Mathematical Biology* **27** 667–680.
- GRIFFITHS, R. C. and MARJORAM, P. (1996). Ancestral inference from samples of Dna sequences with recombination. *Journal of Computational Biology* **3** 479–502.
- GRIFFITHS, R. C. and MARJORAM, P. (1997). An ancestral recombination graph. pp. 257–270 in *Progress in Population Genetics and Human Evolution*, ed. P. Donnelly and S. Tavaré. Ima Volumes on Mathematics and Its Applications, volume 87. Springer, New York.
- GRIFFITHS, R. C. and TAVARE, S. (1994). Sampling theory for neutral alleles in a varying environment. *Philosophical Transactions of the Royal Society of London, Series B (Biological Sciences)* **344** 403–10.
- GRIFFITHS, R. C. and TAVARE, S. (1994). Ancestral inference in population genetics. *Statistical Science* **9** 307–319.

- GRIFFITHS, R. C. and TAVARE, S. (1994). Simulating probability distributions in the coalescent. *Theoretical Population Biology* **46** 131–159.
- HASTINGS, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57** 97–109.
- HUDSON, R. R. (1990). Gene genealogies and the coalescent process. *Oxford Surveys in Evolutionary Biology* **7** 1–44.
- HUDSON, R. R. and KAPLAN, N. L. (1988). The coalescent process in models with selection and recombination. *Genetics* **120** 831–840.
- KAPLAN, N. L., DARDEN, T. and HUDSON, R. R. (1988). The coalescent process in models with selection. *Genetics* **120** 819–829.
- KIMURA, M. (1968). Evolutionary rate at the molecular level. *Nature* **217** 624–626.
- KIMURA, M. (1983). *The Neutral Theory of Molecular Evolution*. Cambridge: Cambridge University Press.
- KINGMAN, J. F. C. (1982). The coalescent. *Stochastic Processes and Their Applications* **13** 235–248.
- KINGMAN, J. F. C. (1982). On the genealogy of large populations. *Journal of Applied Probability* **19A** 27–43.
- KINGMAN, J. F. C. (1982). Exchangeability and the evolution of large populations. pp. 97–112 in Koch, G. and F. Spizzichino, eds. *Exchangeability in Probability and Statistics. Proceedings of the International Conference on Exchangeability in Probability and Statistics*, Rome, 6th–9th April, 1981, in honour of Professor Bruno de Finetti. North-Holland Elsevier, Amsterdam.
- KRONE, S. M. and NEUHAUSER, C. (1997). Ancestral processes with selection. *Theoretical Population Biology* **51** 210–237.
- KUHNER, M. K., YAMATO, J. and FELSENSTEIN, J. (1995). Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. *Genetics* **140** 1421–1430.
- KUHNER, M. K., YAMATO, J. and FELSENSTEIN, J. (1997). Applications of Metropolis-Hastings genealogy sampling. pp. 183–192 in *Progress in Population Genetics and Human Evolution*, ed. P. Donnelly and S. Tavaré. IMA Volumes in Mathematics and its Applications, volume 87. Springer Verlag, Berlin.
- KUHNER, M. K., YAMATO, J. and FELSENSTEIN, J. (1998). Maximum likelihood estimation of population growth rates based on the coalescent. *Genetics*, in press.
- LI, S., PEARL, D. K. and DOSS, H. (1998). Phylogenetic tree construction using Markov chain Monte Carlo. Technical Report No. 583, Department of Statistics, Ohio State University, Columbus, Ohio.
- METROPOLIS, N., ROSENBLUTH, A. W., ROSENBLUTH, M. N., TELLER, A. H. and TELLER, E. (1953). Equation of state calculations by fast computing machines. *Journal of Chemical Physics* **21** 1087–1092.
- NATH, H. B. and GRIFFITHS, R. C. (1996). Estimation in an island model using simulation. *Theoretical Population Biology* **50** 227–253.
- NEUHAUSER, C. and KRONE, S. M. (1997). The genealogy of samples in models with selection. *Genetics* **145** 519–534.
- RANNALA, B. and SLATKIN, M. (1998). Likelihood analysis of disequilibrium mapping, and related problems. *American Journal of Human Genetics* **62** 459–473.
- RODRIGO, A. G. and FELSENSTEIN, J. (1998). Coalescent approaches to Hiv-1 population genetics. in press in *Molecular Evolution of Hiv*, ed. K. A. Crandall. Johns Hopkins University Press, Baltimore.
- TAKAHATA, N. (1990). A simple genealogical structure of strongly balanced allelic lines and trans-species evolution of polymorphism. *Proceedings of the National Academy of Sciences of the USA* **87** 2419–2423.
- WARD, R. H., FRAZIER, B.L., DEW-JAGER, K. and PAABO, S. (1991). Extensive mitochondrial diversity within a single Amerindian tribe. *Proceedings of the National Academy of Sciences*

of the USA **88** 8720–8724.

YANG, Z. (1993). Maximum-likelihood estimation of phylogeny from Dna sequences when substitution rates differ over sites. *Molecular Biology and Evolution* **10** 1396–1401.

YANG, Z. (1994). Maximum likelihood phylogenetic estimation from Dna sequences with variable rates over sites: approximate methods. *Journal of Molecular Evolution* **39** 306–314.

YANG, Z. (1995). A space-time process model for the evolution of Dna sequences. *Genetics* **139** 993–1005.

DEPARTMENT OF GENETICS
UNIVERSITY OF WASHINGTON
Box 357360
SEATTLE, WASHINGTON 98195-7360
JOE@GENETICS.WASHINGTON.EDU

USES OF STATISTICAL PARSIMONY IN HIV ANALYSES

BY KEITH A. CRANDALL

Brigham Young University

Molecular phylogenies have become powerful tools in human epidemiological studies. Because the phylogeny represents the historical relationship of genes through time, it plays an important role in the elucidation of both historical patterns and processes at work on the gene region of interest, and therefore, on the disease associated with that gene region. However, phylogenetically based analyses are only as good as the phylogenies upon which they are based. Two common problems result from the application of phylogenetic techniques to the population genetic level; 1) lack of resolution due to the short divergence times of a population study, and 2) incorrect inference due to the comparison of non-homologous sequence regions resulting from recombination. A population based method for reconstructing historical relationships among gene sequences is statistical parsimony. In this paper, I outline the limitations of traditional methods, outline the advantages and demonstrated superiority of statistical parsimony when divergences among sequences are low. Finally, I demonstrate the multiple applications of this estimation procedure to problems relating to human immunodeficiency virus evolution.

1. Introduction. Recent advances in population genetic theory, especially coalescence theory (Ewens 1990; Hudson 1990; Donnelly and Tavaré 1995), coupled with an expansion of molecular techniques, have allowed detailed phylogenetic information at the population level. Such genealogical relationships are termed gene trees, allele trees, or haplotype trees, in which different haplotypes or alleles are merely unique nucleotide sequences for a specific region of DNA (loosely termed a gene). With these advances, phylogenetic approaches have proven powerful in studying problems in population genetics and human epidemiology. For example, researchers have utilized phylogenies to explore the origin and spread of retroviruses such as HIV-1, HIV-2 and SIV through a population (Hirsch et al. 1989; Gojobori et al. 1990) and to identify transmission events among individuals and between species (Ou et al. 1992; Holmes et al. 1993; Crandall 1995; Sharp et al. 1996). Phylogenetic studies have also been used to examine the population dynamics of viral infections and the associations of host/pathogen (Harvey and Nee 1994; Holmes and Garnett 1994). Phylogenies have played a central role in longitudinal studies examining the diversification of HIV through time (Kuiken et al. 1993; Strunnikova et al. 1995) and how this

Research partially supported by NIH Grant R01-HD34350-01A1 and the Alfred P. Sloan Foundation.

AMS 1991 subject classifications. Primary 92D15, 62G10; secondary 92D10, 92D20.

Key words and phrases. Phylogeny, HIV, parsimony, population genetics.

nucleotide diversity is associated with compartmentalization within the body (Epstein et al. 1991; Ait-Khaled et al. 1995; Kuiken et al. 1995; Strunnikova et al. 1995) and related to gene function (McNearney et al. 1995). Finally, phylogenetic analyses have played a central role in the classification of HIV sequences. Primate lentiviruses have been classified into five distinct lineages, with the HIV-1's representing one of these major lineages. The diversity within this group of HIV-1 sequences has been further subdivided into distinct subtypes (A-G) which represent phylogenetically distinct lineages (Sharp et al. 1994). Clearly then, phylogenetic analyses play a central role in the study of HIV infection, transmission, and diversification.

While there are many examples of the utility of a phylogenetic approach to studies in human epidemiology, all these studies rely on an accurate estimate of phylogenetic relationships. However, traditional methods for estimating phylogenetic relationships (e.g., maximum parsimony, maximum likelihood, and neighbor-joining) have severe limitations at the population genetic level. In this paper, I will; 1) outline the difficulties associated with these traditional methods, 2) present a phylogeny reconstruction method developed specifically to account for these difficulties, and 3) demonstrate the utility of this method to studies of HIV diversity.

2. Statistical parsimony: a new method for HIV phylogenetic analyses. Molecular phylogenies have become powerful tools in human epidemiological studies. Because the phylogeny represents (hopefully) the historical relationship of genes through time, it plays an important role in the elucidation of both historical patterns and processes at work on the gene region of interest, and therefore, on the disease associated with that gene region. When genes are surveyed for variation at the population genetic level, phylogenetic inference can also elucidate current processes affecting the population dynamics of these sequences. The examples cited in the introduction support the utility of phylogenetic analyses in human epidemiology. However, phylogeny based analyses are only as good as the phylogenies upon which they are based. Two common problems result from the application of phylogenetic techniques to the population genetic level; 1) lack of resolution due to the short divergence times of a population study, and 2) incorrect inference due to the comparison of nonhomologous sequence regions resulting from recombination. A population based method for reconstructing historical relationships among gene sequences is statistical parsimony. This population-based approach takes into account these phenomena that violate assumptions of traditional methods. In this section, I outline the limitations of traditional methods, outline the advantages and demonstrated superiority of statistical parsimony when divergences are low. In the following

sections, I then outline the method and demonstrate its application to a number of problems in HIV research.

2.1. Limitations of traditional phylogenetic approaches. Inferences from phylogenetic analyses are only as good as the phylogenies upon which they are based. Phylogenetic analyses of HIV sequences rely on methods developed for systematic biology and therefore are subject to the biases associated with such studies. Traditional methods of phylogeny reconstruction were developed to estimate relationships of higher taxonomic groups, e.g., species, genera, families, etc. Consequently, these methods make at least two major assumptions that are invalid at the population genetic level (Crandall et al. 1994). First, species trees are traditionally regarded as strictly bifurcating. However, in populations, most haplotypes in the gene pool exist as sets of multiple, identical copies because of past DNA replication. In haplotype trees, each gene lineage of the identical copies of a single haplotype is at risk for independent mutation. Consequently, coalescence theory predicts that a single ancestral haplotype will often give rise to multiple descendant haplotypes, thereby yielding a haplotype tree with multifurcations. Second, gene regions examined in populations can undergo recombination. Traditional methods assume recombination does not occur in the region under examination. Furthermore, recombination is an additional reason why the assumption of a strictly bifurcating tree topology is likely to be violated.

Additional differences exist between gene trees at the population level and higher taxa divergences (Pamilo and Nei 1988). Populations typically have lower levels of variation over a given gene region relative to higher taxonomic levels, resulting in fewer characters for phylogenetic analyses. Huelsenbeck and Hillis (1993) have shown that interspecific methods for phylogeny reconstruction perform poorly when few characters are available for analysis. Another difference concerns the treatment of ancestral types. In populations, when one copy of a haplotype in the gene pool mutates to form a new haplotype, it would be extremely unlikely for all the identical copies of the ancestral haplotype to also mutate. Thus as mutations occur to create new haplotypes, they rarely result in the extinction of the ancestral haplotype. The ancestral haplotypes are thereby expected to persist in the population. Indeed, coalescence theory predicts that the most common haplotypes in a gene pool will tend to be the oldest (Watterson and Guess 1977; Donnelly and Tavaré 1986), and most of these old haplotypes will be interior nodes of the haplotype tree (Crandall and Templeton 1993; Castelloe and Templeton 1994). Thus, a method for reconstructing genealogical relationships is needed that takes into account these population genetic phenomena, e.g., statistical parsimony. There are three distinct advantages to this method over traditional methods of phylogeny reconstruction: 1) it has its sta-

tistical power when sequence divergences are low making it complementary to traditional techniques, 2) it offers a quantitative assessment of alternative tree topologies within the 95% confidence set, and 3) it can be applied to sequences that have resulted from recombination and account for recombination in the reconstruction of genealogical relationships. The next section describes how these advantages are realized.

2.2. Estimating intraspecific gene genealogies - Statistical Parsimony. Templeton, Crandall and Sing (1992) have developed a method to estimate within-species gene genealogies based on the probability of multiple mutations at a specific site exhibiting a difference between a pair of haplotypes. This statistical parsimony method is compatible with either nucleotide sequence or restriction site data. The method sets a statistical criterion for the limits of the parsimony assumption; that is, the probability that a nucleotide difference between a specific pair of sequences is due to a single substitution (the parsimonious state) and not the result of multiple substitutions at a single site (the nonparsimonious state) (Templeton et al. 1992). Thus, I use the term parsimony to refer to the minimum number of differences separating two individual sequences rather than a global minimum tree length based on shared derived characters. Hudson (1989) described the probability that a restriction site difference between two randomly drawn individuals from a population is due to a single mutation to be

$$(2.1) \quad H = 1 - \frac{2 \left(\prod_{i=1}^{n-1} \frac{i}{i + r\theta} \right) \left(\sum_{i=1}^{n-1} \frac{r\theta}{i + r\theta} \right)}{\left(\sum_{i=0}^{n-1} \frac{r\theta}{i + r\theta} - \prod_{i=1}^{n-1} \frac{i}{i + r\theta} \right)},$$

where r is the length of the recognition sequence of the restriction enzyme in nucleotides, n is the number of individuals in the entire sample, and $\theta = 4N_e\mu$ where N_e is the inbreeding effective population size and μ is the per nucleotide mutation rate. Thus H provides a probability of multiple hits at a given restriction site (or nucleotide when $r = 1$) or, alternatively, a probability that our data follow the infinite sites model. Intuitively, we would like this probability to change depending upon the number of restriction sites (or nucleotides) sampled, i.e., for a pair of sequences differing by one site but sharing only ten sites the probability of that event being due to a single mutation would be lower than if a pair differed by a single site but shared 100 sites. Using the restriction site model of evolution described by Templeton et al. (1992), we can calculate the total probability that two haplotypes differ at j sites and share the presence of

m sites to be $L(j, m)$, i.e.

$$(2.2) \quad (2q_1)^{j-1}(1-q_1)^{2m+1} \left(1 - \frac{q_1}{br}\right) \left(2 - \frac{q_1[br+1]}{br}\right)^{j-1} \left(1 - 2q_1 \left[1 - \frac{q_1}{br}\right]\right)$$

where q_1 is the probability of a nucleotide change within a block of r nucleotides in the two haplotypes since their respective lineages diverged and b is a parameter to incorporate mutational bias (i.e., $b = 3$ if there are three alternative states for a nucleotide to change to and 2 or 1 if there is some bias in the substitutional pattern for a given region of DNA such that nucleotide substitutions are restricted to 2 or 1 alternative states). Again, L , like H , is giving us the probability of multiple mutational events at those sites differing between a pair of sequences, i.e., we are assigning a probability of the data fitting an infinite sites model. Only, now, we are conditioning that probability on the number of shared sites between two haplotypes. A number of statistical techniques are available to evaluate this expression. Maximum likelihood appears to have boundary value problems at one of the most important evaluations, when $j = 1$; i.e. when two haplotypes differ by a single site. Here the maximum likelihood estimator of q_1 occurs on the boundary condition of 0; i.e., equation (2) reaches its maximum value of 1 when $q_1 = 0$ regardless of the value of m . Therefore the maximum likelihood estimator would always justify the use of parsimony for haplotypes separated by a single difference, despite the results from Hudson (1989) suggesting this might not be the case and empirical results suggesting the same (Crandall et al. 1994). Furthermore, this approach does not support our intuition about the probability adjusting to the number of sites sampled, since the maximum likelihood estimator would always justify sequences differing by a single nucleotide regardless of the number of shared sites.

Instead, we can consider equation (2) as a posterior probability distribution of the data given q_1 and estimate q_1 through a Bayesian analysis. Assuming a uniform prior on q_1 , the Pitman estimator of q_1 is

$$(2.3) \quad \hat{q}_1 = \frac{\int_0^1 q_1 L(j, m) dq_1}{\int_0^1 L(j, m) dq_1}.$$

When $j > 1$ (i.e., there are more than a single difference between haplotypes), deviations from parsimony can occur at other sites in addition to q_1 . We can then estimate q_2 by replacing j with $j - 1$ in equation (2). Likewise, we can perform this calculation iteratively to obtain a set of estimators $\{q_1, \dots, q_j\}$. Then P_j , the probability that two haplotypes differing by j sites but sharing m

sites have a parsimonious relationship, can be estimated by

$$(2.4) \quad \hat{P}_j = \prod_{i=1}^j (1 - \hat{q}_i).$$

With this estimator, the probability that a nucleotide difference between two haplotypes is due to one and only one substitution increases as the number of shared sites between sequences increases. This procedure estimates a set of probable relationships between haplotypes whose cumulative probability is ≥ 0.95 .

After the connection of haplotypes at a given mutational distance, the network of haplotypes established is inspected for evidence of recombination. Recombination can create homoplasies; these are changes introduced in two diverged sequences that are the same, even though the evolution along these lineages was independent (Swofford et al. 1996). Recombination is inferred under two sets of conditions; 1) if two or more homoplasies can be resolved by the inference of a recombination event, or 2) if a single homoplasy involves a mutation of the type that is assumed to evolve in a completely parsimonious fashion, e.g. insertion/deletion events (for examples, see Templeton et al. 1992). The impact of recombination upon the remainder of the analysis depends upon the size of the inferred recombinational region. If only a small number of observations are associated with the recombinational haplotype(s), we simply exclude the recombinants from the analysis. If a large proportion of the data is excluded by this step or recombination appears to be extensive in the region as a whole, we subdivide the region using the ‘approximate’ algorithm given in Hein (1990; 1993). An independent analysis is then performed on each subregion in turn.

The model underlying statistical parsimony assumes independence of sites and allows for biases in substitutional patterns of nucleotide changes (Templeton et al. 1992). The allowance for different mutational biases between each pair of haplotypes being examined results in a lack of a requirement that the mutations be identically distributed; a typical assumption for many reconstruction algorithms. Likewise, the testing for multiple substitutions between a pair of haplotypes assures (at a 95% confidence level) that the infinite alleles model (no site will experience multiple mutations) is not violated by the established relationships. This is important when utilizing results from coalescence theory to refine cladogram probabilities and assign outgroup probabilities (Crandall and Templeton 1993; Castelloe and Templeton 1994; Crandall et al. 1994).

The power of the statistical parsimony procedure is achieved by incorporating the number of shared sites in calculating the probability of multiple mutations at nucleotide positions that differ between a given pair of haplotypes.

Therefore the fewer differences (more shared sites) between a pair of haplotypes, the greater the probability that those few nucleotide substitutions are due only to a single mutational event. This estimation procedure has demonstrated statistical power when reconstructing gene trees and greatly outperforms maximum parsimony when the number of nucleotide substitutions is small and the number of shared positions is large (Crandall 1994), as is the case with many population level sequence data sets. Thus statistical parsimony takes into account the population level phenomena (low levels of divergence, recombination, multifurcations) that violate the assumptions of many traditional estimation procedures, thereby providing a better estimate of genealogical relationships.

It is important to note, when comparing this method to standard tree reconstruction methods, that this method is assessing confidence in connections in a pairwise fashion. This is similar to the bootstrap in that the bootstrap does not give an indication of the confidence in the overall tree topology, but rather in various nodes estimated by the phylogeny reconstruction method (Felsenstein 1985). Standard methods typically do provide some measure of overall fit of a tree (e.g., the likelihood score for a maximum likelihood tree or the number of steps in a parsimony tree). At this point, we have not developed a score to assess overall tree topology. In fact, one of the advantages of this method is that it ignores homoplasy associated with fitting characters to the entire tree and concentrates instead on minimum pairwise connections.

2.3. The general problem of recombination. A fundamental assumption to traditional methods of phylogeny reconstruction is that the set of aligned sequences from which the phylogeny is estimated are homologous, i.e. are similar due to shared ancestry (Hillis 1994). Recombination, which is rarely tested for, results in a direct violation of this assumption. Therefore, recombination in a gene region can cause incorrect phylogenetic inference (Sanderson and Doyle 1992), compromising the power of the phylogenetic approach in epidemiological and population genetic studies. This is true not only for the statistical parsimony method presented here, but for all phylogenetic approaches. Studies in human epidemiology are typically associated with DNA segments that have the possibility of having undergone recombination (McClure 1991). In those few studies that have explored the possibility of recombination, it has been found to have a significant impact on our understanding of the history of gene genealogies and arguments based on these phylogenies (Robertson et al. 1995a, b). Recombination is also an important force in generating genetic diversity within a population. Effects of recombination in HIV sequences have recently been an important consideration in vaccine development (Cammack et al. 1988) and the evolution of drug resistance (Kellam and Larder 1995). Recombination

is also of great utility in quantitative trait loci studies as it narrows the search region of an associated phenotypic effect (Templeton et al. 1992; Templeton and Sing 1993; Crandall 1996a). Thus the ability to accurately detect recombination in a set of aligned sequences is of utmost importance in phylogenetic studies.

A number of statistical techniques have been developed to test for the occurrence of recombination within a given gene region and to determine the bounds of the recombinational event (reviewed in Crandall and Templeton 1999). Few phylogeny reconstruction techniques, one of which is the statistical parsimony procedure (Templeton et al. 1992), have been developed that take into account the possibility of recombination. These techniques differ in their criteria for determining whether or not recombination has occurred and very little is known about the relative performance of these techniques. However, they share the same algorithm for reconstructing histories given the detection of recombination.

2.4. Detecting recombination using statistical parsimony. Recombination is inferred if homoplasies are indicated involving either mutational classes regarded as completely parsimonious (e.g., indels) or if the inference of recombination can resolve two or more homoplasies involving restriction sites or nucleotides. These criteria were suggested and first used by Aquadro et al. (1986) in their study of the alcohol dehydrogenase gene region in *Drosophila melanogaster*. Homoplasies are identified by comparing the network with a phylogeny resulting from a standard maximum parsimony analysis using a program such as PAUP (Swofford 1993). If recombination is indicated, the impact on the analysis depends on the size of the recombinational unit. If the inferred recombination event encompasses a single haplotype or small region, the haplotype is excluded in subsequent steps. If the inferred region is large and encompasses many haplotypes, the region is subdivided into smaller regions with no evidence of recombination. Then separate networks are united by examining the cumulative probability of parsimony for haplotypes that differ by j or $j+1$ mutational steps (equation [9]; Templeton, et al. 1992). If justified, networks are then joined by these minimal connections.

The recombinational criteria have been augmented recently to include an additional test to identify recombination among closely related nucleotide sequences (Crandall and Templeton 1999). The major improvement is a second test for recombination once a candidate for recombination is identified using the multiple homoplasy criterion. The second test looks for statistically significant runs of substitutions within the set of homoplasious characters. The underlying assumption is that if recombination has occurred, the substitutional pattern will reflect the recombination event by an arrangement such that all the substitu-

tions from one parent will come either before or after all the substitutions from the other parent. In general, suppose we have α mutations on one branch and β on the other branch leading to the potential recombinant. We then order them into the α smallest and β largest by nucleotide site number. A perfect match for recombination would have all α smallest on one branch, and all α largest on the other. The probability of getting κ successes in this case (i.e. the number of low site mutations on the low site branch) is given by the hypergeometric distribution:

$$(2.5) \quad \text{Prob } (\kappa \text{ successes}) = \frac{\binom{\alpha}{\kappa} \binom{\beta}{\alpha - \kappa}}{\binom{\alpha + \beta}{\alpha}}$$

When $\kappa < \alpha$, you need to take the sum of the above probabilities from κ equals the observed κ to α to get the appropriate tail probability. We can therefore ask if nucleotide substitutions are ordered as expected in the case of recombination using this equation. This is done for those haplotypes involved in homoplasious connections.

2.5. Accuracy of statistical parsimony. While there are many applications of the statistical parsimony method (see below), the results of these applications are only as good as the method itself. Hillis (1995) has reviewed the four main approaches to exploring the accuracy of phylogeny reconstruction methods; evolutionary simulations (Huelsenbeck 1995), known (observed) phylogenies (Hillis et al. 1992), statistical evaluation (Li and Zharkikh 1995), and congruence studies (Miyamoto and Fitch 1995). Congruence tests are inapplicable for our purposes since the statistical parsimony method constructs gene trees, not species trees. Our method does provide a statistical evaluation of each connection established; however, this evaluation depends upon the model employed by the method. Thus evolutionary simulations and known phylogenies are the two remaining methods to explore the accuracy of the statistical parsimony method. Because this method is designed specifically for phylogenies with low levels of divergence (typically within species phylogenies), assessments cannot be made using “well-supported” phylogenies (Allard and Miyamoto 1992). An additional problem with the “well-supported” phylogeny approach is that it confounds repeatability and accuracy, where repeatability is the probability that a given result will be found again using an alternative reconstruction method or data set and accuracy is the probability that a given result represents the true phylogeny (Hillis and Bull 1993). Recently, the laboratory of Hillis and Bull (White

et al. 1991; Hillis et al. 1992; Bull et al. 1993) introduced an experimental system in which they generated known phylogenies of the bacteriophage T7. These phylogenies can be used to test alternative phylogeny reconstruction techniques under a variety of evolutionary circumstances. With these data, repeatability and accuracy can be partitioned in an analysis of reconstruction techniques.

I have tested the accuracy of the statistical parsimony method using the restriction site data generated by Hillis et al. (1992). To simulate the conditions under which the statistical parsimony method is advertised to perform, i.e. low levels of divergence, we subsampled 16 restriction sites of the 199 variable restriction sites surveyed. I then established relationships based on these sites using both the statistical parsimony method and maximum parsimony. The number of connections between haplotypes was then tallied as well as whether or not the established connection was correct. The results showed that the statistical parsimony performs very near the stated confidence level, i.e. the 95% confidence set of connections established were indeed correct 94% of the time (Crandall 1994). Furthermore, the results showed the statistical parsimony greatly outperforms maximum parsimony in the number of correct connections (Crandall 1994). For example, the percentage of connections inferred correctly for those haplotypes that differed by one or two nucleotides were 91% and 99%, respectively, with statistical parsimony. On the contrary, with maximum parsimony only 23% and 38% of the connections were inferred correctly. This direct comparison using a known phylogeny demonstrated the superiority of the statistical parsimony method in the accurate reconstruction of evolution history.

Obviously, it would be of interest to test our method with additional data from known phylogenies. We are currently pursuing this area of research through the use of computer simulated data sets. Using computer simulation, we can generate known genealogies under a variety of conditions and test the robustness of this method to violations of its assumptions. Additionally, we can incorporate recombination into the genealogy in different topological locations and at different frequencies to test the methods ability to detect recombination and accurately reconstruct the true genealogy. This work is ongoing in my lab at the moment.

2.6. Nested statistical analyses. Templeton et al. (1987) and Templeton and Sing (1993) have developed statistical procedures for detecting significant associations between phenotype and genotype within a cladogram framework. Their procedures utilize the cladogram structure from the above estimation procedure to define a nested statistical design, thereby allowing the clustering of individuals based on genotype rather than phenotype. The statistical analysis allows ambiguity in the cladogram estimation and is compatible with either quantita-

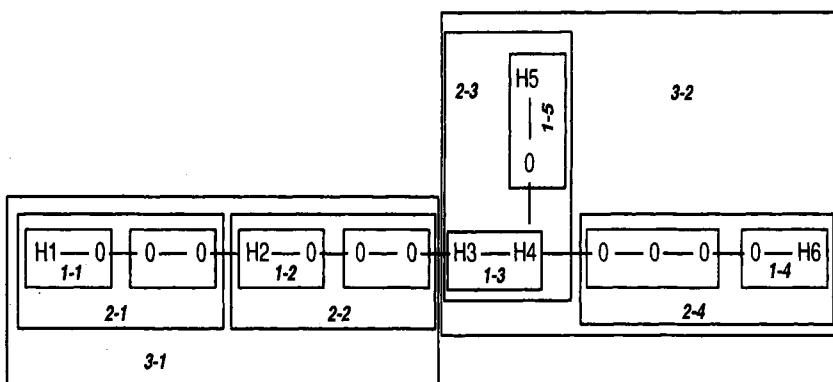


FIG. 1. *A demonstration of the nesting procedure for nucleotide sequence data. H1 through H6 represent the haplotypes under consideration. Lines indicate the mutational pathway interconnecting the six haplotypes with zeros representing missing intermediates. Boxes indicate nesting clades labeled with two numbers in bold and italics. The first number indicates the nesting level and the second is a counter of the clades at that level. Thus, for example, clade **2 – 1** is the first clade formed at the second nesting level. Increasing nesting level corresponds to relatively increasing evolutionary time.*

tive or categorical phenotypes.

The nesting procedure consists of nesting n -step clades within $(n + 1)$ -step clades, where n refers to the number of transitional steps used to define the clade. Thus, n is correlated with, but does not refer to, the number of nucleotide differences separating individual haplotypes. By definition, each haplotype is a 0-step clade. The $(n + 1)$ -step clades are formed by the union of all n -step clades that can be joined together by $(n + 1)$ mutational steps. The nesting procedure begins with tip clades, i.e., those clades with a single mutational connection (e.g., haplotypes H1, H5, and H6 in Figure 1), and proceeds to interior clades. In previous analyses based on restriction site data, missing intermediates were ignored in the nesting procedure as they were inconsequential to these analyses. However, with nucleotide sequence data, there are many more missing intermediates because haplotypes are typically differentiated by more than a single nucleotide difference. These missing intermediates must be considered in the nesting procedure to assure overall consistency. Because these missing intermediates become nested together, the nesting procedure results in a number of empty clades, i.e., two missing intermediates are nested together resulting in a next level clade that represents a missing intermediate as well (see zeros nested together in Figure 1). These next level empty clades are required for the consistency of the nesting procedure to form higher level clades, but can be ignored during subsequent statistical analyses since they contain no observations.

Figure 1 offers an example of the nesting procedure performed with all the

missing intermediates designated by zeros. The 1-step nesting level produces eight clades of which five contain sampled haplotypes. Thus these five clades are labeled **1-1** through **1-5**, where the first number refers to nesting level and the second is a counter for clades containing sampled haplotypes that have been nested at that level. Notice one clade contains three missing intermediates. After nesting in from the H5 and H6 tips, the first missing intermediate to the right of H4 can either nest with the H3-H4 clade or with the clade of missing intermediates to its right. This situation has been termed symmetrically stranded (Templeton and Sing 1993). The placement of the stranded haplotype is initially based on sample size, i.e., it is placed in the clade with the smallest sample size. This results in greater samples within and among clades for hypothesis testing. Therefore, in this example, the missing intermediate is nesting with the other missing intermediates. If both alternatives have the same sample sizes, then one alternative is chosen at random (Templeton and Sing 1993). Now, the 2-step nesting begins with the underlying 1-step clades as the "haplotypes", resulting in four 2-step clades. Nesting continues until the step before all haplotypes are nested into a single clade. Additional rules for nesting with ambiguity are given in Templeton and Sing (1993). The nesting procedure results in hierarchical clades with nesting level directly correlated to evolutionary time, i.e., the lower the nesting level the more recent the evolutionary events relative to higher nesting levels.

3. Applications of the statistical parsimony method to studies of HIV. Statistical parsimony is being used more and more in studies of HIV sequence analysis because of the favorable properties described above for studying sequence variation from closely related sequences. Because the HIV virus has a high mutation rate, sequences from different individuals are often too far diverged for analysis with statistical parsimony. However, this technique is well suited for analyzing sequence variation in HIV isolates from a single patient. Below are a number of examples of such analyses with HIV sequences.

3.1. HIV transmission. The statistical parsimony procedure and associated nested analysis has been used successfully in a number of HIV related studies. One such example is that of the Florida Dentist transmission case. DeBry et al. (1993) reexamined the conclusion reached by Ou et al. (1992) that a Florida dentist infected five of his eight HIV-1 seropositive patients using an alternative model of evolution and additional controls. They criticized the original analysis for using an inappropriate model of evolution (and phylogeny reconstruction technique) and inadequate sampling of local controls. The original analysis used a parsimony optimality criterion with equal weighting of character changes (Ou

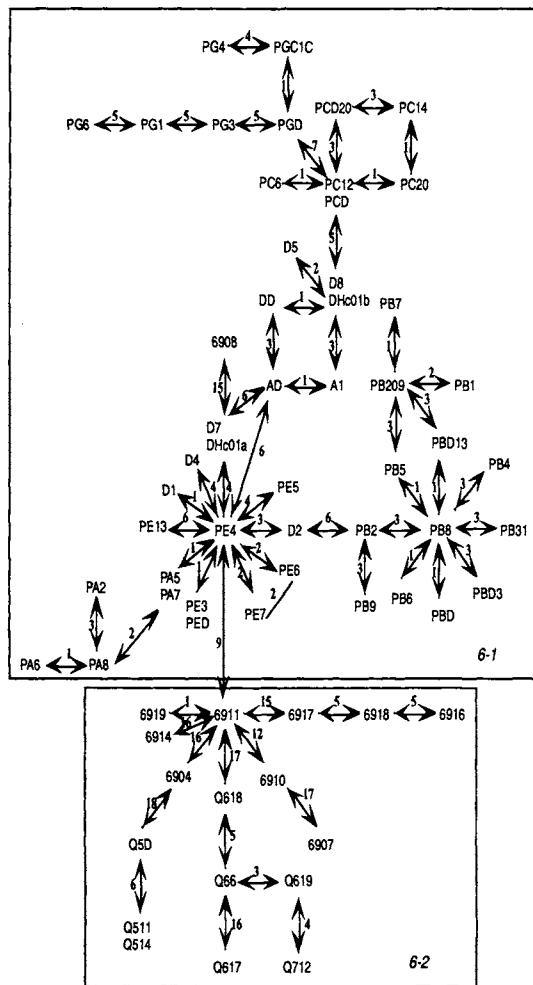


FIG. 2. The main network of HIV sequences from the Florida dentist (labeled Dx) and his patients (labeled Pxx) and local controls. Sequences from patients F, D, and H did not connect to this main network, however sequences from patients A, B, C, and E are clearly epidemiologically linked to the sequences from the dentist. The nested analysis shows that the local controls cluster apart from the dentist and patient sequences through the 6th step nesting level.

et al. 1992). DeBry et al. (1993) argued that because there appears to be rate heterogeneity in substitutions across nucleotide positions, a method should be used that takes rate heterogeneity into account. They used threshold parsimony to accomplish this. The phylogenetic analysis by DeBry et al. (1993) indicated no resolution between the null hypothesis of independent acquisition of the HIV-1 virus versus the alternative of infection via the Florida dentist. The nucleotide sequences in this data set violate the assumptions of the phylogeny reconstruction methods used by both these groups. I reanalyzed the HIV sequences of Ou et al. (1992) with the addition of new sequences from DeBry et al. (1993) using the statistical parsimony procedure, whose assumptions are not violated by the data (Crandall 1995). The resulting cladogram indicated statistical support for the 'dental clade' as originally concluded by Ou et al. (1992) (Fig. 2). Furthermore, a nested statistical analysis gives further support for the 'dental clade' (Crandall 1995). This was the only analysis of these data that included the entire data set. Other analyses have only used partial sequence data because the traditional techniques being used give no resolution for the closely related multiple samples within individuals. The lack of resolution by DeBry et al. (1993) was a result of using a weak analytical procedure, not a shortcoming of the data itself. Thus the statistical parsimony method provided a superior analytical framework relative to previous analyses for a number of reasons; 1) linkages could be made with greater statistical support due to the superior statistical power of the statistical parsimony procedure relative to either maximum parsimony used by Ou et al. (1992) or threshold parsimony used by DeBry et al. (1993), 2) population level phenomena such as multifurcations, interconnections among sequences, and ancestral sequences remaining in the population were more accurately represented in the resulting network than in the bifurcating trees from previous analyses, 3) recombination was tested for and not discovered, as opposed to previous analyses which assumed it did not exist, 4) for the first time, all the sequences relevant to the hypotheses of transmission were analyzed (i.e. the statistical parsimony method can accommodate all sequences, even those that were closely related and therefore ignored by the previous analyses, and 5) the method provided a powerful framework within which I tested hypotheses of transmission to the dental patients in contrast to previous analyses which did not set up an appropriate hypothesis testing framework (Hillis and Huelsenbeck 1994).

Another example where I have used the statistical parsimony method to investigate the occurrence of viral transmissions is that of primate T-cell lymphotropic virus type 1 (PTLV-1) (Crandall 1996b). Unlike HIV-1 sequences, PTLVs have relatively low levels of divergences even among host species of primates. Because of the low level of divergence, traditional phylogeny reconstruc-

tion techniques have not been able to resolve relationships with statistical confidence. Various research groups have presented point estimates of phylogenetic relationships that are suggestive of cross-species transmission of PTLVs among various primate species. However, when bootstrapping procedures (Felsenstein 1985; Hillis and Bull 1993; Zharkikh and Li 1995) are applied to test the reliability of this result, no conclusions can be drawn as nodes uniting PTLVs from different host species are not well supported (Saksena et al. 1993; Koralnik et al. 1994). Using the statistical parsimony method, greater resolution was achieved in establishing phylogenetic relationships among viral sequences. Additionally, because of the hypothesis testing framework associated with this method, hypotheses of cross-species transmission could be appropriately tested. We showed that a range of 11 to 16 cross-species transmissions have occurred throughout the history of these sequences. Additionally, outgroup weights were assigned to haplotypes using arguments from coalescence theory to infer directionality of transmission events. Finally, we compared the results from the statistical parsimony method directly to results obtained from a traditional maximum parsimony approach and found statistical parsimony to be superior at establishing relationships and identifying instances of transmission. We first estimated relationships among 72 sequences of 520 base pairs in length from the *env* gene. The maximum parsimony analysis resulted in over 8,000 most parsimonious trees. The computer memory limited the search to 8,000 trees. Thus, the effectiveness of the maximum parsimony search was restricted due to the ambiguity in the data set (Maddison 1991a,b; Templeton 1992). Furthermore, a bootstrap analysis was impossible, given the difficulty of the initial parsimony search. Therefore, transmission events could not be statistically inferred using the traditional parsimony approach. Yet, using the statistical parsimony approach, 5 time independent networks were estimated with linkages within networks supported at the 95% confidence level or greater. Using these statistically supported relationships, multiple cross-species transmission events were inferred; thereby demonstrating the superiority of the statistical parsimony method in both phylogeny estimation and hypothesis testing. In addition to the inferences concerning cross-species transmission and molecular evolution of the PTLV sequences, I also presented extensions of the nesting procedure (Templeton et al. 1987; Templeton and Sing 1993) and outgroup weighting (Castelloe and Templeton 1994) for haplotypes based on sequence data.

3.2. HIV subtyping. Understanding the global diversity of HIV-1 allows for accurate and predictive modeling of the spread of infection, accurate estimates of the historical spread of HIV, and effective development of vaccines. The diverse forms of HIV-1 have been classified into phylogenetically distinct subtypes

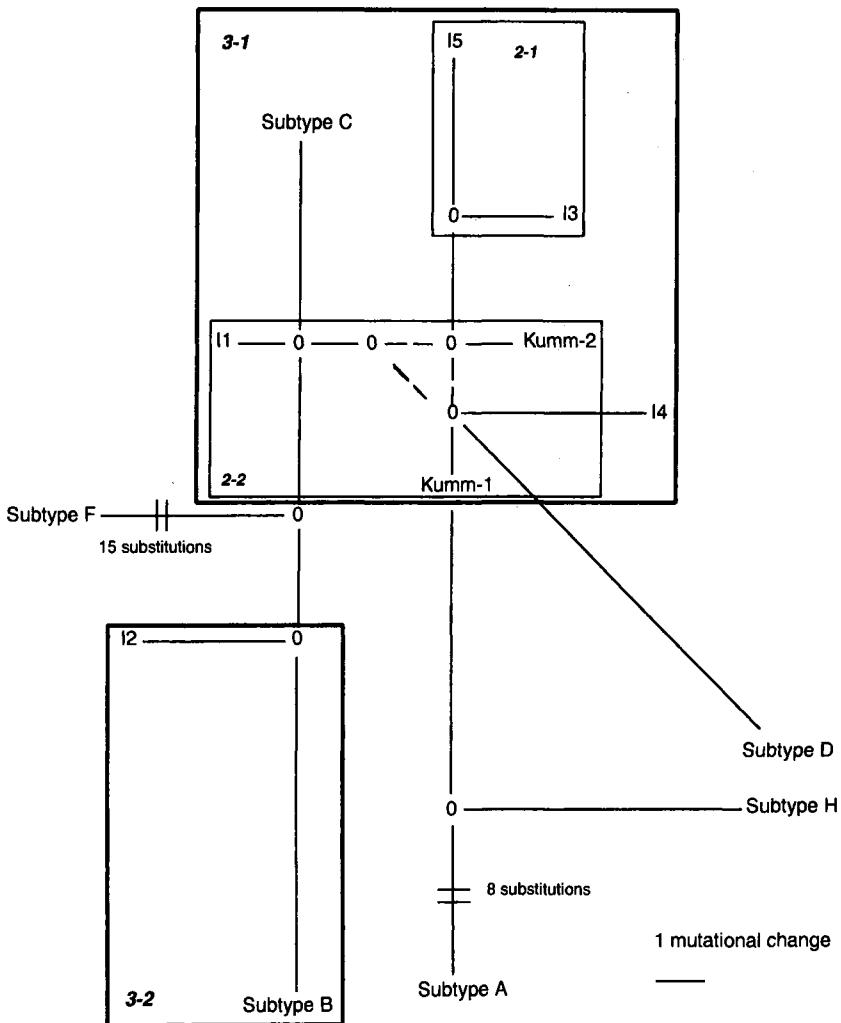


FIG. 3. Network of statistically supported relationships among new sequences (II-15, Kumm-1, Kumm-2, and known subtypes (A-H, except E and G which were so divergent their connection to this main network was not statistically supported). Branch lengths are proportional to the number of nucleotide differences between sequences with the exception of Subtypes F & A whose distances are given on the branches. Dashed lines indicate ambiguity in the given relationships.

(Myers et al. 1993). While the defining of subtypes is preferably done using the sequences of an entire genome of HIV-1, examination of distributional patterns of specific subtypes can be performed using a subset of sequence data. The statistical parsimony procedure can be used effectively to determine HIV-1 sequence subtype when the number of nucleotides that differ among sequences is small. Voevodin et al. (1996) indicated that sequences from new regions of India and Indian and Ethiopian expatriates in Kuwait could clearly be assigned to either subtypes B or C based on a minimum of 231 base pairs from the *gag* gene (Fig. 3). Here the branch lengths are drawn proportional to the amount of change along the branch.

3.3. Longitudinal studies. One of the great utilities of the statistical parsimony method is in longitudinal studies. In such studies, one is interested in the partitioning of variation in HIV sequences within a patient over time. Typically, such studies have sequences that are < 5% diverged, making them ideal candidates for the statistical parsimony method. One such study examined variation in the *nef* region of HIV-1 (McNearney et al. 1995). They found that deletions and rearrangements were more common in late than early stages of disease progression. Additionally, they found continued sequence evolution in HIV-1 quasispecies with *nef* deletions suggesting that *nef*-deleted quasispecies are capable of replication *in vivo* (McNearney et al. 1995).

4. Software availability. We are currently working on a program that will implement the statistical parsimony procedure in its entirety. This program is still at least three months away from distribution. However, we do have a Mathematica package that will calculate the probabilities given in equations (1-4). This package is available from the author upon request.

5. Summary. Phylogenetic approaches have proven powerful in comparative biology at the population level and higher taxonomic levels. However, traditional methods for estimating phylogenetic relationships (e.g., maximum parsimony, maximum likelihood, and neighbor-joining) assume recombination has not occurred in a set of aligned sequences. Additionally, traditional methods assume the history of the sequences can be adequately represented by a bifurcating (or multifurcating) tree topology. Recombination directly violates this assumption resulting in reticulate relationships which can be represented by networks. Because of these assumption violations, recombination in a gene region can cause incorrect phylogenetic inference, compromising the power of the phylogenetic approach in evolutionary studies. Thus, the ability to accurately detect recombination is of utmost importance in phylogenetic studies.

The examples given above demonstrate two important attributes of the statistical parsimony approach. First, in multiple studies where statistical parsimony has been compared directly to traditional techniques, it has always been found to be superior at estimating phylogenetic relationships. Because of this, it has also been superior at testing hypotheses based on these phylogenies. The additional power in the statistical parsimony procedure comes from taking into account population biological phenomena such as recombination. Second, statistical parsimony has a wide range of application in HIV studies. This, combined with its superior performance, make it a highly desirable method of analysis.

Acknowledgements. I thank Joe Felsenstein and an anonymous reviewer for providing helpful comments to improve the manuscript. This work was supported by the Alfred P. Sloan Foundation and the National Institutes of Health.

REFERENCES

- AIT-KHALED, M., McLAUGHLIN, J. E., JOHNSON, M. A. and EMERY, V. C. (1997). Distinct HIV-1 long terminal repeat quasispecies present in nervous tissues compared to that in lung, blood and lymphoid tissues of an AIDS patient. *AIDS* **9** 675–683.
- ALLARD, M. W. and MIYAMOTO, M. M. (1992). Perspective: Testing phylogenetic approaches with empirical data, as illustrated with the parsimony method. *Molecular Biology and Evolution* **9** 778–786.
- AQUADRO, C. F., DESSE, S. F., BLAND, M. M., LANGLEY, C. H. and LAURIE-AHLBERG, C. C. (1986). Molecular population genetics of the alcohol dehydrogenase gene region of *Drosophila melanogaster*. *Genetics* **114** 1165–1190.
- BULL, J. J., CUNNINGHAM, C. W., MOLINEUX, I. J., BADGETT, M. R. and HILLIS, D. M. (1993). Experimental molecular evolution of bacteriophage T7. *Evolution* **47** 993–1007.
- CAMMACK, N., PHILLIPS, A., DUNN, G., PATEL, V. and MINOR, P. D. (1988). Intertypic genomic rearrangements of poliovirus strains in vaccinees. *Virology* **167** 507–514.
- CASTELLOE, J. and TEMPLETON, A. R. (1994). Root probabilities for intraspecific gene trees under neutral coalescent theory. *Molecular Phylogenetics and Evolution* **3** 102–113.
- CRANDALL, K. A. (1994). Intraspecific cladogram estimation: Accuracy at higher levels of divergence. *Systematic Biology* **43** 222–235.
- CRANDALL, K. A. (1995). Intraspecific phylogenetics: Support for dental transmission of human immunodeficiency virus. *Journal of Virology* **69** 2351–2356.
- CRANDALL, K. A. (1996a). Identifying links between genotype and phenotype using marker loci and candidate genes. In *The Impact of Plant Molecular Genetics*. (B. W. S. Sobral, eds.) 137–157. Birkhauser, Boston.
- CRANDALL, K. A. (1996b). Multiple interspecies transmissions of human and simian T-cell leukemia/lymphoma virus type I sequences *Molecular Biology and Evolution* **13** 115–131.
- CRANDALL, K. A. and TEMPLETON, A. R. (1993). Empirical tests of some predictions from coalescent theory with applications to intraspecific phylogeny reconstruction. *Genetics* **134** 959–969.
- CRANDALL, K. A. and TEMPLETON, A. R. (1999). Statistical methods for detecting recombination. In *Evolution of HIV*. (K. A. Crandall, ed.) in press. The Johns Hopkins University Press, Baltimore, MD.
- CRANDALL, K. A., TEMPLETON, A. R. and SING, C. F. (1994). Intraspecific phylogenetics: problems and solutions. In *Models in Phylogeny Reconstruction*. (R. W. Scotland, D. J. Siebert and D. M. Williams, eds.) 273–297. Clarendon Press, Oxford, England.

- DEBRY, R. W., ABELE, L. G., WEISS, S. H., HILL, M. D., BOZAS, M., LORENZO, E., GRAEBNITZ, F. and RESNICK, L. (1993). Dental HIV transmission? *Nature* **361** 691.
- DONNELLY, P. and TAVARE, S. (1986). The ages of alleles and a coalescent. *Advances in Applied Probability* **18** 1–19.
- DONNELLY, P. and TAVARE, S. (1995). Coalescents and genealogical structure under neutrality. *Annual Review of Genetics* **29** 401–421.
- EPSTEIN, L. G., KUIKEN, C., BLUMBERG, B. M. et al. (1991). HIV-1 V3 domain variation in brain and spleen of children with AIDS: tissue-specific evolution within host-determined quasispecies. *Virology* **180** 583–590.
- EWENS, W. J. (1990). Population genetics theory—the past and the future. In *Mathematical and Statistical Developments of Evolutionary Theory* (S. Lessard, ed.) 177–227. Kluwer Academic Publishers, New York, NY.
- FELSENSTEIN, J. (1985) Phylogenies and the comparative method. *American Naturalist* **125** 1–15.
- GOJOBORI, T., MORIYAMA, E. N., INA, Y., IKEO, K., MIURA, T., TSUJIMOTO, H., HAYAMI, M. and YOKOYAMA, S. (1990). Evolutionary origin of human and simian immunodeficiency viruses. *Proceedings of the National Academy of Sciences USA* **87** 4108–4111.
- HARVEY, P. H. and NEE, S. (1994). Phylogenetic epidemiology lives. *Trends in Ecology and Evolution* **9** 361–363.
- HEIN, J. (1990). Reconstructing evolution of sequences subject to recombination using parsimony. *Mathematical Bioscience* **98** 185–200.
- HEIN, J. (1993). A heuristic method to reconstruct the history of sequences subject to recombination. *Journal of Molecular Evolution* **36** 396–405.
- HILLIS, D. M. (1994). Homology in molecular biology. In *Homology: The Hierarchical Basis of Comparative Biology*. (B. K. Hall, ed.) 339–368. Academic Press, Inc., New York.
- HILLIS, D. M. (1995). Approaches for assessing phylogenetic accuracy. *Systematic Biology* **44** 3–16.
- HILLIS, D. M. and BULL, J. J. (1993). An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Systematic Biology* **42** 182–192.
- HILLIS, D. M., BULL, J. J., WHITE, M. E., BADGETT, M. R. and MOLINEUX, I. J. (1992). Experimental phylogenetics: Generation of a known phylogeny. *Science* **255** 589–591.
- HILLIS, D. M. and HUELSENBECK, J. P. (1994). Support for dental HIV transmission. *Nature* **369** 24–25.
- HIRSCH, V. M., OLMSTED, R. A., MURPHEY-CORB, PURCELL, R. H. and JOHNSON, P. R. (1989). An African primate lentivirus (SIVsm) closely related to HIV-2. *Nature* **339** 389–391.
- HOLMES, E. C. and GARNETT, G. P. (1994). Genes, trees and infections: molecular evidence in epidemiology. *Trends in Ecology and Evolution* **9** 256–260.
- HOLMES, E. C., ZHANG, L. Q., SIMMONDS, P., ROGERS, A. S. and BROWN, A. J. L. (1993). Molecular investigation of human immunodeficiency virus (HIV) infection in a patient of an HIV-infected surgeon. *Journal of Infectious Diseases* **167** 1411–1414.
- HUDSON, R. R. (1989). How often are polymorphic restriction sites due to a single mutation? *Theoretical Population Biology* **36** 23–33.
- HUDSON, R. R. (1990). Gene genealogies and the coalescent process. *Oxford Surveys in Evolutionary Biology* **7** 1–44.
- HUELSENBECK, J. P. (1995). Performance of phylogenetic methods in simulation. *Systematic Biology* **44** 17–48.
- HUELSENBECK, J. P. and HILLIS, D. M. (1993). Success of phylogenetic methods in the four-taxon case. *Systematic Biology* **42** 247–264.
- KELLAM, P. and LARDER, B. A. (1995). Retroviral recombination can lead to linkage of reverse transcriptase mutations that confer increased zidovudine resistance. *Journal of Virology* **69** 669–674.
- KORALNIK, I. J., BEORI, E., SAXINGER, W. C., MONICO, A. L., FULLEN, J., GESSAIN, A., GUO, H.-G., GALLO, R. C., MARKHAM, P., KALYANARAMAN, V., HIRSCH, V., ALLAN, J., MURTHY, K., ALFORD, P., SLATTERY, J. P., O'BRIEN, S. J. and FRANCHINI, G. (1994). Phyloge-

- netic associations of human and simian T-cell leukemia/lymphotropic virus type I strains: Evidence for interspecies transmission. *Journal of Virology* **68** 2693–2707.
- KUIKEN, C. L., GOUDSMIT, J., WEILLER, G. F., ARMSTRONG, J. S., HARTMAN, S., PORTEGIES, P., DEKKER, J. and CORNELISSEN, M. (1995). Differences in human immunodeficiency virus type 1 V3 sequences from patients with and without AIDS dementia complex. *Journal of Virology* **76** 175–180.
- KUIKEN, C. L., ZWART, G., BAAN, E., COUTINHO, R. A., VAN DEN HOEK, J. A. R. and GOUDSMIT, J. (1993). Increasing antigenic and genetic diversity of the V3 variable domain of the human immunodeficiency virus envelope protein in the course of the AIDS epidemic. *Proceedings of the National Academy of Sciences USA* **90** 9061–9065.
- LI, W.-H. and ZHARKIKH, A. (1995). Statistical tests of DNA phylogenies. *Systematic Biology* **44** 49–63.
- MADDISON, D. R. (1991a). African origin of human mitochondrial DNA reexamined. *Systematic Zoology* **40** 355–363.
- MADDISON, D. R. (1991b). The discovery and importance of multiple islands of most-parsimonious trees. *Systematic Zoology* **40** 315–328.
- MCCLURE, M. A. (1991). Evolution of retroposons by acquisition or deletion of retrovirus-like genes. *Molecular Biology and Evolution* **8** 835–856.
- MCMEARNEY, T., HORNICKOVA, Z., TEMPLETON, A., BIRDWELL, A., ARENS, M., MARKHAM, R., SAAH, A. and RATNER, L. (1995). Nef and LTR sequence variation from sequentially derived human immunodeficiency virus type 1 isolates. *Virology* **208** 388–398.
- MIYAMOTO, M. M. and FITCH, W. M. (1995). Testing species phylogenies and phylogenetic methods with congruence. *Systematic Biology* **44** 64–76.
- MYERS, G., BERZOFSKY, J. A., KORBER, B., SMITH, R. F. and PAVLAKIS, G. N. (1993). Human retroviruses and AIDS. Department of Theoretical Biology and Biophysics, Los Alamos National Laboratory.
- OU, C.-Y., CIESIELSKI, C. A., MYERS, G., BANDEA, C. I., LUO, C.-C., KORBER, B. T. M., MULLINS, J. I., SCHOCHELMAN, G., BERKELMAN, R. L., ECONOMOU, A. N., WITTE, J. J., FURMAN, L. J., SATTE, G. A., MACINNES, K. A., CURRAN, J. W., JAFFE, H. W., GROUP, L. I. and GROUP, E. I. (1992). Molecular epidemiology of HIV transmission in a dental practice. *Science* **256** 1165–1171.
- PAMILO, P. and NEI, M. (1988). Relationships between gene trees and species trees. *Molecular Biology and Evolution* **5** 568–583.
- ROBERTSON, D. L., HAHN, B. H. and SHARP, P. M. (1995a). Recombination in AIDS viruses. *Journal of Molecular Evolution* **40** 249–259.
- SAKSENA, N. K., HERVE, V., SHERMAN, M. P., DURAND, J. P., MATHIOT, C., MULLER, M., LOVE, J. L., LEGUENNO, B., SINOUSSI, F. B., DUBE, D. K. and POIESZ, B. J. (1993). Sequence and phylogenetic analyses of a new STLV-I from a naturally infected Tantalus monkey from Central Africa. *Virology* **192** 312–320.
- SANDERSON, M. J. and DOYLE, J. J. (1992). Reconstruction of organismal and gene phylogenies from data on multigene families: Concerted evolution, homoplasy, and confidence. *Systematic Biology* **41** 4–17.
- SHARP, P. M., ROBERTSON, D. L., GAO, F. and HAHN, B. H. (1994). Origins and diversity of human immunodeficiency viruses. *AIDS* **8** S27–S42.
- SHARP, P. M., ROBERTSON, D. L. and HAHN, B. H. (1996). Cross-species transmission and recombination of 'AIDS' viruses. In *New Uses for New Phylogenies* (P. H. Harvey, A. J. L. Brown, J. M. Smith and S. Nee, eds.) 134–152. Oxford University Press, Oxford.
- STRUNNIKOVA, N., RAY, S. C., LIVINGSTON, R. A., RUBALCABA, E. and VISCIIDI, R. P. (1995). Convergent evolution within the V3 loop domain of human immunodeficiency virus type 1 in association with disease progression. *Journal of Virology* **69** 7548–7558.
- SWOFFORD, D. L. (1993). PAUP: Phylogenetic Analysis Using Parsimony. 3.1.1. Smithsonian Institution, Washington, D. C.

- SWOFFORD, D. L., OLSEN, G. J., WADDELL, P. J. and HILLIS, D. M. (1996). Phylogenetic Inference. In *Molecular Systematics* (D. M. Hillis, C. Moritz and Mable, B. K., eds.) 407–514. Sinauer Associates, Inc., Sunderland, MA.
- TEMPLETON, A. R. (1992). Human origins and analysis of mitochondrial DNA sequences. *Science* **255** 737.
- TEMPLETON, A. R., BOERWINKLE, E. and SING, C. F. (1987). A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data I. Basic theory and an analysis of alcohol dehydrogenase activity in *Drosophila*. *Genetics* **117** 343–351.
- TEMPLETON, A. R., CRANDALL, K. A. and SING, C. F. (1992). A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. III. Cladogram estimation. *Genetics* **132** 619–633.
- TEMPLETON, A. R. and SING, C. F. (1993). A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. IV. Nested analyses with cladogram uncertainty and recombination. *Genetics* **134** 659–669.
- VOEVODIN, A., CRANDALL, K. A., CRANDALL, SETH, P. and MUFTI, S. A. (1996). HIV type 1 subtypes B and C from new regions of India and Indian and Ethiopian expatriates in Kuwait. *AIDS Research and Human Retroviruses* **12** 641–643.
- WATTERSON, G. A. and GUESS, H. A. (1977). Is the most frequent allele the oldest? *Theoretical Population Biology* **11** 141–160.
- WHITE, M. E., BULL, J. J., MOLINEUX, I. J. and HILLIS, D. M. (1991). Experimental phylogenies from T7 bacteriophage. In *Proceedings of the Fourth International Congress of Systematics and Evolutionary Biology*. (E. Dudley, eds.) 935–943. Dioscorides Press, Portland, Oregon.
- ZHARKIKH, A. and W.-H. LI (1995). Estimation of confidence in phylogeny: The complete-and partial bootstrap technique. *Molecular Phylogenetics and Evolution* **4** 44–63.

DEPARTMENT OF ZOOLOGY
574 WIDTSE Building
BRIGHAM YOUNG UNIVERSITY
PROVO, UT 84602-5255
KEITH_CRANDALL@BYU.EDU

LINEAR ESTIMATORS FOR THE EVOLUTION OF TRANSPOSABLE ELEMENTS

BY PAUL JOYCE¹, LINETTE FOX, N. CAROL CASAVANT, AND
HOLLY A. WICHMAN

*Department of Mathematics, Division of Statistics, University of Idaho,
 Department of Zoology and Genetics, Iowa State University and Department of
 Biological Sciences, University of Idaho*

Pairwise differences and segregating sites are two measures of sequence divergence often used to estimate the rate of evolution. There are other measures of sequence divergence. Which method is most appropriate? Motivated by a study of the evolution of transposable elements, we develop a new and more precise method for estimating the rate of evolution. We apply our method to LINE-1 data from Casavant *et al.* (1996).

1. Introduction. The motivation for this paper grew out of a very curious discovery made while analyzing some DNA sequence data. See Fox (1997). Under certain model assumptions for the evolution of transposable elements (mobile repetitive DNA found dispersed throughout the genome), we considered two simple measures of sequence divergence to estimate the rate of evolution. One estimate was based on the number of pairwise differences and the other was based on the number of segregating sites. We showed that the estimator based on the number of pairwise differences was inconsistent (variance of the estimator does not go to zero), while the segregating sites method was consistent. This is not at all surprising. The same story is true for the well studied neutral coalescent model, see Watterson (1975) and Tajima (1983). However, not only was it demonstrated that the segregating sites estimator is consistent under our model assumptions, but the variance of the estimator goes to zero like $1/n$, where n is the sample size. For evolutionary models involving DNA data it is unusual for estimators to have such good asymptotic properties.

The most curious discovery came when we applied each method to data. We found that in most cases of biological relevance, the pairwise difference estimator actually outperformed the segregating sites estimator. In this paper we resolve this apparent paradox.

We begin with a brief description of the relevant biology followed by the assumptions of the single master model used to analyze the data. We then

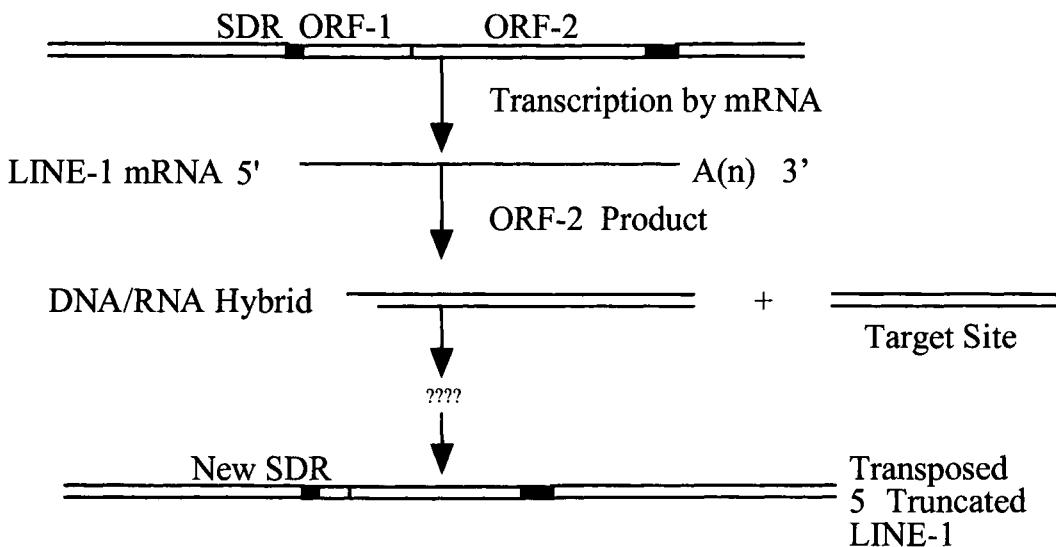
¹This research is supported by the National Science Foundation grant 96-26764.

AMS 1991 subject classifications. Primary 62F05; secondary 60G42, 60F15.

Key words and phrases. Coalescent, transposable elements, best linear unbiased estimator (BLUE).

FIG. 1. *LINE retrotransposition.*

Inserted Functional LINE-1



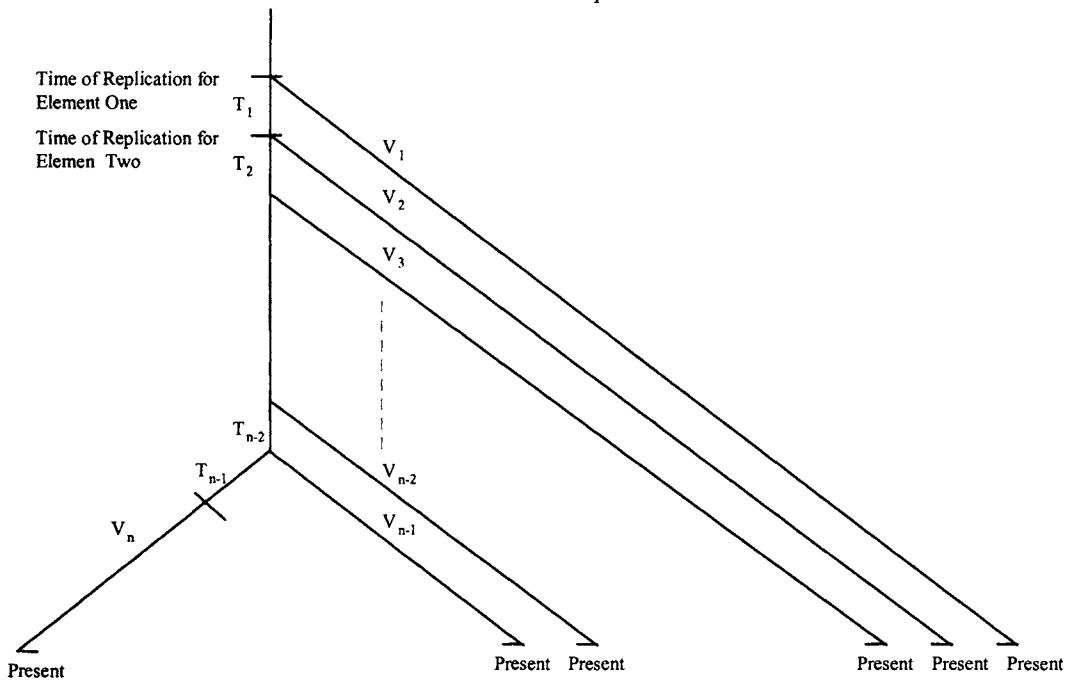
discuss pairwise difference and segregating sites estimators in the context of our model. The properties of both estimators are best understood when considered as members of a class of estimators called linear unbiased. We thus develop the theory of linear unbiased estimation in this context. Finally, we apply our new method to data. Fu (1994) also considered unbiased estimation in the context of the neutral coalescent model.

2. Mobile repetitive DNA. Mobile repetitive DNA sequences are found dispersed throughout the genome. LINEs (Long interspersed nuclear elements) have the capability of copying and inserting the copy into the genome at some other site by a process called retrotransposition. In this process DNA encodes RNA, RNA uses reverse transcription to code complementary DNA, and the DNA is integrated back into the chromosome at a different site in the genome. Figure 1 demonstrates one possible method of retrotransposition for LINEs. The functional LINE-1 with two open reading frames (ORF-1 and ORF-2) is transcribed by mRNA. The short direct repeats along the chromosomal DNA are shown by the filled in boxes and the open reading frames of the LINE-1 element by the open boxes. After transcription, RNA, which is shown as a single line, encodes a complementary DNA by reverse transcription. The reverse transcriptase protein in the ORF-2 region catalyzes the reverse transcription of

the RNA to form cDNA. The DNA/RNA hybrid integrates with the target site. The first open reading frame codes for a binding protein. This binding protein binds the DNA to the integration site. Retrotransposition usually produces a 5' truncated LINE-1. At integration, the RNA is detached. When the RNA detaches, the DNA often folds back on itself and primes second strand synthesis. The loop formed by this synthesis is broken to allow cDNA to synthesize to the chromosomal DNA. The result is a 5' truncation. Once integrated, the cDNA is ligated to the chromosomal DNA at the 3' end. Repair synthesis builds the second strand of DNA on the homologous chromosome.

2.1. Master copy model. The Master Copy Model assumes that one or a few elements in the genome have the capacity to replicate and all other elements are pseudogenes. A version of this model dates back to Kaplan and Hudson (1989). They developed an equilibrium master copy model where the number of copies reaches equilibrium due to the balance between duplication and deletion. They showed their model was consistent with the *Alu* divergence data. Recent studies of *Alu* and other SINE (short interspersed nuclear elements) families is consistent with a transient master copy model (Tachida 1993, 1996) that considers successive waves of expansion. We are interested in master copy models that are consistent with data from LINE-1 families in mammalian genomes, in particular the deer mouse, *Peromyscus* (Casavant *et al.* 1996). Unlike the *Alu* SINE data, there is evidence of fairly young LINE-1. This indicates that the LINE-1 elements under study may be in the midst of an expansion period. The purpose is to develop statistical methodology that can be applied to master copy data.

A simple mathematical description of the master copy model is the following. Consider a population of elements that is generated by a single master element giving birth at a constant rate. After a fixed time t has evolved, sample n individuals at random. The rooted tree (rooted at the time of the first offspring) describing the relationship between individuals in the sample, will have one main branch with all of the offspring branches emanating from the main branch (see Figure 2) The expected age of a randomly chosen element is $t/2$. Conditional on the tree, place marks along the branches according to a Poisson process of rate θ_m . The marks represent observed mutations that have occurred over evolutionary time. The Poisson process is independent along each branch. Count the number of marks on each of the branches. Assume the tree topology together with the number of marks on each branch is observable. Based on this observation, the problem is to estimate the parameter $\theta = \theta_m t$, the mean number of marks accumulated by a randomly chosen element. We rescale time so that the (rescaled) age of the master element is 1.

FIG. 2. *Element replication.*

Let V_i be the age of the i th oldest replicate in a sample of size n . The difference between the age of element i , V_i , and the next youngest element, V_{i-1} , is denoted by T_i . The age of the youngest element in the sample is V_{n-1} and $V_{n-1} = T_{n-1}$. Let P_i be the number of private mutations accumulated by the i th element over the time period V_i . Let S_i be the number of shared variants between the i th and $i-1$ th element accumulated over the time period T_i . Under the assumption of constant rate of element replication, the age of a randomly chosen element is uniformly distributed. So the ages of the elements are distributed according to the order statistics of the uniform. However, the master element will not appear in the sample. For this reason, we cannot determine which among the two youngest elements in the sample is indeed the youngest. This means that the coalescent time T_{n-1} is on average $2/(n+1)$, where as all other coalescent times are $E(T_i) = 1/(n+1)$ for $i < n-1$.

It will be convenient to use vector notation. Vectors will be column vectors unless superscripted by ' for transpose: thus we write

$$\mathbf{T}' = (T_1, T_2, \dots, T_{n-1})$$

and similarly

$$\mathbf{B}' = (V_1, V_2, \dots, V_{n-1}, T_1, T_2, \dots, T_{n-1}).$$

It is convenient to record the private and shared mutations in the following order

$$\mathbf{S}' = (P_1, P_2, \dots, P_{n-1}, S_1, S_2, \dots, S_{n-2}, P_n).$$

Let $\mu = E(\mathbf{T})$. We denote the variance of the $(n - 1) \times 1$ vector \mathbf{T} by the $(n - 1) \times (n - 1)$ matrix, $\Sigma = \text{Var}\mathbf{T}$, where $(\text{Var}\mathbf{T})_{ij} = \text{Cov}(T_i, T_j)$.

The branch lengths of a tree are related to the coalescent times, in that each branch is a sum of coalescent times. We may view the vector of branch lengths as a linear transformation of the coalescent times. If there are $n - 1$ coalescent times, there will be $2(n - 1)$ branches to the tree. As one traces the ancestry of the individuals, each coalescence introduces two new branches.

The linear transformation between coalescent times and branch lengths is given by an $(2n - 2) \times (n - 1)$ matrix \mathbf{c} , where the entries of \mathbf{c} are $c_{ij} = 1$ or 0,

$$\mathbf{c}\mathbf{T} = \mathbf{B}.$$

For the master locus model, relationship between branch lengths and coalescent times is given by

$$\mathbf{c} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 0 & 1 & \cdots & 1 \\ \vdots & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ \vdots & & & \ddots \\ 0 & 0 & \cdots & 1 \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

where $\mathbf{c}\mathbf{T} = \mathbf{B}$,

$$\mu' = \left(\frac{1}{n+1}, \frac{1}{n+1}, \dots, \frac{1}{n+1}, \frac{2}{n+1} \right)$$

and

$$\Sigma = \begin{pmatrix} \frac{n}{(n+1)^2(n+2)} & \frac{-1}{(n+1)^2(n+2)} & \cdots & \frac{-2}{(n+1)^2(n+2)} \\ \frac{-1}{(n+1)^2(n+2)} & \frac{n}{(n+1)^2(n+2)} & \cdots & \frac{-2}{(n+1)^2(n+2)} \\ \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \frac{-2}{(n+1)^2(n+2)} & \frac{-2}{(n+1)^2(n+2)} & \cdots & \frac{2(n-1)}{(n+1)^2(n+2)} \end{pmatrix}$$

3. Pairwise differences versus segregating sites. Using the tree in Figure 2 one can formulate an unbiased estimate of the mean number of mutations on a randomly selected element using the segregating sites method. Private mutations P_i occur along the elements over the age of the element according to a Poisson process, and shared mutations S_i occur along the master between replication events. The total number of segregating sites S is

$$S = \sum_{i=1}^{n-2} S_i + \sum_{i=1}^n P_i.$$

If θ_m is the mutation rate, then an unbiased estimator for the parameter $\theta = \theta_m t$ can be calculated to be

$$\hat{\theta}_{ss} = \frac{2(n-1)}{n^2 + 3n - 2} S.$$

We omit the details. The variance of the unbiased estimate using the segregating sites estimator is given as

$$\text{Var}(\hat{\theta}_{ss}) = \frac{n^4 + 4n^3 + 5n^2 + 26n - 24}{3(n^2 + 3n - 2)^2(n+2)} \theta^2 + \frac{2(n+1)}{n^2 + 3n - 2} \theta.$$

When time is distributed uniformly, the times between replication events are not independent, which complicates the variance. For a detailed derivation of this formula see Fox (1997). An asymptotic formula with only the leading terms is easier to absorb

$$\text{Var}(\hat{\theta}_{ss}) \approx \frac{1}{3n} \theta^2 + \frac{2}{n} \theta.$$

For large values of n (relative to mean number of mutations accumulated by a randomly chosen element) the segregating sites method produces a small variance and is consistent.

The pairwise difference estimator sums the mutations for all pairs of elements. The sum of the pairwise differences, D , can be written as

$$D = (n - 1) \sum_{i=1}^n P_i + \sum_{i=1}^{n-1} i(n - i) S_i$$

The pairwise difference estimator counts some mutations more often than others. An example illustrates the higher weights given to some mutations. If ten elements are sampled from the population and sorted by their times of replication, shared mutations that occur between the time of the replication of the 5th element and the replication of the 6th element are counted 25 times. Moreover, shared mutations between the replication of the ninth and tenth elements are counted nine times. Thus, the shared mutations between the 5th and 6th elements are counted more often and given a higher weight.

An unbiased estimator of θ based on the mean number of pairwise differences is given by

$$\hat{\theta}_{pd} = \frac{3}{4} \bar{D}.$$

The variance for the pairwise differences unbiased estimator is given by the following equation

$\text{Var}(\hat{\theta}_{pd})$

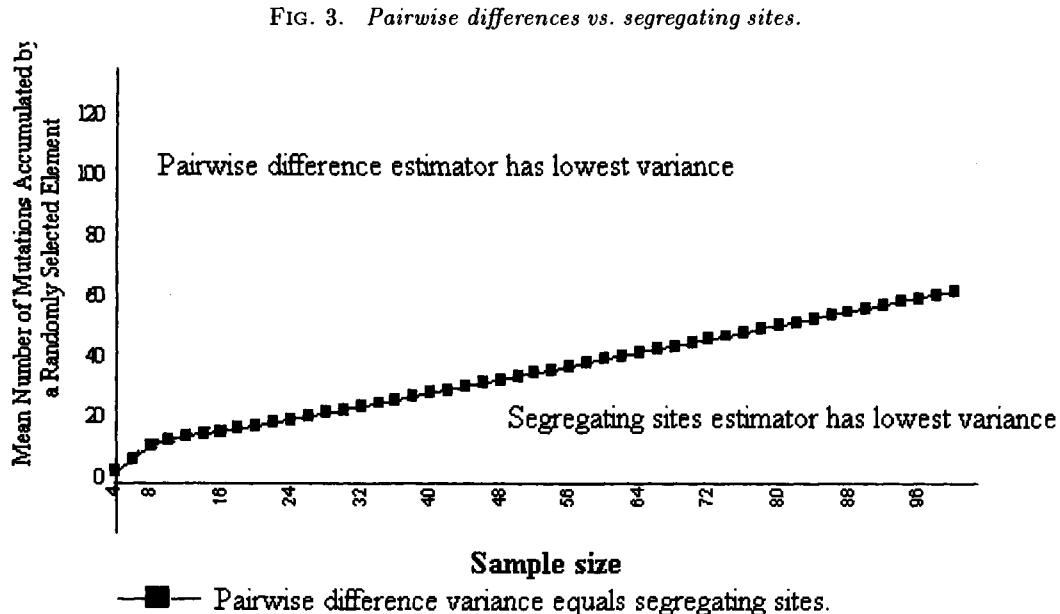
$$\begin{aligned} &= \left(\frac{(2n-1)(2n^3 + 15n^2 - 41n + 6)}{20n(n-1)(n+2)(n+1)^2} + \frac{9n^3}{4n^2(n-1)^2(n+1)^2(n+2)} \right) \theta^2 \\ &\quad + \left(\frac{9}{8n} + \frac{3(n^2 + 1)}{40n(n-1)} \right) \theta. \end{aligned}$$

An asymptotic formula including only leading terms may be easier to absorb

$$\text{Var}(\hat{\theta}_{pd}) \approx \frac{1}{5n} \theta^2 + \frac{3}{40} \theta.$$

Note that the variance of the pairwise difference estimator can never be smaller than $3\theta/40$, regardless of the sample size.

We introduced two familiar methods for estimating the parameter θ . One was based on the number of pairwise differences and the other on the number of segregating sites. It is not surprising to learn that the estimator based on



pairwise differences is inconsistent (that is, the error of the estimate does not go to zero as the sample size increases). One familiar with these types of problems may also expect that the estimate of θ based on the segregating sites is consistent. However, unlike the results of standard coalescent models,

1. the asymptotic properties of the segregating sites estimator are quite good

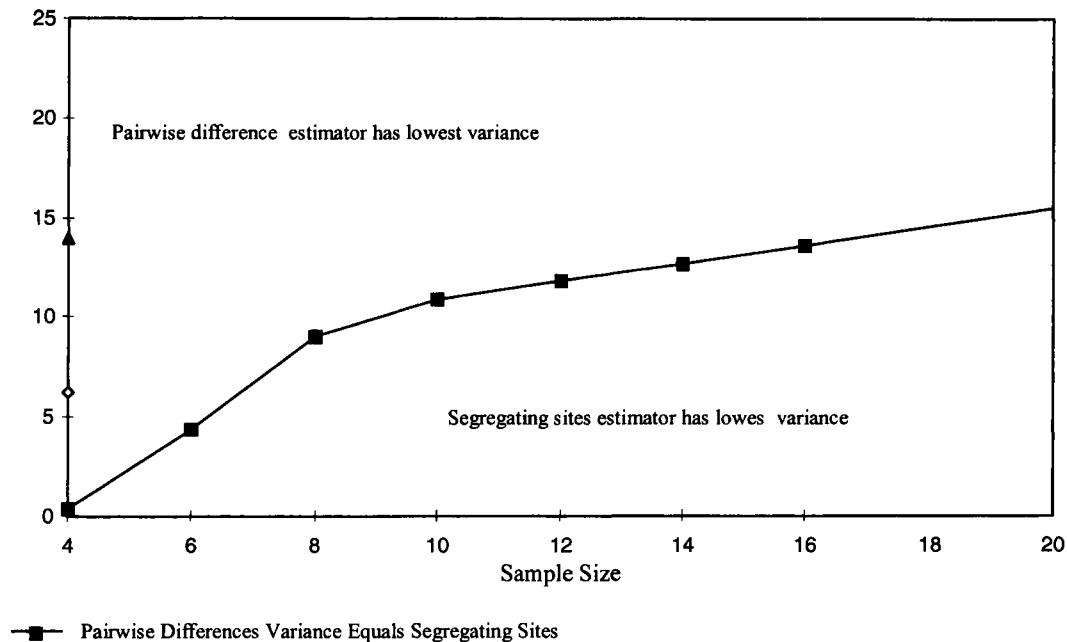
$$\text{Var}(\hat{\theta}_{ss}) \rightarrow 0$$

as $n \rightarrow \infty$ like $1/n$

2. for sample sizes and parameter values of practical interest, the inconsistent estimator based on pairwise differences actually outperforms the estimator based on segregating sites. All points above the line in the graphs below (Figures 3, 4) represent (n, θ) where the pairwise difference method outperforms segregating sites.

Resolving the apparent contradiction between statements (1) and (2) leads to the theory of linear unbiased estimators, and a new and better estimation procedure.

3.1. *Analyzing the sources of error.* If $\hat{\theta}$ is any estimate of the parameter θ , then we use the following conditioning argument to calculate the variance of $\hat{\theta}$

FIG. 4. *Pairwise differences vs. segregating sites (small sample size).*

given by

$$\text{Var}(\hat{\theta}) = \text{Var}(E(\hat{\theta}|\mathbf{B})) + E(\text{Var}(\hat{\theta}|\mathbf{B})).$$

The above variance formula is central to understanding linear unbiased estimation. There are two sources of error in any estimate of θ . The first source is due to the stochastic nature of the coalescent process. That is, coalescent times are random. The second is due to the stochastic nature of the mutation process, that is, mutations accumulate over time on each branch according to a Poisson process.

The first source of error is measured by the quantity

$$\text{Var}[E(\hat{\theta}|\mathbf{B})].$$

We refer to this as the **coalescent error**, because it is due to the stochastic nature of coalescent. The second source is measured by the quantity

$$E[\text{Var}(\hat{\theta}|\mathbf{B})].$$

We refer to this as the **Poisson error**.

We will see that there is always a trade off between the two sources of error. If one chooses an estimate that has small coalescent error, it is likely to have larger Poisson error.

TABLE 1

Comparing the variance for segregating sites, pairwise differences, and CBLUE. The entries are the coalescent errors (CE), Poisson errors (PE) and the variances of the estimate (V).

θ	Sample Size	Segregating Sites			Pairwise Differences			CBLUE		
		CE	PE	V	CE	PE	V	CE	PE	V
9	9	2.20	1.70	3.90	2.03	1.89	3.93	0.82	2.62	3.43
17	6	11.00	4.57	15.58	10.29	4.76	15.05	6.02	6.04	12.06
18	14	6.14	2.28	8.43	5.26	2.91	8.17	1.45	4.23	5.68
15	7	7.52	3.53	11.04	7.06	3.75	10.81	3.57	4.93	8.50

It can be shown that the segregating sites estimator has the smallest Poisson error of all linear unbiased estimators. However, it has larger coalescent error than the estimator based on pairwise differences. Table 1 below shows that the coalescent error is often the dominant term for the variance of the estimate. Our new estimation procedure minimizes the coalescent error and thus outperforms pairwise differences and segregating sites in most cases. Our preliminary results show that this new method compares favorably to maximum likelihood.

4. Coalescent best linear unbiased estimator (CBLUE). The purpose of this section is to consider unbiased estimators of θ that can be described as linear combinations of the observed changes on the branches. Let \mathbf{x} be an $(2n - 2) \times 1$ vector of weights. We consider estimates of θ of the form

$$\hat{\theta} = \sum_{i=1}^{2n-2} x_i S_i = \mathbf{x}' \mathbf{S}.$$

If we assume that $\hat{\theta}$ is an unbiased estimator, then we have a linear constraint on the set of possible weight vectors \mathbf{x} . Note that

$$E(\hat{\theta}) = E(\mathbf{x}' \mathbf{S}) = \theta E(\mathbf{x}' \mathbf{B}) = \theta E(\mathbf{x}' \mathbf{c} \mathbf{T}) = \theta \mathbf{x}' \mathbf{c} \boldsymbol{\mu}.$$

Therefore, if $\hat{\theta}$ is unbiased, then $E(\hat{\theta}) = \theta$ implies $\mathbf{x}' \mathbf{c} \boldsymbol{\mu} = 1$. We now calculate

the variance of $\hat{\theta}$ as

$$\begin{aligned}
 \text{Var}(\hat{\theta}) &= E[\text{Var}(\hat{\theta}|\mathbf{B})] + \text{Var}[E(\hat{\theta}|\mathbf{B})] \\
 &= E[\text{Var}(\mathbf{x}'\mathbf{S}|\mathbf{B})] + \text{Var}[E(\mathbf{x}'\mathbf{S}|\mathbf{B})] \\
 &= E[\mathbf{x}'\text{Var}(\mathbf{S}|\mathbf{B})\mathbf{x}] + \text{Var}[\mathbf{x}'E(\mathbf{S}|\mathbf{B})] \\
 &= \theta[\mathbf{x}'E(\text{diag}(\mathbf{B}))\mathbf{x}] + \theta^2\text{Var}(\mathbf{x}'\mathbf{B}) \\
 &= \theta[\mathbf{x}'E(\text{diag}(\mathbf{c}\mathbf{T}))\mathbf{x}] + \theta^2\text{Var}(\mathbf{x}'\mathbf{c}\mathbf{T}) \\
 &= \theta[\mathbf{x}'E(\text{diag}(\mathbf{c}\mathbf{T}))\mathbf{x}] + \theta^2[\mathbf{x}'\mathbf{c}\text{Var}(\mathbf{T})\mathbf{c}'\mathbf{x}]
 \end{aligned}$$

where $\text{diag}(\mathbf{B})$ is a diagonal matrix with $(\text{diag}(\mathbf{B}))_{ii} = B_i$ and $\text{diag}(\mathbf{B})_{ij} = 0$ for $i \neq j$.

Let $\mathbf{y} = \mathbf{c}'\mathbf{x}$ and let $\mathbf{M} = E(\text{diag}(\mathbf{c}\mathbf{T}))$ then we can write

$$(4.1) \quad \text{Var}(\hat{\theta}) = \theta[\mathbf{x}'\mathbf{M}\mathbf{x}] + \theta^2[\mathbf{y}'\Sigma\mathbf{y}].$$

Since the estimator must be unbiased we have the constraint $\boldsymbol{\mu}'\mathbf{y} = 1$. An estimator that minimizes the coalescent error is found by minimizing the quadratic form $\mathbf{y}'\Sigma\mathbf{y}$, subject to the linear constraint $\boldsymbol{\mu}'\mathbf{y} = 1$. Thus there will be exactly one \mathbf{y} that makes the coalescent error minimum. Notice that $\mathbf{y} = \mathbf{c}'\mathbf{x}$ is a system of $n - 1$ linear equations with $2n - 2$ unknowns. Because there are typically many solutions, there will be many choices of \mathbf{x} that minimize the coalescent error. We then pick among these choices the one that minimizes the Poisson error. We call this the coalescent best estimators.

Definition. Let $\mathcal{C} = \{\hat{\theta} = \mathbf{x}'\mathbf{S} \mid \mathbf{x}'\mathbf{c}\boldsymbol{\mu} = 1, \text{Var}(E(\hat{\theta}|\mathbf{B})) \leq \text{Var}(E(\tilde{\theta}|\mathbf{B})) \text{ for all unbiased linear estimators } \tilde{\theta}\}$. $\hat{\theta}_c$ is the coalescent best linear unbiased estimator (CBLUE) if

1. $\hat{\theta}_c \in \mathcal{C}$
2. $E(\text{Var}(\hat{\theta}_c|\mathbf{B})) \leq E(\text{Var}(\tilde{\theta}|\mathbf{B}))$ for all $\tilde{\theta} \in \mathcal{C}$.

Lemma 1. *The collection of unbiased estimators that minimize the coalescent error \mathcal{C} is a linear subspace of \mathbb{R}^{2n-2} given by $\mathcal{C} = \{\hat{\theta} = \mathbf{x}'\mathbf{S} \mid \mathbf{x} \text{ is a solution to the following linear system } \frac{\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}{\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}} = \mathbf{c}'\mathbf{x}\}$, where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the mean and variance of the coalescent times.*

Proof. It follows from equation (4.1) that if \mathbf{y} minimizes the quadratic form $\mathbf{y}'\Sigma\mathbf{y}$ subject to the linear constraint, $\boldsymbol{\mu}'\mathbf{y} = 1$, then any solution to the linear equations $\mathbf{y} = \mathbf{c}'\mathbf{x}$ will produce a $\hat{\theta}$ in \mathcal{C} . We need only to show that the solution to the minimization problem is $\mathbf{y} = \frac{\Sigma^{-1}\boldsymbol{\mu}}{\boldsymbol{\mu}'\Sigma^{-1}\boldsymbol{\mu}}$. We use the method of Lagrange multipliers. Define

$$g(\mathbf{y}) = \mathbf{y}'\Sigma\mathbf{y} - \lambda(\mathbf{y}'\boldsymbol{\mu} - 1).$$

Then the derivative of g is given by

$$\frac{dg}{d\mathbf{y}}(\mathbf{y}) = 2\Sigma\mathbf{y} - \lambda\boldsymbol{\mu}.$$

Setting the derivative equal to zero and solving for the critical number gives

$$\mathbf{y} = \frac{\lambda}{2}\Sigma^{-1}\boldsymbol{\mu}.$$

Since $\boldsymbol{\mu}'\mathbf{y} = 1$ implies $\frac{\lambda}{2}\boldsymbol{\mu}'\Sigma^{-1}\boldsymbol{\mu} = 1$. Therefore, $\lambda/2 = 1/(\boldsymbol{\mu}'\Sigma^{-1}\boldsymbol{\mu})$, implying

$$\mathbf{y} = \frac{\Sigma^{-1}\boldsymbol{\mu}}{\boldsymbol{\mu}'\Sigma^{-1}\boldsymbol{\mu}}.$$

This completes the proof since $\mathbf{y} = \mathbf{c}'\mathbf{x}$ by definition. ■

Theorem 1. *There exists a unique CBLUE, $\hat{\theta}_c$, for θ . If \mathbf{T} is the vector of coalescent times, with $E(\mathbf{T}) = \boldsymbol{\mu}$, $\text{Var}(\mathbf{T}) = \Sigma$ and $E(\text{diag}(\mathbf{cT})) = \mathbf{M}$ then the CBLUE estimate is given by $\hat{\theta}_c = \mathbf{S}'\mathbf{x}$, where*

$$(4.2) \quad \mathbf{x} = \mathbf{M}^{-1}\mathbf{c} (\mathbf{c}'\mathbf{M}^{-1}\mathbf{c})^{-1} \frac{\Sigma^{-1}\boldsymbol{\mu}}{\boldsymbol{\mu}'\Sigma^{-1}\boldsymbol{\mu}}.$$

Proof. It follows from the Lemma 1 and equation (4.1) that the CBLUE will be the \mathbf{x} that minimizes the quadratic form $\mathbf{x}'\mathbf{M}\mathbf{x}$, subject to $n - 1$ linear constraints induced by the linear equation $\mathbf{y} = \mathbf{c}'\mathbf{x}$, where $\mathbf{y} = \frac{\Sigma^{-1}\boldsymbol{\mu}}{\boldsymbol{\mu}'\Sigma^{-1}\boldsymbol{\mu}}$. Again we use the method of Lagrange multipliers. This time the Lagrange multiplier is a $(n - 1) \times 1$ vector denoted by $\boldsymbol{\lambda}$. Define

$$h(\mathbf{x}) = \mathbf{x}'\mathbf{M}\mathbf{x} - (\mathbf{x}'\mathbf{c} - \mathbf{y}')\boldsymbol{\lambda}.$$

then the derivative of h is given by

$$\frac{dh}{d\mathbf{x}}(\mathbf{x}) = 2\mathbf{M}\mathbf{x} - \mathbf{c}\boldsymbol{\lambda}$$

TABLE 2
Estimates and standard deviations of Peromyscus data.

Estimation Method	<i>Peromyscus</i> Species	Sample Size	Mean Number of Mutations Per Element Estimate	Estimated Standard Deviation
Segregating sites	<i>P. californicus</i>	9	8.68	1.92
	<i>P. maniculatus</i>	6	16.42	3.83
Pairwise differences	<i>P. californicus</i>	9	8.63	1.92
	<i>P. maniculatus</i>	6	17.35	3.95
CBLUE	<i>P. californicus</i>	9	7.90	1.67
	<i>P. maniculatus</i>	6	18.08	2.89

Setting the derivative equal to zero and solving for \mathbf{x} gives

$$\mathbf{x} = \frac{1}{2}\mathbf{M}^{-1}\mathbf{c}\boldsymbol{\lambda}.$$

Substituting the above \mathbf{x} into the linear constraint $\mathbf{y} = \mathbf{c}'\mathbf{x}$ and solving for $\boldsymbol{\lambda}/2$ gives

$$\frac{1}{2}\boldsymbol{\lambda} = (\mathbf{c}'\mathbf{M}^{-1}\mathbf{c})^{-1}\mathbf{y}$$

which implies

$$\mathbf{x} = \mathbf{M}^{-1}\mathbf{c}(\mathbf{c}'\mathbf{M}^{-1}\mathbf{c})^{-1}\mathbf{y}. \blacksquare$$

We applied the CBLUE estimator given by $\hat{\theta}_c = \mathbf{S}'\mathbf{x}$ to two lineages of LINE-1 for two species of *Peromyscus* (deer mouse). The results are given in Table 2 below. Note that CBLUE method is a significant improvement over pairwise differences and segregating sites. A more complete analysis of the data can be found in Joyce *et al.*, in preparation.

5. Conclusion. The master locus model can be used to estimate rates of evolution and make comparisons for LINE-1 data. However, using traditional measures of sequence divergence to estimate the evolutionary parameters leads to the following puzzling conclusion. While the method of pairwise differences leads to an inconsistent estimator, it is in many cases more precise than the consistent estimator based on the number of segregating sites.

The puzzle is resolved when one realizes that the variance of a linear estimator is of the form

$$a_n\theta^2 + b_n\theta.$$

If θ is relatively large, then the $a_n\theta^2$ term will dominate the error in the estimate. This term is smaller for pairwise differences than for segregating sites.

By choosing an estimator that minimizes the dominant term of the error, one can often improve on both methods. This estimator is called the coalescent best linear unbiased estimator CBLUE.

While the CBLUE estimator was demonstrated for the master locus model, it applies in a more general setting. However, it can be shown that there does not exist a uniformly best linear unbiased estimator.

Acknowledgements. The authors would like to thank the American Mathematical Society for sponsoring the meeting 'Statistics in Molecular Biology,' at the University of Washington, Seattle in June 1997. A special thanks goes to Dr. Françoise Seillier-Moiseiwitsch for all her efforts in organizing the conference and editing this proceedings.

REFERENCES

- CASAVANT, N. C., SHERMAN, A. N. and WICHMAN, H. A. (1996). Two Persistent LINE-1 Lineages in *Peromyscus* Have Unequal Rates of Evolution. *Genetics* **142** 1289–1298.
- FU, Y.-X. (1994). A Phylogenetic Estimator of Effective Population Size or Mutation Rate. *Genetics* **136** 685–692.
- FOX, L. (1997). Statistical Methods for Analyzing Transposable Elements. Unpublished Masters Thesis. University of Idaho.
- JOYCE, P., FOX, L., CASAVANT, N. C., WICHMAN, H. A. and FOSTER, J. Statistical models and methods for LINE-1 evolution, in preparation.
- KAPLAN, N. L. and HUDSON, R. R. (1989). An Evolutionary Model for Highly Repeated Interspersed DNA Sequences. In *Mathematical Evolutionary Theory*, Feldman, M. W. (ed). Princeton University Press, New Jersey, 301–314.
- TACHIDA, H. (1996). A Population Genetic Study of the Evolution of SINEs II Sequence Evolution Under the Master Copy Model. *Genetics* **143** 1033–1042.
- WATTERSON, G. A. (1975). On the Number of Segregating Sites in Genetical Models Without Recombination. *Theoretical Population Biology* **7** 256–276.

DIVISION OF STATISTICS
 DEPARTMENT OF MATHEMATICS
 UNIVERSITY OF IDAHO
 MOSCOW, ID 83844-1103
 JOYCE@UIDAHO.EDU
 LFOX@PSG.HS.EDU

DEPARTMENT OF PLANT PATHOLOGY
 IOWA STATE UNIVERSITY
 AMES, IA 50011
 CASAVANT@IASTATE.EDU

DEPARTMENT OF BIOLOGICAL SCIENCES
 UNIVERSITY OF IDAHO
 MOSCOW, ID 83844-3051
 HWICHMAN@UIDAHO.EDU

A CONDITIONAL APPROACH TO THE DETECTION OF CORRELATED MUTATIONS

BY MAHA C. KARNOUB, FRANCOISE SEILLIER-MOISEIWITSCH¹ AND PRANAB K. SEN

Glaxo Wellcome Inc., Research Triangle Park, NC and University of North Carolina at Chapel Hill

Some genomes mutate quickly. Studying their mutation process may allow us to understand the selection pressures these genomes are undergoing. Considering simultaneous mutations at several sites may provide insights into protein structure. We consider the situation where the frequency table for the amino-acid pairs can be summarized by a 2×2 table. We develop a test for mutational linkage between two positions conditionally on the consensus pair. We illustrate the use of this test with sequences from the V3 loop of the envelope gene from the human immunodeficiency virus.

1. Introduction. The genome of a retrovirus like the human immunodeficiency virus (HIV) evolves at a fast pace: the high mutation rate (due to the error-prone reverse transcriptase and recombination between the two RNA strands) is compounded to the high rate of replication (~ 300 cycles of replication per year). Some of these substitutions confer a survival advantage by enabling some mutants to escape the immune system. Some may cause a phenotypic change (such as cell tropism and virulence). These random mutations thus persist in the viral population. Some subsist only when a substitution at another position occurs. These *linked* mutations may simply maintain structure (and thus viability) or may again be beneficial to the virus.

When the structure of a viral protein is unknown, detecting such double mutations may help in inferring pairs of amino acids that interact and are therefore more likely to be in close spatial proximity. Further, for sequence data, more often than not the probabilistic model underlying analytical methods relies on the assumption that positions undergo independent mutation processes. The methodology introduced in this paper can serve to check this assumption. Its violation leads to the overestimation of genetic distances and thus to erroneous phylogenetic reconstructions [Seillier-Moiseiwitsch et al. (1998)].

In Section 2, we consider specific pairs of sites and test the number of dou-

¹Her research was funded in part by the National Science Foundation (DMS-9305588), the American Foundation for AIDS Research (70428-15-RF) and the National Institutes of Health (R29-GM49804 and P30-HD37260).

AMS 1991 subject classifications. Primary 62F03 ; secondary 62P10, 92D20.

Key words and phrases. Correlation, independence, mutation, phylogeny.

TABLE 1
Contingency table for sequence data.

		Position 2			
		V	W		
Position 1		D	n_{11}, p_{11}	n_{12}	$n_{1.}$
		E	n_{21}	n_{22}	$n_{2.}$
			p_{21}	p_{22}	$p_{2.}$
			$n_{.1}$	$n_{.2}$	n
			$p_{.1}$	$p_{.2}$	

ble mutations away from the *consensus* (i.e., the most frequent configuration) conditioning on the total number of sequences and on the consensus pair; a conventional test for independence is appraised in this context. The methodological developments are presented in Sections 3, 4, and 5, with some details relegated to the Appendix. We present some simulation results in Section 5 and analyze a set of HIV-1 sequences (Section 6).

2. The set-up. Consider a specific pair of sites along the sequences, say Position 1 and Position 2. Assume that at Position 1, across all sequences, amino acids D and E are present while Position 2 exhibits V and W . In the contingency table summarizing the data (Table 1), the $(1, 1)$ -cell contains the number of sequences with the consensus configuration. Let n_{ij} be the number of sequences in row i (Position 1) and column j (Position 2), and p_{ij} the probability of having this configuration. Any sequence in the first row or the first column but not in the $(1, 1)$ -cell sustained a single mutation away from the consensus. The others had two mutations.

In the context of viral sequences, it is reasonable to treat these sequences as independent, at least as a first assumption. Indeed, replication cycles are short, and each cycle generates a number of substitutions. Hence, when all sequences are sampled from different individuals, many rounds of replication separate any two viruses. Each position had many opportunities to mutate. Also, functionality of the resulting protein drives viral survival. Thus, whether a specific position is allowed or required to change depends on the amino-acid composition at other locations. In a sense, the consensus is "rediscovered" after

each alteration through structural linkage. This selection pressure and the high viral turn-over overwhelm ancestral relationships.

We are interested in testing excess or paucity of double mutations under the assumption of independence of mutations at the two positions. The random variable of interest is N_{22} , the number of sequences with double mutations. In Fisher's exact (conditional) test [Bishop et al. (1975), p.364], the marginal totals $n_{1..}$, $n_{.1}$, $n_{2..}$, and $n_{.2}$ are conditioned upon and the conditional probability of the observed counts is given by the hypergeometric law

$$Pr(N_{ij} = n_{ij}, i, j = 1, 2 \mid n_{1..}, n_{.1}, n_{2..}, n_{.2}) = \frac{n_{1..}! n_{2..}! n_{.1}! n_{.2}!}{n! n_{11}! n_{12}! n_{21}! n_{22}!} .$$

The exact one- and two-sided tests using the collection of all possible tables with these marginal totals and this probability distribution can be constructed. Alternatively, a large-sample (normal) approximation can be applied as follows.

The estimate of the expected value of N_{22} is

$$\hat{E}(N_{22}) = n \hat{p}_{22} .$$

Under the null hypothesis of independence among positions,

$$\hat{p}_{22} = \frac{n_{2..} n_{.2}}{n^2} \quad \text{and} \quad \hat{E}(N_{22}) = \frac{n_{2..} n_{.2}}{n} .$$

The exact variance conditional on the marginal totals is

$$Var\left(N_{22} - \hat{E}(N_{22})\right) = \frac{n_{1..} n_{.1} n_{2..} n_{.2}}{n^2 (n-1)} .$$

The test statistic is thus

$$T_n = \left(N_{22} - \hat{E}(N_{22}) \right) / \sqrt{\frac{n_{1..} n_{.1} n_{2..} n_{.2}}{n^2 (n-1)}} .$$

Under H_0 , $T_n \sim \mathcal{N}(0, 1)$,

and the critical level of this test statistic is close to the normal percentile point corresponding to the significance level.

In the present context, the marginal totals are not fixed as we do not have any knowledge about the total number of mutations with a specific nucleotide or amino acid at any one of the positions. Since in our set-up, we condition on the consensus cell frequency N_{11} ($= n_{11}$) (see Section 3 for further motivation); hence, Fisher's exact (conditional) test is not appropriate here. However, we may add that as the conditional distribution of T_n , given the marginal totals, is asymptotically $\mathcal{N}(0, 1)$, in probability (under H_0), and thus free of the marginal

totals, the asymptotic unconditional null distribution of T_n is also $\mathcal{N}(0, 1)$. Consequently, a test based on T_n will have a specified level of significance if the N_{ij} 's are all large. Nevertheless, in terms of power properties it might not compare favourably with alternative tests based on the conditional distribution given N_{11} and N . In passing, we may remark that the traditional optimality (viz., UMP) property that may be attributed to Fisher's exact test (against some specific parametric alternative) may not be tenable in our set-up where the alternative hypotheses are not only of more complex nature but also not entirely of a parametric flavour. In fact, in our case, a UMP test may not exist. For this reason, we have recourse to a direct approach based on the conditional distribution, given N_{11} and N along with the additional information that N_{11} is the maximum cell-count among the four cells. This is the main theme of the current study.

3. Distribution of the cell counts. The standard distributions associated with cell counts in contingency tables are [Bishop et al. (1975)]: the Poisson model, obtained with a sampling plan that has no restrictions on the total sample size, the multinomial model with a fixed total sample size, and independent multinomial distributions for the rows (with fixed row totals) or independent multinomial distributions for the columns (with fixed column totals). For these sampling models, the marginal totals are sufficient statistics for testing the independence of two factors. Further, under the assumption of independence of the factors, the maximum-likelihood estimates under the above sampling processes exist, are unique and, if none of the marginal totals is 0, they are equal. In fact, when the total sample size is fixed, the multinomial and the Poisson schemes are equivalent [Bishop et al. (1975)]. The Poisson model is usually preferred when some of the events are rare, i.e. some of the cell counts are small. In view of the nature of our data, we adopt this model here. Indeed, the measured average error rate per site for HIV-1 reverse transcriptase is between 10^{-4} and 10^{-3} .

We let $\mathbf{N} = (N_{11}, N_{12}, N_{21}, N_{22})'$, and make the following assumptions:

1. the cell counts N_{ij} are independent and have Poisson distributions with parameters λ_{ij} , i.e.,

$$(3.1) \quad P\{\mathbf{N} = \mathbf{n}\} = \prod_{i=1}^2 \prod_{j=1}^2 \frac{e^{-\lambda_{ij}} \lambda_{ij}^{n_{ij}}}{n_{ij}!} \quad \mathbf{n} \geq \mathbf{0}$$

2. $\lambda_{11} \gg \lambda_{ij}$ for all $(i, j) \neq (1, 1)$ and the λ_{ij} 's are large (e.g., greater than 5).

These two assumptions basically ensure that (i) N_{11} is a maximum with probability close to 1, and (ii) the multinormality approximations hold for the distribution of the N_{ij} 's.

Let $\lambda = \sum_{i,j} \lambda_{ij}$. We consider the null hypothesis (of independence of mutations)

$H_0 : \lambda_{ij} = \lambda \alpha_i \beta_j$ for all (i, j) , $\alpha_1 + \alpha_2 = 1$, $\beta_1 + \beta_2 = 1$,
against the alternative hypothesis

$H_a : \lambda_{ij}$'s are not factorizable this way,
i.e., the mutations at the two positions are not stochastically independent.

We test the above hypothesis conditionally on the consensus pair. Conditioning on the consensus pair allows us to identify the polymorphisms (i.e., the pairs made up of the consensus amino acid at one position and a mutation at the other position) and the double mutations. The goal of the test is to identify pairs of positions that show a propensity to change together. The reason is two-fold. First, the consensus pair codes for a part of the protein that plays its function well. Departures at one or the other of the two amino acids (but not at both simultaneously) indicate changes that do not disturb the structure so much that the product is no longer a functioning protein. These single departures can be viewed as "noise". Double mutations, on the other hand, come about either as a rescue mechanism for a slightly deleterious single mutation or as a change in phenotype (acquisition of a different cell tropism, for example). Hence, identifying correlated substitutions serves as an exploratory investigation of structure modifications and thus of alterations in phenotypes. Second, usually, positions that interact are in the same three-dimensional vicinity. Hence, from these correlated pairs, one can infer some information about the structure of the protein. For these two purposes determining the consensus pair is crucial to identifying changes. Further, due to genetic drift and independent selection pressures at each position, α_1 and β_1 are not constant over time. Thus, conditioning on N and N_{11} enable us to separate these effects from selection that act on both positions simultaneously.

Let $n_* = n - n_{11}$ and $\lambda_* = \lambda - \lambda_{11}$. Under Assumptions 1 and 2 above, we show that if the λ_{ij} 's are large

$$(3.2) \quad P\{N_{11} > \max(N_{12}, N_{21}, N_{22})\} \rightarrow 1$$

(see Appendix). Next, consider the joint distribution of the cell counts, given $N_{11} = n_{11}$ and the fixed total sample size n :

$$\begin{aligned} P\{N_{11} = n_{11}, N_{12} = n_{12}, N_{21} = n_{21}, N_{22} = n_{22} | N_{11} = n_{11}\} \\ = P\{N_{12} = n_{12}, N_{21} = n_{21}, N_{22} = n_{22}\} \\ = e^{-\lambda_*} \left(\frac{\lambda_{12}^{n_{12}} \lambda_{21}^{n_{21}} \lambda_{22}^{n_{22}}}{n_{12}! n_{21}! n_{22}!} \right), \end{aligned}$$

which follows readily from (3.1). As a result, the distribution of $N_* = N_{21} + N_{12} + N_{22}$, given $N_{11} = n_{11}$, is

$$P\{N_* = n_* \mid N_{11} = n_{11}\} = \frac{e^{-\lambda_*} \lambda_*^{n_*}}{n_*!}.$$

Further, note that given $N_{11} = n_{11}$, $N_* = n_*$, we have $N = n = n_* + n_{11}$. Hence, from the above two equations, we obtain that

$$P\{N_{11} = n_{11}, N_{12} = n_{12}, N_{21} = n_{21}, N_{22} = n_{22} \mid N_{11} = n_{11} \text{ & } N = n\}$$

$$(3.3) \quad = \frac{n_*!}{n_{12}! n_{21}! n_{22}!} \left(\frac{\lambda_{12}}{\lambda_*} \right)^{n_{12}} \left(\frac{\lambda_{21}}{\lambda_*} \right)^{n_{21}} \left(\frac{\lambda_{22}}{\lambda_*} \right)^{n_{22}}.$$

This is a trinomial distribution for $\mathbf{N}_* = (N_{12}, N_{21}, N_{22})'$, with parameters n_* and $\frac{\lambda_{12}}{\lambda_*}$, $\frac{\lambda_{21}}{\lambda_*}$, and $\frac{\lambda_{22}}{\lambda_*}$. Let $\nu_{ij} = \frac{\lambda_{ij}}{\lambda_*}$ for $(i, j) \neq (1, 1)$, $\boldsymbol{\nu} = (\nu_{12}, \nu_{21}, \nu_{22})'$ and $\mathbf{D} = \text{Diag}(\nu_{12}, \nu_{21}, \nu_{22})$. Then

$$(3.4) \quad \begin{aligned} E(\mathbf{N}_* \mid N_{11} = n_{11}, N_* = n_*) &= n_* \boldsymbol{\nu} \\ \text{Var}(\mathbf{N}_* \mid N_{11} = n_{11}, N_* = n_*) &= n_* [\mathbf{D} - \boldsymbol{\nu} \boldsymbol{\nu}']. \end{aligned}$$

At this stage, we invoke (3.2) and (3.3), and claim that as n increases,

$$(3.5) \quad \begin{aligned} P\{N_{12} = n_{12}, N_{21} = n_{21}, N_{22} = n_{22} \mid N_{11} = n_{11}, N_{11} = \max \text{ and } N = n\} \\ \approx P\{(N_{12} = n_{12}, N_{21} = n_{21}, N_{22} = n_{22} \mid N_{11} = n_{11}, N_* = n_*\}. \end{aligned}$$

Therefore, the moment results in (3.4) can also be shown to be good approximations for the conditional case where in addition N_{11} is the maximum.

4. Parameter estimation. By virtue of (3.5), under Assumptions 1 and 2, we shall work with the approximate likelihood function (given $N_{11} = n_{11}$, $N_* = n_*$ and that N_{11} is the maximum cell count among the four cells):

$$(4.1) \quad \begin{aligned} P\{N_{12} = n_{12}, N_{21} = n_{21}, N_{22} = n_{22} \mid N_{11} = n_{11} = \max, N = n\} \\ \equiv \frac{n_*!}{n_{12}! n_{21}! n_{22}!} \nu_{12}^{n_{12}} \nu_{21}^{n_{21}} \nu_{22}^{n_{22}}. \end{aligned}$$

Recall that, under H_0 , $\lambda_{ij} = \lambda \alpha_i \beta_j$ for $i = 1, 2$, and $j = 1, 2$. Without any loss of generality, we set $\alpha_1 + \alpha_2 = 1 = \beta_1 + \beta_2$. With this simplification, we can express $\nu_{ij} = \lambda_{ij}/\lambda_*$ in terms of the two unknowns, α_1 and β_1 . The trinomial law in (4.1) has effectively two degrees of freedom (DF) for a goodness-of-fit (GOF) test statistic. The conventional Pearsonian GOF test for the conditional law in

(4.1) would result in 0 DF, and hence, would not be usable. To eliminate this impasse, we recall that N_{11} is held fixed (along with other statistical information contained in its marginal distribution), and further that

$$P\{N_{11} = n_{11} \mid N = n\} = \binom{n}{n_{11}} \left(\frac{\lambda_{11}}{\lambda}\right)^{n_{11}} \left(1 - \frac{\lambda_{11}}{\lambda}\right)^{n-n_{11}},$$

which under the null hypothesis H_0 reduces to

$$(4.2) \quad P\{N_{11} = n_{11} \mid N = n, H_0\} = \binom{n}{n_{11}} (\alpha_1 \beta_1)^{n_{11}} (1 - \alpha_1 \beta_1)^{n-n_{11}}.$$

Therefore, letting $\theta = \alpha_1 \beta_1$, from (4.2), we obtain a MLE or BAN estimator of θ :

$$\hat{\theta}_n = \frac{n_{11}}{n}.$$

As such, we work with the conditional model in (4.1) incorporating the additional restraint that

$$(4.3) \quad \alpha_1 \beta_1 = \hat{\theta}_n = \frac{n_{11}}{n}.$$

Now we have effectively only one unknown parameter, α_1 say, and hence, the classical inference theory for categorical models can be called upon.

The log-likelihood of α_1 , based on (4.1) and the constraint (4.3), is

$$\begin{aligned} L_0(\alpha_1) &= C + n_{12} \log \alpha_1 + n_{12} \log \left(1 - \frac{\hat{\theta}_n}{\alpha_1}\right) + n_{21} \log (1 - \alpha_1) \\ &\quad + n_{21} \log \frac{\hat{\theta}_n}{\alpha_1} + n_{22} \log (1 - \alpha_1) + n_{22} \log \left(1 - \frac{\hat{\theta}_n}{\alpha_1}\right), \end{aligned}$$

where C comprises the terms that do not depend on α_1 . Then

$$\begin{aligned} \frac{\partial L_0(\alpha_1)}{\partial \alpha_1} &= -\frac{n_{2.}}{\alpha_1} + \frac{n_{.2}}{\alpha_1 - \hat{\theta}_n} - \frac{n_{2.}}{1 - \alpha_1} = \frac{n_{.2}}{\alpha_1 - \hat{\theta}_n} - \frac{n_{2.}}{\alpha_1(1 - \alpha_1)}, \\ (4.4) \quad \frac{\partial^2 L_0(\alpha_1)}{\partial \alpha_1^2} &= \frac{n_{2.}}{\alpha_1^2} - \frac{n_{.2}}{\left(\alpha_1 - \hat{\theta}_n\right)^2} - \frac{n_{2.}}{(1 - \alpha_1)^2}. \end{aligned}$$

In the region where $\frac{n_{2.}}{n} = \alpha_2 + O_p(1/\sqrt{n})$ and $\frac{n_{.2}}{n} = \beta_2 + O_p(1/\sqrt{n})$, it follows by routine computations that

$$\frac{\partial^2 L_0(\alpha_1)}{\partial \alpha_1^2} < 0, \text{ in probability as } n \rightarrow \infty.$$

We obtain the solution to the estimating equation in (4.4) as (in probability)

$$(4.5) \quad \tilde{\alpha}_1 = \frac{1}{2} \left\{ 1 - \frac{n_{2.}}{n_{.2}} + \sqrt{\left(1 - \frac{n_{2.}}{n_{.2}} \right)^2 + 4 \hat{\theta}_n \frac{n_{2.}}{n_{.2}}} \right\}.$$

Further,

$$\tilde{\alpha}_2 = 1 - \tilde{\alpha}_1, \quad \tilde{\beta}_1 = \frac{\hat{\theta}_n}{\tilde{\alpha}_1}, \quad \tilde{\beta}_2 = 1 - \tilde{\beta}_1.$$

We incorporate these estimators in our proposed test statistic in Section 5.

5. The new binomial test. Our proposed test statistic, based on the restraint that

$$\tilde{\alpha}_1 \tilde{\beta}_1 = \hat{\theta}_n = \frac{n_{11}}{n},$$

is

$$\tilde{Z}_{22} = \frac{1}{\sqrt{n}} \left(N_{22} - n_* \frac{\tilde{\alpha}_2 \tilde{\beta}_2}{\left(1 - \tilde{\alpha}_1 \tilde{\beta}_1 \right)} \right) = \frac{1}{\sqrt{n}} \left(N_{22} - n \tilde{\alpha}_2 \tilde{\beta}_2 \right).$$

Note that the original 2×2 table has 3 DF, and hence, having the estimators $\hat{\theta}_n$ and $\tilde{\alpha}_1$ ($\tilde{\beta}_1 = \hat{\theta}_n / \tilde{\alpha}_1$), we have effectively one DF. For this reason, it suffices to use only either \tilde{Z}_{22} or \tilde{Z}_{12} or \tilde{Z}_{21} , defined analogously. However, since we are interested in double mutations, \tilde{Z}_{22} is intuitively more appealing.

We may provide a natural interpretation of \tilde{Z}_{22} in terms of the κ coefficient for agreement (no mutation or double mutation, in our set-up) for categorical data. Following Cohen (1960) [see also Landis & Koch (1977)], we define

$$\kappa = \frac{\pi_o - \pi_e}{1 - \pi_o}$$

where π_o is the probability of no or double mutation and π_e is the hypothetical probability of the same under the baseline constraints that $p_{11} = \alpha_1 \beta_1$ and $p_{22} = (1 - \alpha_1)(1 - \beta_1)$ ($= \alpha_2 \beta_2$). Thus, here

$$\pi_o = p_{11} + p_{22} \quad \text{and} \quad \pi_e = p_{11} + (1 - \alpha_1)(1 - \beta_1),$$

so that $1 - \pi_e = \alpha_1 + \beta_1 - 2\alpha_1\beta_1 = \alpha_1\alpha_2 + \beta_1\beta_2 + (\alpha_1 - \beta_1)^2$. Note that by virtue of the consensus pairing, $p_{11} = \alpha_1\beta_1$, \tilde{Z}_{22} is the sample counterpart of the numerator of κ . As we shall see later, though we have not included the

denominator, the factor $(\alpha_1 + \beta_1 - 2\alpha_1\beta_1)$ shows up in the sampling variance of \tilde{Z}_{22} . The main reason for not including the denominator in \tilde{Z}_{22} is that $(\alpha_1 + \beta_1 - 2\alpha_1\beta_1)$ can be very small when both α_1 and β_1 are close to 1 (as is the case here), which might make the κ coefficient look much inflated.

In order to find the critical level for the test statistic, we need to obtain an expression for its sampling variance under H_0 . Let

$$Z_{ij} = \frac{1}{\sqrt{n}} \left(N_{ij} - n \frac{\lambda_{ij}}{\lambda} \right), \quad i = 1, 2, \quad j = 1, 2.$$

Note that under H_0

$$Z_{ij} = \frac{1}{\sqrt{n}} (N_{ij} - n \alpha_i \beta_j).$$

Hence, appealing to the multinomial law in (3.1), we have under H_0 ,

$$Z_{ij} \sim \mathcal{N}(0, \alpha_i \beta_j (1 - \alpha_i \beta_j)), \quad i = 1, 2, \quad j = 1, 2.$$

Let $U_n = \frac{N_{2.}}{N_{.2}}$. Noting that by (4.5), $U_n \tilde{\beta}_2 = \tilde{\alpha}_2$, we rewrite

$$\tilde{Z}_{22} = \frac{1}{\sqrt{n}} (N_{22} - n U_n^{-1} \tilde{\alpha}_2^2).$$

Also, we show in Appendix 2 that

$$(5.1) \quad \frac{N_{2.}}{N_{.2}} = \frac{\alpha_2}{\beta_2} + \frac{1}{\sqrt{n} \beta_2^2} (Z_{2.} \beta_2 - Z_{.2} \alpha_2) + O_p(n^{-1})$$

with $\mu = \alpha_2/\beta_2$. Therefore, we obtain from the above equation that

$$U_n^{-1} = \frac{1}{\mu} - \frac{1}{\sqrt{n} \mu^2 \beta_2} (Z_{2.} - \mu Z_{.2}) + O_p(n^{-1}).$$

We also rewrite $\tilde{\alpha}_1$ as

$$\tilde{\alpha}_1 = \frac{1}{2} \left\{ 1 - U_n + \sqrt{(1 - U_n)^2 + 4 \hat{\theta}_n U_n} \right\} = g(U_n, \hat{\theta}_n), \text{ say,}$$

where by direct substitution, it follows that

$$\begin{aligned} g(\mu, \hat{\theta}_n) &= \alpha_1, \\ g'(\mu, \hat{\theta}_n) &= -\frac{\alpha_1(1 - \beta_1)^2}{\alpha_1 + \beta_1 - 2\alpha_1\beta_1}. \end{aligned}$$

From the last four equations, we obtain

$$\begin{aligned}\tilde{Z}_{22} &= \frac{1}{\sqrt{n}} \left\{ N_{22} - \frac{n(1-\alpha_1)}{\mu} \left([1 - \frac{(U_n - \mu)}{\mu}] [1 + \alpha_1 - 2g'(\mu, \hat{\theta}_n)(U_n - \mu)] \right. \right. \\ &\quad \left. \left. + O_p(n^{-1}) \right) \right\} \\ &= d_{12}Z_{12} + d_{21}Z_{21} + d_{22}Z_{22} + O_p(n^{-1/2}),\end{aligned}$$

where

$$\begin{aligned}d_{12} &= -\frac{(\beta_1 - \alpha_1)(1 - \alpha_1)}{\alpha_1 + \beta_1 - 2\alpha_1\beta_1}, \quad d_{21} = \frac{(\beta_1 - \alpha_1)(1 - \beta_1)}{\alpha_1 + \beta_1 - 2\alpha_1\beta_1} \\ d_{22} &= \frac{\alpha_1(1 - \alpha_1) + \beta_1(1 - \beta_1)}{\alpha_1 + \beta_1 - 2\alpha_1\beta_1}.\end{aligned}$$

Let E_0 denote the expected value and Var_0 the variance under H_0 . Then

$$\begin{aligned}E_0(\tilde{Z}_{22}) &= 0 + O(n^{-1/2}) = o(1), \\ Var_0(\tilde{Z}_{22}) &= \sum_{(i,j) \neq (1,1)} (d_{ij})^2 \alpha_i \beta_j - \frac{n}{n_*} (\alpha_2 \beta_2)^2 = \gamma_{22} \text{ (say).}\end{aligned}$$

We may note that when $\alpha_1 = \beta_1$,

$$d_{12} = d_{21} = 0 \quad \text{and } Var_0(\tilde{Z}_{22}) = \alpha_2^2(1 - \alpha_1^2 - \alpha_2^2)/(1 - \alpha_1^2).$$

In summary, based on the consensus count and the resulting 2×2 table, we propose the following test statistic

$$(5.2) \quad \frac{N_{22} - n \tilde{\alpha}_2 \tilde{\beta}_2}{\sqrt{n \hat{\gamma}_{22}}}$$

where $\hat{\gamma}_{22}$ is obtained from γ_{22} by substituting the estimates of the α_i 's and β_j 's as obtained earlier into the d_{ij} 's.

Table 2 shows the results of simulations performed to assess the empirical validity of the above statistic. For each sample size and underlying probability distribution, 1,000 contingency tables were generated under the hypothesis of independence of rows and columns. \underline{r} and \underline{c} denote the row and column probability vector, respectively. From the data in each of the simulated contingency tables, the above test statistic was computed. Then for each of these tables, \underline{r} and \underline{c} were estimated, and these estimates were utilized to construct 1,000 bootstrap further tables, keeping the (1,1)-cell fixed. The empirical distribution of the test statistics calculated from the bootstrap tables is used as the reference

TABLE 2

*Results of 1,000 simulations under the assumption of independence between rows and columns. × * indicates that a percentage is two standard deviations away from its expected value and ** that it is three standard deviations away.*

Sample Size	1	5	10	90	95	99
$r=(.60,.40), c=(.60,.40)$						
100	0.1 **	0.3 **	1.7 **	0.1 **	0 **	0 **
500	0	0.1 **	0.4 **	0.1 **	0 **	0 **
1,000	0 **	0 **	0 **	0.1 **	0 **	0 **
$r=(.70,.30), c=(.60,.40)$						
100	0.6	3.1 *	7.3 *	13.2 **	7.8 **	2.0 **
200	0.8	5.5	10.9	12.5 *	7.5 **	2.0 **
300	0.5	4.1	9.8	13.3 **	7.7 **	2.1 **
400	0.8	5.2	9.4	9.5	5.8	1.2
$r=(.80,.20), c=(.70,.30)$						
100	0.2 *	3.5 *	9.0	12.3 *	7.3 **	2.2 **
200	0.3 *	4.8	10.0	10.8	6.6 *	1.9 *
300	0.6	4.3	8.5	9.2	5.3	1.4

distribution of the observed statistic. The entries of Table 2 are the percentages of the test statistics falling below the 1st, 5th and 10th percentiles and above the 90th, 95th and 99th percentiles of the bootstrap reference distribution distribution. These simulations show that, though the (1,1)-cell is assigned the largest frequency, the test is not appropriate for the scenario where $r = c = (.60, .40)$. For the other probabilistic set-ups, the observed numbers get closer to their expected values as the sample size increases. Hence, the asymptotic result derived in this paper is achieved with sample sizes of 300 to 400, depending on the underlying distribution.

6. Data analysis. We consider 141 HIV-1 T-cell-adapted sequences. Each sequence is 35 amino acids long and spans the V3 loop of the envelope protein gene. This region varies highly in composition across individuals and within individuals, and has been found to be functionally important [Korber et al. (1993), Potts et al. (1993)]. This data set does not contain duplicate sequences and a person contributes a single sequence. To avoid detecting linkages that, while declared statistically significant, have little scientific importance, we only considered pairs of positions for which the double-mutation count is at least 4.

For illustration purposes, we show in Table 3 the counts for positions 5 and 10. These data were aggregated into a 2×2 table (Table 4). The results appear in Table 5. The bootstrap distribution for the test statistics is computed by the BC_a method [Efron & Tibshirani (1993)].

TABLE 3
Contingency table for positions 5 and 10.

		Position 10					
		K	R	N	S	Q	G
Position 5	N	87	27	1	2	2	1
	H	0	3	0	0	0	0
	S	4	4	0	0	0	0
	G	4	0	0	0	0	0
	K	0	1	0	0	0	0
	D	1	1	0	0	0	0
	Y	1	0	0	1	1	0

TABLE 4
Aggregated table for positions 5 and 10.

	K	&
N	87	33
&	10	11

The global variation of the virus is divided into *groups* and *clades* within a group (groups O, M and N; clades A-J within group M). Each clade/group exhibits a different consensus sequence. Sequences from the same clade cluster in the same geographic region. These clades may result from major natural selection pressures. The above test is helpful in detecting parallel evolution within a clade (here, the sequences belong to clade B which covers North America, Western Europe and Thailand). Were one to analyze sequences from different clades together, the linkages would reflect correlations in the consensus sequences. Then a binomial test like (5.2) is of little value as it merely consider the number of double mutations, i.e. the number in each clade (and thus depends on the sampling procedure and not on a biological mechanism). More relevant would be χ^2 -square-type tests which consider specific amino-acid pairings.

7. Discussion. Our interest lies in investigating whether departures from the consensus amino acids at two positions are correlated. The assessment of whether there has been, at a specific position, a substitution away from the consensus clearly relies on the knowledge of this consensus. This is essentially the reasoning behind our conditioning on the number of pairs in the consensus cell. This conditioning affects the probability law for the test statistic by basically reducing its variance. Further, there are two statistical arguments in favour of conditioning on N_{11} . First, in a large $r \times c$ contingency table (potentially 20 \times 20) with many empty and low-frequency cells, the exact test conditional on

TABLE 5
Results for 141 HIV-1 sequences.

Positions	Test statistics	p-values
5 , 10	2.442	.005 < p < .01
7 , 22	0.324	> .1
8 , 10	5.291	< .0005
10, 23	1.354	.05 < p < .1
10 , 24	2.833	.001 < p < .005
10 , 34	3.243	.001 < p < .005
12 , 19	1.926	.01 < p < .05
12 , 22	1.245	> .1
12, 29	1.340	> .1
14 , 19	9.210	< .0005
14 , 20	4.481	< .0005
19 , 20	6.647	< .0005
19 , 22	-0.898	> .1
19 , 32	2.513	.005 < p < .01
21 , 22	2.781	.001 < p < .005
21 , 24	4.306	< .0005
22 , 23	3.296	< .0005
22 , 24	5.548	< .0005
22 , 34	0.536	> .1
23 , 24	5.721	< .0005

fixed marginals has little power. The table thus needs to be reduced. Here, to do so, we use the consensus cell. Hence, since this consensus is data dependent, it affects the resulting structure of the table and probability model. Second, the size of N_{11} affects the power of the test. It is indeed of a different order of magnitude to the other entries (in the example of Section 6, N_{11} is 87 and the next largest entry is 33). By conditioning on N_{11} , we gain power.

Through a simple test for the equality of two binomial proportions, one can verify that the consensus pair is indeed truly maximal. When λ_{11} is not very large compared to the other λ_{ij} 's, there is no obvious consensus pair and thus this conditioning cannot be applied. One then needs to resort to the usual χ^2 -test for independence in contingency tables. However, the absence of a clear consensus often indicates, in our experience, that two or more populations of sequences are represented in the sample. Then the analysis proceeds by considering each subpopulation separately.

The theory developed in this paper relies on the independence of the sequences. This assumption is violated in many instances. However, it would be straightforward to alter the reference distribution so that it would take into

account the evolutionary process undergone by the sequences under study. This could be done by simulating the evolutionary process and generating a large number of sets of sequences. The test statistic is computed on each set, to construct a reference distribution. For the sequences utilized in Section 6, the replication rate for HIV is very high (once every one to two days). Each replication introduces one to ten substitutions along the whole genome. As many rounds of replication are likely to separate two sequences (i.e., the total time to their most recent ancestor along the two branches), sharing the same amino acid at a polymorphic site is due to structure restriction and not phylogenetic relationships. Thus, these sequences can be regarded as independent, and the test introduced here can be applied.

Acknowledgements. We thank the referees for helpful comments.

REFERENCES

- BISHOP, Y., FIENBERG, S. and HOLLAND, P. (1975). *Discrete Multivariate Analysis: Theory and Practice*. The MIT Press, Cambridge, Massachusetts.
- COHEN, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurements* **20** 37–46.
- EFRON, B. and TIBSHIRANI, R. (1993). *An Introduction to the Bootstrap*. Chapman and Hall, New York.
- KARNOUB, M. (1997). Understanding Dependencies among Mutations along the HIV Genome, PhD thesis, Department of Biostatistics, University of North Carolina at Chapel Hill.
- KORBER, B.T.M., FARBER, R.M., WOLPERT, D.H. and LAPEDES, A.S. (1993). Covariation of mutations in the V3 loop of the human immunodeficiency virus type 1 envelope protein: An information-theoretic analysis. *Proceedings of the National Academy of Sciences U.S.A.* **90** 7176–7180.
- LANDIS, J.R. and KOCH, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics* **33** 159–174.
- POTTS, K.E., KALISH, M.L., BANDEA, C.I., ORLOFF, G.M., ST. LOUIS, M., BROWN, C., MALANDA, N., KAVUKA, M., SCHOCHELMAN, G., OU, C. and HEYWARD, W.L. (1993). Genetic diversity of human immunodeficiency virus type 1 strains in Kinshasa, Zaïre. *AIDS Research and Human Retroviruses* **9** 613–618.
- SEILLIER-MOISEIWITSCH, F., PINHEIRO, H., KARNOUB, M. and SEN, P.K. (1998). Novel methodology for quantifying genomic heterogeneity. In *Proceedings of the Joint Statistical Meetings, Anaheim, California, August 1997*.

Appendix 1: Proof of (3.2)

Note that

$$\begin{aligned}
 P\{N_{11} \text{ is maximum}\} \\
 &= P\{N_{ij} < N_{11} \quad \forall (i, j) \neq (1, 1)\} \\
 &= E\{P\{N_{ij} < N_{11} \quad \forall (i, j) \neq (1, 1) | N_{11}\}\} \\
 &= E\{P\{N_{12} < N_{11} | N_{11}\} P\{N_{21} < N_{11} | N_{11}\} P\{N_{22} < N_{11} | N_{11}\}\}
 \end{aligned}$$

where for large λ_{ij} 's, we use the square-root transformation on the Poisson variates. Thus, under Assumption 2 in Section 3, for each $(i, j) \neq (1, 1)$,

$$\begin{aligned} P\{N_{ij} < N_{11} | N_{11}\} \\ &= P\{\sqrt{N_{ij}} - \sqrt{\lambda_{ij}} < \sqrt{N_{11}} - \sqrt{\lambda_{11}} | N_{11}\} \\ &= P\{W_{ij} < W_{11} + 2(\sqrt{\lambda_{11}} - \sqrt{\lambda_{ij}}) | W_{11}\} \end{aligned}$$

where the W_{ij} are independent and asymptotically normally distributed with zero mean and unit variance, and therefore are all bounded in probability. On the other hand, under Assumption 2, $\sqrt{\lambda_{11}} - \sqrt{\lambda_{ij}}$ is large and positive for any $(i, j) \neq (1, 1)$. Hence, the above probability converges to 1 as λ_{11} increases, which satisfies Assumption 2. Therefore, their product also converges to 1, and hence, being a bounded random variable, their expectation has the same limit, i.e. 1.

Appendix 2: Proof of (5.1)

$$\begin{aligned} \frac{N_{2.}}{N_{.2}} &= \frac{\sqrt{n} Z_{2.} + n \alpha_2}{\sqrt{n} Z_{.2} + n \beta_2} \\ &= \frac{\alpha_2}{\beta_2} + \frac{1}{\sqrt{n} \beta_2^2} \frac{Z_{2.} \beta_2 - Z_{.2} \alpha_2}{1 + \frac{Z_{.2}}{\sqrt{n} \beta_2}}, \end{aligned}$$

where noting that $Z_{.2} = O_p(1)$, we have

$$\left\{ 1 + \frac{Z_{.2}}{\sqrt{n} \beta_2} \right\}^{-1} = 1 - \frac{Z_{.2}}{\sqrt{n} \beta_2} + O_p(n^{-1}),$$

and hence

$$\frac{N_{2.}}{N_{.2}} = \frac{\alpha_2}{\beta_2} + \frac{1}{\sqrt{n} \beta_2^2} (Z_{2.} \beta_2 - Z_{.2} \alpha_2) + O_p(n^{-1}).$$

GLAXO WELLCOME INC.
FIVE MOORE DRIVE
P.O.Box 13398
RESEARCH TRIANGLE PARK, NC 27709-3398
MAC31876@GLAXOWELLCOME.COM

DEPARTMENT OF BIOSTATISTICS
SCHOOL OF PUBLIC HEALTH
UNIVERSITY OF NORTH CAROLINA
CHAPEL HILL, NC 27599-7400
SEILLIER@BIOS.UNC.EDU
PKSEN@BIOS.UNC.EDU

CORRELATED MUTATIONS IN MODELS OF PROTEIN SEQUENCES: PHYLOGENETIC AND STRUCTURAL EFFECTS

BY ALAN S. LAPEDES¹, BERTRAND G. GIRAUD, LONCHANG LIU AND GARY D. STORMO

Los Alamos National Laboratory, Santa Fe Institute, Service Physique Théorique, DSM, C.E.N. Saclay and University of Colorado

Covariation analysis of sets of aligned sequences for RNA molecules is relatively successful in elucidating RNA secondary structure, as well as some aspects of tertiary structure [Gutell et al. (1992)]. Covariation analysis of sets of aligned sequences for protein molecules is successful in certain instances in elucidating certain structural and functional links [Korber et al. (1993)], but in general, pairs of sites displaying highly covarying mutations in protein sequences do not necessarily correspond to sites that are spatially close in the protein structure [Gobel et al. (1994), Clarke (1995), Shindyalov et al. (1994), Thomas et al. (1996), Taylor & Hatrick (1994), Neher (1994)]. In this paper we identify two reasons why naive use of covariation analysis for protein sequences fails to reliably indicate sequence positions that are spatially proximate. The first reason involves the bias introduced in calculation of covariation measures due to the fact that biological sequences are generally related by a non-trivial phylogenetic tree. We present a null-model approach to solve this problem. The second reason involves linked chains of covariation which can result in pairs of sites displaying significant covariation even though they are not spatially proximate. We present a maximum entropy solution to this classic problem of “causation versus correlation”. The methodologies are validated in simulation.

1. Introduction. Analysis of sets of aligned sequences, such as RNA or protein sequences, is a common procedure in bioinformatic analysis. Various methods have been developed to describe aligned sequences: “consensus” sequences which are determined by the most conserved symbol in each sequence position; “profiles” [Gribskov et al. (1987)] which represent the probability distribution of symbols in each position, and can also include inserts and deletes with fixed position independent penalties; and “hidden Markov models” [Krogh et al. (1994)], which represent single site probability distributions as well as position dependent probability distributions for insertions and deletions. Correlation analysis extends such methods to consideration of the probability distribution for pairs of symbols in all possible pairs of positions in the sequence. “Mutual information”, a measure of correlation for discrete symbols [Cover &

¹Research supported by the Department of Energy under contract W-7405-ENG-36

AMS 1991 subject classifications. Primary 62F03 ; secondary 62P10, 92D20.

Key words and phrases. Correlated mutations, phylogeny, interaction, structure from sequence.

Thomas (1991)], quantifies the covariation of mutations for pairs of positions in biological sequences Gutell et al. (1992), Korber et al. (1993)].

Mutual information can be expressed in numerous equivalent ways, some of which derive from information theory, hence the name. In this paper we will not use information theoretic expressions involving entropy [see e.g. Korber et al. (1993)], but will instead use the following formula

$$M = \sum_{ab} P_{ab} \log P_{ab} / P_a P_b$$

where P_{ab} denotes the pairwise probability distribution for symbols in a pair of sequence positions, and a and b represent the possible base or amino acid symbols of the sequence. P_a is the single site probability distribution for the first member of the pair, and P_b is the single site probability distribution for the second member of the pair. This expression may be interpreted as the log-likelihood ratio for the data for a specific pair of positions to arise from the independent (factorized) distribution versus a pairwise distribution.

To apply this formula one needs to estimate from the data the individual pairwise and single site probability distributions. Given a set of sequences which are assumed to be *i.i.d* (independent and identically distributed) samples from a probability distribution, then one can independently estimate each pairwise probability distribution for every pair of positions by frequency counting – this estimate results from a maximum likelihood analysis independently applied to each pair of positions. Marginalizing the estimate of the pairwise distribution yields the estimate for single site probabilities.

In Section 3 we will examine the effect on estimates of mutual information when the sequences used to estimate each individual pairwise probability distribution are not themselves independent samples, but are instead related via shared ancestry described by a phylogenetic tree. Other work addressing phylogenetic effects may be found in references [Altschul et al. (1989), Sibbald & Argos (1990), Gerstein et al. (1994), Hennikoff & Hennikoff (1994), Thompson et al. (1994)].

Our approach is to define a null model, which is based on evolution down an assumed known phylogenetic tree with independent mutations in different sequence positions. By incorporating the tree into the hypothesis of independent evolution of sites, we can determine a threshold value of mutual information from the null model, such that any values of mutual information seen in the real data which are over threshold have a very low probability of coming from the null model. In other words, a threshold is determined such that pairs of sites yielding mutual information values over threshold probably did not result

from independent evolution down the phylogenetic tree, i.e. they really are correlated. This approach simultaneously deals with two issues, (1) since mutual information is a positive semi-definite quantity, any estimate from finite data can only overestimate the mutual information. Put differently, positive values of mutual information will result even if sites are independent, purely due to fluctuations inherent in a finite sample size, and (2) non-trivial phylogenetic trees amplify the finite sample size effect, hence independent evolution of sites down a non-trivial phylogenetic tree will result in higher mutual information values than evolution down a simple star phylogeny. The null model technique addresses both these issues.

After addressing finite sample and phylogenetic effects, another important effect remains. In Section 4 we address this problem. The problem can be stated in various ways. One statement is that there can exist “chains” of covarying pairs of positions. For example, sequence position 3 may be correlated with position 23 (because these positions are spatially close in the folded structure), position 23 may be correlated with position 33 (because these positions are close in the folded structure), and position 33 may be correlated with position 43 (because these positions are close in the folded structure). Sequence position 3 would then typically be correlated with sequence position 43 due to the chaining of correlations between the two positions. However, sequence position 3 and sequence position 43 need not be spatially close in the folded structure, and an inference that they were close based on significant covariation between the positions can be in error. This “chaining effect” is the cause of many of the errors that occur when attempting to deduce spatially close positions in protein sequences using a covariation analysis. This effect, and the associated errors, is not as pronounced for RNA sequences because the specific bonding and saturation of Watson-Crick pairs tends to prevent chains of correlated mutations. We present a solution to the chaining problem for protein sequences which is validated in model simulations.

Physicists will recognize the “chaining effect” as “correlation at a distance” in spin systems [Stanley (1971), Binney et al (1992)], of which the one dimensional Ising spin model in a heat bath is a favorite example. The Ising model is a one dimensional chain of two-state spins, with each spin having a local physical interaction with only the spins on either side of it. Nevertheless, significant non-local correlations occur between spins that are separated by large distances even though the physical interaction is strictly a local nearest-neighbor interaction. More generally, one might have a spin system with physical interactions between designated sites described by a “contact matrix”, C_{ij} . If the spins have a direct physical contact, and hence a direct interaction, then $C_{ij} = 1$, and otherwise $C_{ij} = 0$. A potential matrix, P , describes the energetic contribution of two con-

tacting spins at i and one at j . Typically, this matrix is not position dependent, i.e. two “up” spins always have the same energetic contribution no matter where they are located (and similarly “down” spins, or mixed “up/down” spins). In proteins this matrix will be a twenty by twenty symmetric matrix describing the energetic contributions of two amino acids in contact. The probability of a configuration of spins (or amino acids), as represented by the Gibbs distribution, is proportional to $\exp -(\text{Energy})$, where *Energy* is the energy of the configuration (obtained from P and C_{ij}). In Section 4, we address the question: how can one use single site and pairwise probability information (as embodied in e.g. correlation measures) to estimate the contact matrix of local physical interaction?

Statisticians will recognize this question as being related to the inference of parameters, i.e. P and C_{ij} , occurring in the discrete multivariate probability distribution representing the probability of the sequence as a whole, given just estimates of the first and second order moments of the distribution. Clearly, under the special assumption that each site evolves independently of other sites then it is easy to estimate the probability distribution for the sequence as a whole using maximum likelihood techniques. However, this assumption utilizes only the first order moments, and ignores the second order moments. To also include the second order moments (as embodied in the observed correlations) we develop a maximum entropy analysis in Section 4. This analysis determines the unique probability distribution for the sequence as a whole, which has the given first and second order moments (i.e. correlations) and also has maximal entropy.

The problem of determining a probability distribution given just a finite number of moments is ill-posed – there are many solutions. The additional constraint of maximal entropy makes the solution unique. The maximal entropy constraint may be viewed as the plausible restriction that the distribution results in the observed correlations, but is otherwise as “flat”, or as “simple”, as possible. It is this simplicity constraint which limits the number of non-zero coefficients, and allows one to deduce a small set of local interaction parameters which can account for nonlocal correlations induced by the “chaining effect”, as we demonstrate in model simulations.

2. General models of evolution.

2.1. *Evolution with independent sites.* Models describing independent evolution of bases, such as the Jukes-Cantor [Jukes & Cantor (1969)] model and its variants [Kimura (1980)], can be extended to describe the independent evolution of amino acids [Kishino et al. (1990), Hasegawa & Fujiwara (1993)] by

incorporating PAM matrices [Dayhoff et al. (1978)] in definition of the transition rates. Let $P(t)$ be the probability distribution for the amino acids at a site evolving according to a Kimura style model of independent evolution of amino acid sites [Kimura (1980)]. To fix ideas, consider the simplest situation where $P(t)$ is a 20-vector satisfying the following simple equation (compare Hillis et al. (1995)):

$$dP(t)/dt = \begin{pmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,20} \\ a_{2,1} & a_{2,2} & \dots & a_{2,20} \\ \vdots & \vdots & \ddots & \vdots \\ a_{20,1} & a_{20,2} & \dots & a_{20,20} \end{pmatrix} P(t)$$

where

$$a_{ii} = (-19\alpha) \quad \text{and} \quad a_{ij} = (\alpha), i \neq j$$

The probability distribution for the sequence as a whole is the product of the independent single site probabilities, above. Kishino et al. (1990) and Hasegawa & Fujiwara (1993) extend the above model to incorporate the propensities of amino acids to mutate to those of a similar physico-chemical nature [Dayhoff et al. (1978)] and to have different equilibrium probabilities. The analytical solution to equations such as the above results in the familiar exponential time dependence of the probabilities, which can take a quite complicated form even though the equation satisfied by the probabilities is simple. In practice, sequences can be evolved numerically via Monte Carlo, not analytically, where the probability to accept a mutation is related to the values a_{ij} . The Monte Carlo procedure allows one to numerically simulate solutions to the general evolution equation, including situations where complicated interactions are introduced between sites (see below), and no analytic solution is possible.

2.2. Evolution with interacting sites. The Chapman-Kolmogorov equation for jump processes (the “Master Equation” to physicists) generalizes the above simple stochastic evolution equation to nonindependent i.e. interacting sites. It subsumes all possible discrete state evolution models. Let x_i be 20-state objects representing amino acids located at sequence position i . The Master Equation balances the probability of transitioning into a configuration, x , of the system from a configuration y , with the probability of transitioning out of configuration x of the system to all possible configurations y .

$$\frac{dP(x_1, x_2, \dots, x_n)}{dt} = \sum_{y_1 \dots y_n} \omega(x_1 \dots x_n; y_1 \dots y_n) P(y_1 \dots y_n) - \sum_{y_1 \dots y_n} \omega(y_1 \dots y_n; x_1 \dots x_n) P(x_1 \dots x_n)$$

Here, $\omega(x_1 \dots x_n; y_1 \dots y_n)$ is the instantaneous transition rate from configuration $y_1 \dots y_n$ to configuration $x_1 \dots x_n$. Only one mutation occurs in any infinitesimal time interval and hence the configuration x differs from the configuration y at only one site. The $\sum_{y_1 \dots y_n}$ are sums over all configurations y differing from configuration x in (all) single positions. This of course does not mean that the sites are evolving independently. The above single site, independent Kimura model, or any other discrete state standard evolution model which assumes independence, is recovered when the transition rates $\omega(x_1 \dots x_n; y_1 \dots y_n)$ depend only on single sites or are constants. The Master Equation is the most general expression possible for discrete state evolution models, and it encompasses non-independent evolution assumptions by choice of a suitable $\omega(x_1 \dots x_n; y_1 \dots y_n)$, as we illustrate below. If sites do not evolve independently (where analytic solutions are possible), then a numerical solution of the Master Equation via Monte Carlo is possible – a method of solution familiar to physicists in Monte Carlo analysis of interacting spin systems.

2.3. Defining the transition rates for interacting sites. Assume that an “interaction energy” function, $E()$, exists which defines the energy of a sequence based on pairwise interactions. To motivate the concept of such an interaction energy for protein sequences one may think of the classic pairwise “contact potentials” used in threading and inverse folding investigations [Sippl (1990), Sippl (1993)] however the arguments given below are independent of the exact form of the potential. Such pairwise “contact potentials” are of proven utility in relating sequence to structure. A pairwise potential based on an assumed energy of interaction provides the simplest possible model of evolution with interacting sites and thus provides the simplest possible generalization beyond the standard assumption of independent evolution of sites. Transition rates are related to the energy, $E()$, of configurations by a standard argument from statistical mechanics [Stanley (1971), Glauber (1963)] which we won’t repeat in detail here. We remark that it is clear that such a relation should exist from the following two observations:

- In equilibrium, where $\frac{dP(x_1, x_2, \dots, x_n)}{dt} = 0$, each configuration will occur with the Boltzman probability $\propto \exp(-E)$.
- In equilibrium, where $\frac{dP(x_1, x_2, \dots, x_n)}{dt} = 0$, the Master Equation yields a relation between the transition rate ω and the equilibrium probability, which involves E .

The energy defined by *contact potentials* used in inverse folding/threading problems (incorporating simple physico-chemical characteristics of pairwise amino acids interactions) motivates the form of E we will explore below, however

the formalism developed here is not limited to such potentials. In analogy to pairwise contact potentials, we define a model energy as: $E = \sum_{ij} P(A_i^\alpha, A_j^\beta) C_{ij}$ where $P(A_i^\alpha, A_j^\beta)$ is a fixed potential matrix defining the interaction energy of amino acid α at site i and amino acid β at site j . In inverse folding/threading investigations the $20 * 20$ symmetric matrix, P , is derived from the statistics of contacting amino acids observed in x-ray crystal structure data [Sippl (1990)]. In our model simulations, below, this $20 * 20$ matrix is chosen to have random elements between -1 and 1. C_{ij} is a “contact matrix” describing the structure of a “protein”, with element $C_{ij} = 1$ if the amino acid at site i is in contact with the amino acid at site j , and zero otherwise. In inverse folding/threading investigations “contact” is typically defined by a condition such as: the distance between the C^α atoms of residue i and residue j is less than 8 Angstroms. In our simulation C_{ij} was chosen to be a random, symmetric matrix of zeros and ones, with the average number of “contacts” per “amino acid” user specifiable. Contact potentials derived from inverse folding studies, and contact matrices derived from x-ray crystal structures of real proteins will be investigated in later work.

3. Phylogenetic effects. Sequences related by a phylogenetic tree do not constitute *i.i.d* samples. Hence estimation of pairwise probabilities by a frequency counting approximation, resulting from a maximum likelihood analysis which (falsely) assumes independence of the sequence samples, can be biased. Note that there are two uses of the concept of “independence” in this paper: (1) the assumption that the individual biological sequences are *i.i.d*, and (2) the assumption that individual positions in the sequences evolve independently, i.e. with no interaction between the positions. Of course, these two uses are quite different, and should not be confused.

In this section we present a null-model approach to handle phylogenetic bias in estimation of covariation and validate it in simulation. Given a phylogenetic tree, and a model for independent evolution of sites to be described below, we evolve sequences down the given tree numerous times using the independence model for sequence evolution of Section 2. A histogram is compiled for the resulting mutual information values which are calculated between all pairs of sequence positions. These mutual information values will be different from zero, even though the sites are evolving independently, due to (a) finite sample size effects (the mutual information is a positive semi-definite quantity and any fluctuation due to finite sample size can therefore only result in positive mutual information) and (b) effects of the phylogenetic tree (the bifurcations of a typical phylogenetic tree tend to amplify finite sample fluctuations). At a bifurcation point of the tree the state of the sequence is duplicated, and the two copies are

subsequently independently evolved.

The null model procedure described above determines a threshold mutual information value, such that if any mutual information value calculated for the real sequence data exceeds the threshold value, then it is very unlikely that such a value could have arisen from the null model of "given phylogenetic tree and independent evolution of sites". However, the conclusion that the mutual information between a pair of sites was unlikely to have arisen from the null model of independence does not necessarily mean those sites are directly physically interacting. A second procedure is needed which is able to disentangle

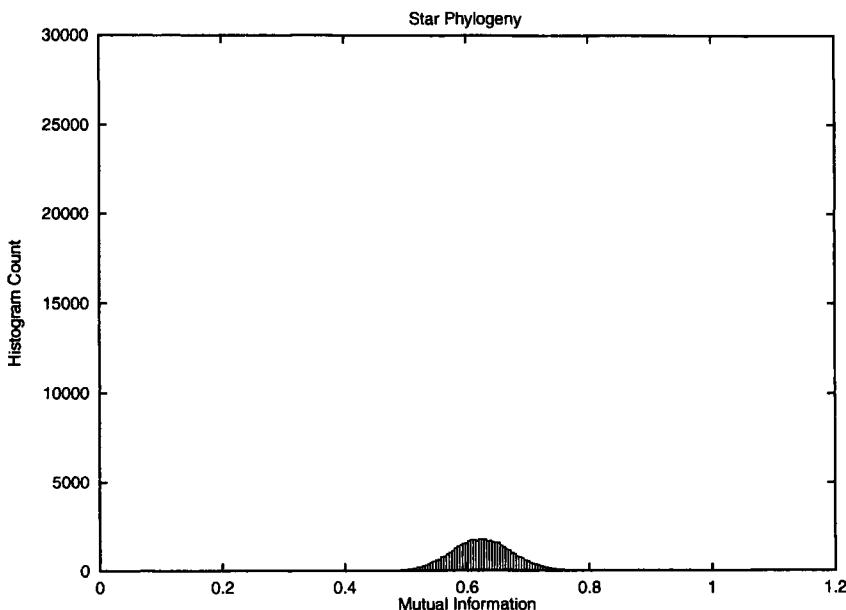


FIG. 1. *Histogram of mutual information between all pairs of sites in 256 "amino acid" sequences at the leaves of a star phylogeny. The sequences are 100 "amino acids" long. The total branch length from the root to the leaves is 64 time units. Ten separate runs with different random root sequences have been averaged to create the histogram.*

long chains of correlation to determine which sites are correlated due to direct interaction, and which sites are correlated due to (possibly long) indirect chains of interaction. This procedure is introduced in Section 4.

3.1. *The null model: Independent evolution of sites down a given phylogenetic tree.* Consider a model simulation in which 100 amino acid long sequences are evolved using a Kimura style independent site evolution model [Kimura (1980), Kishino et al. (1990)] (see Section 2), with mutations occurring independently in different positions. Any amino acid can mutate with equal probability to

any other amino acid. We show that non-trivial phylogenetic trees "create" mutual information between sites, even when explicit interactions between sites is absent. This is because the topology of the tree magnifies the effects of finite sample size on estimation of mutual information. In the simulation we will consider a binary "tree" of 8 levels with each branch length 8 time units long,

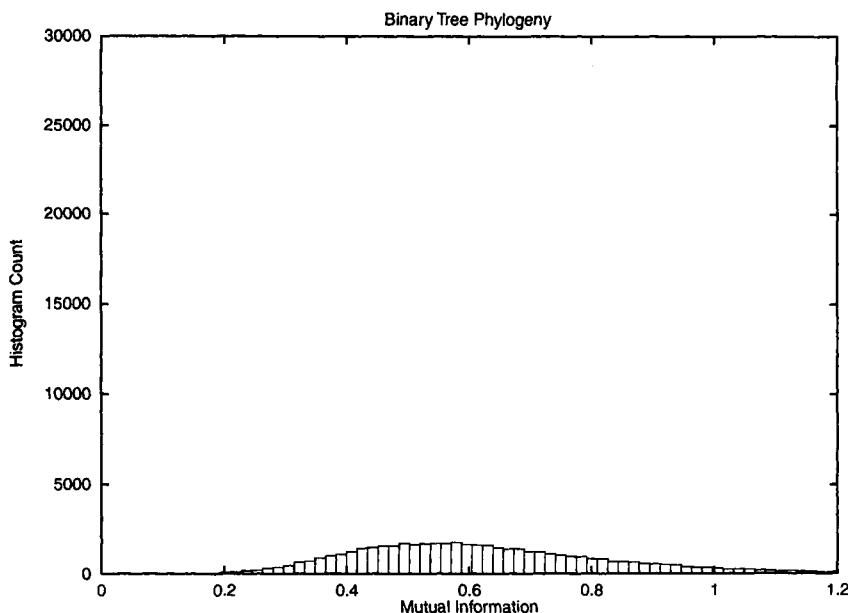


FIG. 2. *Histogram of mutual information between all pairs of sites in 256 "amino acid" sequences at the leaves of a tree phylogeny which is a balanced binary tree of eight levels. The sequences are 100 "amino acids" long. The individual branch lengths are of length eight, resulting in a total branch length from root to leaves which is identical to the star phylogeny of Figure 1 i.e. 64 time units. Ten separate runs with different random root sequences have been averaged to create the histogram.*

resulting in 256 different "amino acid" sequences of length 100 at the leaves. The total time from root to leaves is 64 time units. We use a binary tree merely as a familiar example of a tree with non-trivial topology which can be compared to a star phylogeny having a trivial topology. Phylogenetic trees of actual biological sequences will generally not be balanced binary trees and details will differ depending on the detailed tree topology. We will calculate the mutual information between all pairs of positions of the sequences at the nodes of the tree, and compare this calculation to that of a star phylogeny with 256 children and total branch length equal to that of the binary tree above, i.e. 64 time units. To evolve a sequence down a given phylogenetic tree the state of the sequence is duplicated at each bifurcation point of the tree, and the two copies are stochastically and independently evolved from the common ancestor.

The histogram of mutual information for the star phylogeny and the tree phylogeny are shown in Figures 1 and 2 respectively. It may be observed that there is a non-zero probability of achieving higher mutual information values (even though all sites are evolving independently) in the tree phylogeny, as opposed to the star phylogeny. A null model threshold based on the star phylogeny, i.e. based on an incorrect threshold which results from ignoring the real phylogenetic tree, would be too low and result in false conclusions of non-independence. On the other hand, if a null model threshold is chosen based on the correct phylogenetic tree, then sites which were truly independent, but evolved down the given tree and hence associated with an amplification of finite sample size effects, will be detected as being independent. This is quantified in the following section, where we introduce a specific interaction between sites, and show by explicit simulation that the specificity of predicting non-independent sites by evaluation of mutual information is increased when knowledge of the correct phylogenetic tree is used to create the null model.

3.2. Validating the null model. Various attempts to “weight” sequences in a manner related to the tree to correct for bias exist in the literature [Altschul et al. (1989), Sibbald & Argos (1990), Gerstein et al. (1994), Hennikoff & Hennikoff (1994), Thompson et al. (1994)]. However, to our knowledge such approaches have not been validated in model simulations where the interaction between designated sites is under the investigator’s control. Here, we validate the null-model approach described above, in a model world where we can test the ability to predict interacting sites based on observed correlations. To create the model world:

- (a) select a phylogenetic tree, here a balanced binary tree of eight levels with 256 leaves, such as used in Figure 2. Various branch lengths of the tree will be considered in separate runs, ranging from extremely short lengths, to lengths that are sufficiently long to have sequences evolve to equilibrium.
- (b) evolve sequences via Monte Carlo using a *non-independent* model (see Section 2) with a selected C_{ij} and potential matrix P to generate sequences which play the role of “sequences observed in Nature”, and which have sites that are truly interacting. The connection matrix for results reported here has every “amino acid” contacting three other amino acids chosen at random. The potential matrix P for results reported here was chosen to be a symmetric, twenty by twenty matrix, with elements chosen at random from a flat distribution between negative one and one.

- (c) Calculate the mutual information between pairs of sites in these “real” sequences.

Next, we create the null model of “a given tree and independent evolution of

sites" by evolving sequences via Monte Carlo down the selected tree (always assumed known to the investigator), but using the independent model of evolution where each "amino acid" has equal probability to mutate to any other "amino acid". We determine the threshold of the null model to be such that mutual

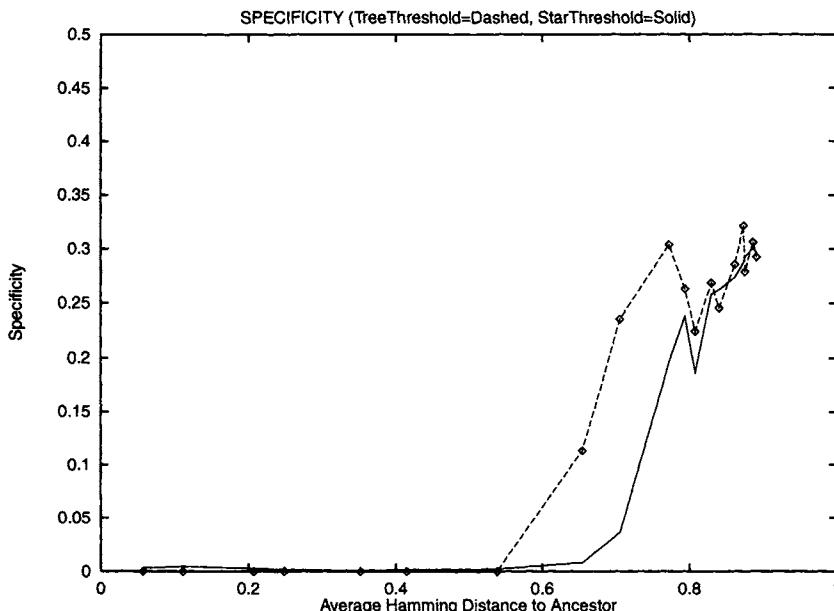


FIG. 3. *Specificity of the prediction of contacts based on over-threshold values of the mutual information is plotted as a function of the average Hamming distance of the 256 children to the root. Length 100 "amino acid" sequences were evolved down a phylogenetic tree with a binary tree topology of eight levels (256 child sequences). Dashed curve: threshold determination of the null model was correctly based on the binary tree topology. Solid curve: threshold determination was incorrectly based on a star phylogeny. If the topology of the balanced binary phylogenetic tree is ignored and a star phylogeny is incorrectly assumed, then the specificity of predictions is seen to be significantly lower than that achieved using a null model which correctly incorporates the phylogenetic tree.*

information values which exceed the threshold are very unlikely to have been generated by the null model (i.e., the threshold is in the far tail of the histogram of mutual information values). Hence, mutual information values for pairs of positions as calculated in the "real" data which exceed threshold are very unlikely to have been generated by the null model of independence of mutations.

Mutual information values between pairs of positions i and j in the "sequences observed in Nature" which are over the null model threshold are predicted to have contact matrix element, $C_{ij} = 1$, i.e. are predicted to be spatially close. To verify the predictions, an "experiment" may be performed in the model world to determine the "real" values of C_{ij} . Of course, this experiment is as simple as viewing the file containing the original values chosen for C_{ij} in

step (b) above, which was used in the Monte Carlo evolution that generated the "sequences from Nature".

This procedure results in "specificity" and "sensitivity" plots, Figures 3 and 4 for the prediction of $C_{ij} = 1$. Specificity is defined to be the percentage of predicted contacts that were actual contacts, i.e. that were defined in step (b) above to have $C_{ij} = 1$. Sensitivity is defined to be the percentage of actual contacts that were predicted to be contacts. Figures 3 and 4 show the result of numerous runs with varying branch lengths ranging from short (one time unit) to long (five hundred time units). The two separate extremes of very short branch lengths where essentially no evolutionary mutation takes place, and very long branch lengths where equilibrium can be reached within one branch of the tree, show little difference as should be expected. For intermediate branch lengths the effects of the phylogenetic tree become evident.

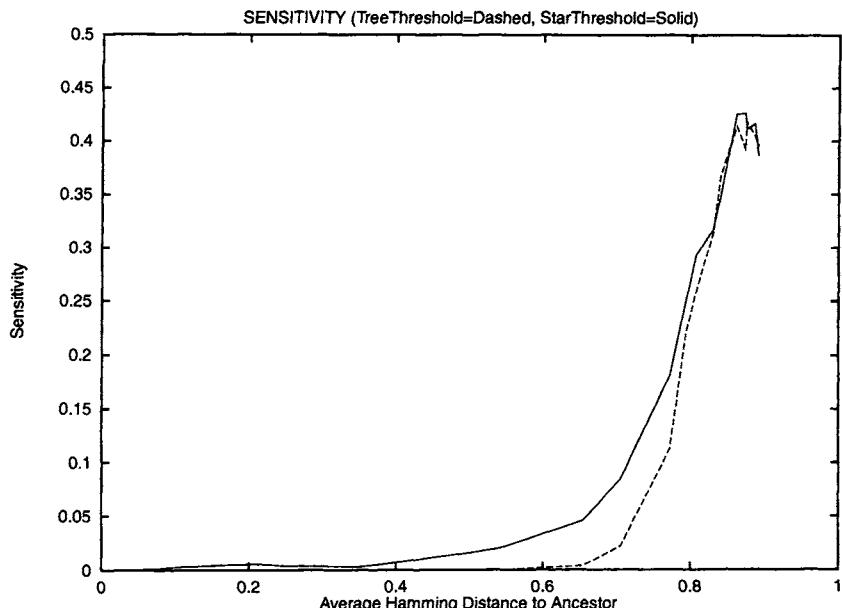


FIG. 4. *Sensitivity of the prediction of contacts based on over-threshold values of the mutual information is plotted as a function of the average Hamming distance of the 256 children to the root. Length 100 "amino acid" sequences were evolved down a phylogenetic tree with a star topology of 256 child sequences and total branch length equal to that of the binary tree phylogeny. Dashed curve: threshold determination of the null model was correctly based on the binary tree topology. Solid curve: threshold determination was incorrectly based on a star phylogeny. Choosing a threshold based on a null model correctly incorporating the binary phylogenetic tree, as opposed to incorrectly assuming a star phylogeny, enhances specificity at the expense of sensitivity.*

Note that choosing the threshold of the null model by using the topology of the given phylogenetic tree does significantly improve specificity, compared to a

null model threshold determined by ignoring tree topology and naively using a star phylogeny. However a significant number of errors clearly still remain. These are addressed in the following section. The sensitivity for predicting correct contacts is decreased, see Figure 4, if the correct phylogenetic tree is incorporated into the construction of the null model. A decrease in sensitivity with enhanced specificity, is preferable to enhanced sensitivity at the expense of specificity. In Section 4 we introduce a new methodology to use the observed first and second order moments to predict physical contacts which is not based on simple thresholding of mutual information or correlation measures.

4. Structural effects. The origin of the specificity errors in the simulation investigated in Section 3 are due to a covariation versus causation phenomenon. Consider the following situation: Site A physically interacts and covaries with site B; site B physically interacts and covaries with site C; but site A and site C do not physically interact. Site A can covary with site C in spite of no physical interaction between A and C. This effect of *chained covariation* is known as “correlation at a distance” or “order at a distance” in the analysis of interacting spin systems [Stanley (1971), Binney et al (1992)]. How can one disentangle causation (direct physical interaction) from chained covariation (order at a distance)? We present in the following a maximum entropy approach to this problem.

4.1. Maximum entropy analysis. Although protein sequences can be hundreds of amino acids long, typically a much smaller number of amino acids display significant covariation, perhaps on the order of ten to twenty amino acids depending on the length of sequence examined. In this section we will consider, for reasons of simplicity only ten potentially interacting sites, and we will also restrict consideration to two-state “amino acids”. The algorithms developed here scale reasonably with the number of potentially interacting amino acids, and with the number of states per “residue” (see below), but in general the algorithms require a non-trivial amount of computation time. Hence, for the following model simulations which we use to explain and to validate the algorithms, we report results for smaller systems where results can be obtained easily. However, the algorithms do have a practical scaling behavior and are applicable to larger systems.

Consider the simplest situation of evolution down a star phylogeny. Phylogenetic tree effects due to more complicated tree topology can also easily be accommodated. Consider a star phylogeny with five hundred leaves containing two-state sequences of length ten, obtained by evolution to near equilibrium (10000 time steps) using the following connectivity matrix where on average each site is connected to three other sites:

$$\begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ . & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ . & . & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ . & . & . & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ . & . & . & . & 0 & 0 & 1 & 0 & 1 & 1 \\ . & . & . & . & . & 0 & 0 & 1 & 0 & 1 \\ . & . & . & . & . & . & 0 & 1 & 0 & 0 \\ . & . & . & . & . & . & . & 0 & 0 & 0 \\ . & . & . & . & . & . & . & . & 0 & 0 \\ . & . & . & . & . & . & . & . & . & 0 \end{pmatrix}$$

The two by two interaction matrix, P , for a two-state system contains three independent components, which to use the language of spins, can be described as: up-up, up-down (and equivalently down-up), and down-down. A “ferromagnetic” interaction matrix, which we use for this example, has the up-up and down-down values assigned positive one, and the up-down (equivalently down-up) value assigned negative 1. Allowing more general potential matrices for two-state systems merely has the effect of adding new terms to the energy that are linear in the spins (“external magnetic fields” in spin language), which serve only to bias the final equilibrium single site probabilities and do not illuminate chaining phenomena.

The correlation matrix, $\langle (x_i - \bar{x}_i)(x_j - \bar{x}_j) \rangle$, and the contact matrix are represented graphically in Figure 5. Solid lines between pairs of sites represent those pairs which are connected via the contact matrix (above). Dashed lines between pairs of sites represent pairs which achieved a correlation (absolute value) over 0.3. This threshold value for representation was chosen because such values occurred less than one time in one hundred, as computed in one hundred simulations of evolution down the same star phylogeny, but using independent evolution of sites. Hence dashed lines between sites represent correlations that are very improbable to have occurred in the null model of independent evolution of sites, and yet are not caused by direct connections between sites.

Note that significant covariation exists between many pairs of sites that are not physically connected. Sites (1,6), as well as sites (2,8), see Figure 5 are examples. Note that sites (1,2) are physically connected, as are sites (2,6), and hence a chain of covariation, (1,2) (2,6), can form which leads to significant correlation between disconnected sites such as (1,6). In larger systems one can have extended chains. Correlation between sites (2,8) is also significant even though they are not physically connected. Although we prefer to call the general mechanism by which disconnected sites can co-vary, “chained correlation”, there is

also another interpretation. The disconnected sites (2,8) can display significant correlation because sites (1,2) are physically connected, as are sites (1,8). In this situation, correlation can exist between sites (2,8), even though they are

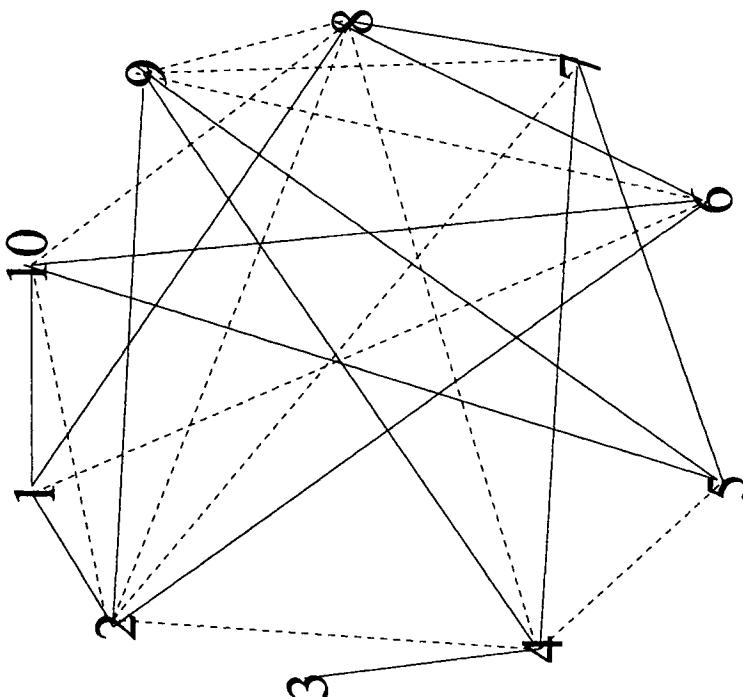


FIG. 5. *Correlations (dashed lines) between sites such as (1,6) and (2,8) occur because of the chaining effect. Sites (1,8), (1,2) and (2,6) have physical connections (solid lines). Sites (1,6) and (2,8) are not connected, yet still exhibit statistically significant correlation. Statistically significant values of correlation were computed by performing one hundred simulations of evolution down the same star phylogeny, but with independent evolution of sites. A correlation greater than 0.3 (absolute value) has less than a one in one hundred chance of occurring if sites are truly independent.*

not physically connected, because of a common “driving cause”, which is the connection of both site (2) and site (8) to site (1). Such chained correlation effects are examples of the classic conundrum of “covariation versus causation”.

It is clear that in situations where there can be extended interactions between sites, such as in proteins (but not so much in RNA, where once a base pair forms, it is unusual to form many other base pairs), then extended, complicated chains can occur which leads to correlations between sites that are not physically connected. Attempting to predict the connectivity matrix based on the correlation matrix would in general result in poor specificity. One must disentangle the chains of covariation and solve, in our specific case at least, the classic conundrum of causation versus covariation. As we show below, an effec-

tive solution is to estimate the parameters of the simplest model probability distribution which yields the observed correlations.

Classic Problem: Given estimates of first and second moments of a probability distribution (as used to estimate the correlation, above) determine the probability distribution which has maximal entropy, i.e. which is the “flattest” or “simplest” distribution satisfying the observational constraints. This problem would be ill-posed without the additional constraint of “simplicity” i.e. maximum entropy – many probability distributions exist which agree with any given moments.

Classic Solution: Maximizing the entropy subject to the constraints of given first and second moments results in the classic [Tikochinsky et al. (1984), Levine & Tribus (1979)] form for P :

$$P(x) = \frac{\exp - (\sum_i \lambda_i x_i + \sum_{ij} \lambda_{ij} x_i x_j)}{Z}$$

where the λ 's are Lagrange multiplier used to implement the constraints, and Z normalizes P to unity. The constraints are satisfied at the minimum of the following function F , considered to be a function of the λ 's:

$$F = \log Z + \sum_i \lambda_i \bar{x}_i + \sum_{ii} \lambda_{ii} \bar{x}_i x_j$$

where \bar{x}_i and $\bar{x_i x_j}$ represent the observed first and second order moments, respectively. For related investigations with a similar functional form see [Heumann et al. (1995)].

4.2. Application to model simulation. Comparing the resulting form of $P(x)$, above, to the contact potential used to generate the simulated data, we see that the reconstructed parameters λ_{ij} should be zero for non-connected sites, and equal to the appropriate element of the potential matrix, P , for connected sites. When applied to the above model simulation, the reconstructed parameters are:

Note that even though the correlation between non-connected sites, such as (1,6) and also (2,8) can be high, that the reconstructed parameter values above are generally low between non-connected sites, and are high (bold face values) between connected sites. The maximum entropy procedure identifies the dashed lines of Figure 5 as a chaining phenomenon. Finite sample effects accounts for the remaining “noise” in elements of the matrix which should have value zero (because of zero connectivity), and in the elements which should have absolute value of one (due to nonzero connectivity and the values used in the ferromagnetic potential matrix, P).

4.3. Practical considerations. F is a nonlinear function of the variables λ . It is possible to prove that F has a unique, global minimum by using standard inequalities of information theory [Cover & Thomas (1991)]. Evaluating the minimum of F by e.g. gradient descent with respect to the variables λ , results in an expression involving the first and second order moments of the model distribution evaluated at intermediate (i.e. non-extremal) values of λ . Thus, the numerical procedure to solve for the parameters λ involves successive rounds of Monte Carlo evolution, followed by a small change in the λ 's in the gradient direction, followed by Monte Carlo to evaluate the new expectations etc. It may be seen by direct differentiation of F with respect to the λ 's that this process converges when the numerically computed expectations agree with the specified expectations \bar{x}_i and $\bar{x}_i \bar{x}_j$. Evaluation of the first and second order moments at each intermediate value of the λ 's would be prohibitively expensive if all states were enumerated exhaustively. However, standard techniques of importance sampling, long familiar to physicists performing Monte Carlo simulation of spin systems [Binney et al (1992)], and now popular in bioinformatic investigations [Lawrence et al. (1993)], are an accurate and efficient alternative to exhaustive enumeration of all states.

4.4. Conceptual considerations. An assumed pairwise interaction energy may not accurately model the *fitness function* that is optimized in Nature. We emphasize that although we have motivated our discussion of correlated mutations using analogies to pairwise contact potentials (because we believe that some aspects of fitness are related to the match of sequence to structure as represented in pairwise contact potentials), the formalism is not limited to the protein contact potentials in use today. Indeed, the second order interactions we allow are perfectly general. In the maximum entropy formalism the second order interactions are determined by the observed correlations and as such provide the first logical step beyond independence of sites.

For practical reasons, it will be of interest (but is not necessary) to pursue the

utility of pairwise contact potentials in evolutionary analysis by using the form of pairwise potential matrices, P , to restrict the variability of the λ_{ij} parameters above. This can be done by assuming that the λ_{ij} parameters are the product of a fixed twenty by twenty amino acid interaction matrix (related to the potential of pairwise contact potential investigations), P , multiplied by a variable contact matrix, C_{ij} . The maximum entropy formalism performs an implicit search over C_{ij} , which even in the simple example considered here involves an implicit search over 2^{45} or approximately 10^{13} discrete contact matrices. This illustrates the power of the formalism. Other heuristic search techniques could also be used, such as genetic algorithms or use of Monte Carlo methods to search over the large space of discrete contact matrices, C_{ij} . We note that there is an assumption that the mutations do not change the C_{ij} , i.e. that the protein backbone remains relatively unchanged in spite of the amino acid mutations. Examples of this abound in Nature, including e.g. the hundreds of variable globin sequences which share a well conserved backbone structure.

Other issues deserving investigation include:

- the effects of mis-specification of the assumed model for the probability of a sequence: suppose that the “true” fitness function according to which real sequences are evolved in Nature includes, e.g., third order terms in addition to second order terms, and hence these terms will influence the observed values of second order correlations. If one observes just second order correlations then how accurately will the second order terms of the fitness function be recovered in a model which ignores third order terms? In other words, how “structurally stable” is the formalism?
- robustness of the reconstructed parameters to noise or sampling error in the original estimation of the moments from data: limited data will produce errors in the estimated moments. How stable are the reconstructed parameters to the presence of such errors?
- the assumption of evolution to equilibrium: if sequences in Nature are observed at times before equilibrium is reached, how will this affect the reconstructed parameters (obtained under an assumption of equilibrium)?

5. Conclusions. We have addressed two issues in covariation analysis of biosequences, (1) the effect of nontrivial phylogenetic trees on the estimation of mutual information, and (2) the effect of protein structure on propagation of correlations to sites that are not structurally linked. Regarding issue (1): Naive application of covariation analysis to biological sequences related by a phylogenetic tree can give misleading results. A non-trivial phylogenetic tree can amplify finite sample size fluctuations, making it appear that significant covariation exists between pairs of sites, when in fact all sites are evolving in-

dependently of each other. A null model procedure was introduced to address this problem. Regarding issue (2): Covariation between disconnected pairs of sites in sequences can result from possibly long chains of co-variation and from "common cause" effects, and not from causation (i.e. not from structural links). Chained covariation makes the prediction of structural links using naive application of covariation analysis prone to error. A technique involving maximum entropy reconstruction of the parameters for the probability distribution of the sequences was developed, and was validated in model simulations where accurate recovery of the structural links was achieved.

We remark that additional errors will probably remain even after addressing phylogenetic and chaining effects. The origins of such errors can be diverse, such as possibly critical relationships between certain amino acids that are required to maintain the folding pathway. However, addressing the phylogenetic and chaining effects should go a long way towards improving accuracy of prediction of spatial contacts in families of varying protein sequences.

The conclusion that causation (direct structural links) can be distinguished from covariation, by fitting parameters to an assumed model, stands independent of the particular models and simulations used here to illustrate the point. Our goal in this paper is to lay the conceptual foundation for protein structure determination via analysis of covarying mutations, by using models describing the probability distribution of the sequences as a whole, and by constraining the probability distributions with simplicity criteria such as maximum entropy.

Acknowledgments. The authors would like to thank the Santa Fe Institute, where part of this work was performed.

REFERENCES

- ALTSCHUL, S., CARROLL, R. and LIPMAN, D. (1989). Weights for data related by a tree. *Journal of Molecular Biology* **207** 647–653.
- BINNEY, J., DOWRICK, N., FISHER, A. and NEWMAN, M. (1992). *The Theory of Critical Phenomena*. Oxford University Press, Oxford.
- COVER, T. and THOMAS, J. (1991). *Elements of Information Theory*. Wiley Series in Telecommunications, New York.
- CLARKE, N. (1995). Covariation of residues in the homeodomain sequence family. *Protein Science* **4** 2269–2278.
- DAYHOFF, M., SCHWARTZ R. and ORCUTT, B. (1978). A Model of Evolutionary Change in Proteins, pps. 345–352 in *Atlas of Protein Sequence and Structure*, vol. 5, suppl. 3, Natl. Biomedical Res. Found., Silver Spring, Maryland.
- GERSTEIN, M., SONNHAMMER, E. and CHOTHIA, C. (1994). Volume changes in protein evolution. *Journal of Molecular Biology* **235** 1067–1078.
- GLAUBER, R. (1963). Time-dependent statistics of the Ising Model. *Journal of Mathematical Physics* **4** 294.
- GOBEL, U., SANDER, C., SCHNEIDER, R. and VALENCIA, A. (1994). Correlated mutations and residue contacts in proteins. *Proteins: Structure, Function, Genetics* **18** 309–317.

- GRIBSKOV, M., McLACHLAN, A. and EISENBERG, D. (1987). Profile analysis: Detection of distantly related proteins. *Proceedings of the National Academy of Sciences of the USA* **84** 4355–4358.
- GUTELL, R.R., POWER, A., HERTZ, G.Z., PUTZ, E. and STORMO, G.D. (1992). Identifying constraints on the higher-order structure of RNA: continued development and application of comparative sequence analysis methods. *Nucleic Acids Research* **20** 5785–5795.
- HASEGAWA, M. and FUJIWARA, M. (1993). Relative efficiencies of the maximum likelihood, maximum parsimony, and neighbor joining methods for estimating protein phylogeny. *Molecular Phylogenetics and Evolution* **2** 1–5.
- HENNIKOFF, S. and HENNIKOFF, J. (1994). Position-based sequence weights. *Journal of Molecular Biology* **243** 574–578.
- HEUMANN, J., LAPEDES, A. and STORMO, G. (1995). Alignment of regulatory sites using neural networks to maximize specificity. In: *Proceedings of the 1995 World Congress on Neural Networks II*, 771–775.
- HILLIS, D., MORITZ, C. and MABLE, B. (1995). *Molecular Systematics* (second edition). Sinauer Associates Inc., Sunderland, MA.
- JUKES, T. and CANTOR, C. (1969). *Evolution of protein molecules*. In Munro, H. (ed.) *Mammalian Protein Metabolism*, Academic Press, New York.
- KIMURA, M. (1980). A simple method of estimating evolutionary rate of base substitutions through comparative analysis of nucleotide sequences. *Journal of Molecular Evolution* **16** 111–120.
- KIMURA, M. (1983). *The Neutral Theory of Evolution*. Cambridge University Press, Cambridge, England.
- KISHINO, H., MIYATA, T. and HASEGAWA, M. (1990). Maximum likelihood inference of protein phylogenies and the origin of chloroplasts. *Journal of Molecular Evolution* **31** 151–160.
- KORBER, B., FARBER, R., WOLPERT, D. and LAPEDES, A. (1993). Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: An information theoretic analysis. *Proceedings of the National Academy of Sciences of the USA* **90** 7176–7180.
- KROGH, A., BROWN, M., MIAN, I., SJOLANDER, K. and HAUSSLER, D. (1994). Hidden Markov models in computational biology: Applications to protein modeling. *Journal of Molecular Biology* **235** 1501–1531.
- LAWRENCE, C., ALTSCHUL, S., BOGUSKI, M., LIU, J., NUEWALD, A. and WOOTON, J. (1993). Detecting subtle sequence signals: A Gibbs sampling strategy for multiple alignment. *Science* **262** 208–214.
- LEVINE, R. and TRIBUS, M. (1979). The maximum entropy formalism. *Physical Review* **106** 620.
- NEHER, E. (1994). How frequent are correlated changes in families of protein sequences? *Proceedings of the National Academy of Sciences of the USA* **91** 98–102.
- SHINDYALOV, I., KOLCHANOV, N. and SANDER, C. (1994). Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? *Protein Engineering* **7** 349–358.
- SIBBALD, P. and ARGOS, P. (1990). Weighting aligned protein or nucleic acid sequences to correct for unequal representation. *Journal of Molecular Biology* **216** 813–818.
- M. J. SIPPL (1990). Calculation of conformational ensembles from potentials of mean Force – An approach to the knowledge-based prediction of local structures in globular proteins. *Journal of Molecular Biology* **213** 859–883.
- M. J. SIPPL (1993). Boltzmann's principle, knowledge-based mean fields and protein folding. *Journal of Computer-Aided Molecular Design* **7** 473–501.
- STANLEY, H. (1971). *Introduction to Phase Transitions and Critical Phenomena*. The International Series of Monographs on Physics, Oxford University Press Inc., Oxford and New York.
- TAYLOR, W. and HATRICK, K. (1994). Compensating changes in protein multiple sequence alignments. *Protein Engineering* **7** 341–348.
- THOMAS, D., CASARI, G. and SANDER C. (1996). The prediction of protein contacts from multiple sequence alignments. *Protein Engineering* **9** 941–948.
- THOMPSON, J., HIGGINS, D. and GIBSON, T. (1994). Improved sensitivity of profile searches through the use of sequence weights and gap excision. *CABIOS* **10** 19–29.

TIKHOCHINSKY, N., TISHBY, N. and LEVINE, R. (1984). Alternate approach to maximum entropy inference. *Physical Review A* **30** 2638-2644.

THEORETICAL DIVISION
LOS ALAMOS NATIONAL LABORATORY
LOS ALAMOS, NM, 87545
ASL@LANL.GOV
LIU@LANL.GOV

SERVICE PHYSIQUE THÉORIQUE
DSM, C.E.N. SACLAY
Gif/YVETTE, FRANCE 91191
GIRAUD@SPHT.SACLAY.CEA.FR

SANTA FE INSTITUTE
1399 HYDE PARK ROAD
SANTA FE, NM 87501
ASL@SANTAFE.EDU

MOLECULAR, CELLULAR AND DEVELOPMENTAL BIOLOGY
UNIVERSITY OF COLORADO
BOULDER, COLORADO
STORMO@BEAGLE.COLORADO.EDU

LARGE COMPOUND POISSON APPROXIMATIONS FOR OCCURRENCES OF MULTIPLE WORDS

BY GESINE REINERT¹ AND SOPHIE SCHBATH²

Department of Statistics, UCLA and Unité de Biométrie, INRA, 78352 Jouy-en-Josas, France.

A compound Poisson process approximation for the number of occurrences of multiple words in a sequence of letters is derived, where the letters are assumed to be independent and identically distributed. Using the Chen-Stein method, a bound on the error in the approximation is provided. For rare words, this error tends to zero as the length of the sequence increases to infinity. As an application the efficiency of the approximation for the number of occurrences of rare stem-loop motifs in DNA sequences is illustrated.

1. Introduction. When searching a database for the occurrence of a combination of several words within a sequence, the typical Poisson approximation used by programs like BLAST is no longer valid, as overlapping words may be dependent on each other. Here a compound Poisson approximation for the multiple occurrences of short words within a sequence is derived. Using the Chen-Stein method for Poisson process approximation, an explicit error bound for the approximation is given, improving those obtained by Schbath (1995a) for a single rare word. The approximation error increases with the amount of overlap between the words. The results are applied to the occurrences of stem-loop motifs. Another application might be a set of words coding for the same amino-acid sequence.

In general, consider a finite sequence S of letters chosen independently from a finite alphabet \mathcal{A} . The main example will be the four-letter DNA alphabet $\{\text{A, C, G, T}\}$ but the results are valid for general finite alphabets such as $\{0, 1\}$ or the 20-letter amino acid alphabet. An abundant literature exists on the asymptotic distribution of the number of occurrences of a single word in such a sequence S . A normal approximation, valid for frequent words, is presented by Prum et al. (1995). A compound Poisson approximation is obtained in Arratia et al. (1990), Geske et al. (1995) and Schbath (1995a) for the number of occurrences of a rare word, whereas the number of clumps of a rare word is approximated by a Poisson variable (as a rule of thumb, a word is rare if its length

¹This work was partially supported by NSF grant DMS-9505075.

²This work was partially supported by NATO and by NSF grant BIR-9504393.

AMS 1991 subject classifications. Primary 60F05 ; secondary 92D20.

Key words and phrases. Chen-Stein method, stem-loop motifs, compound Poisson approximation, occurrences of multiple words.

is at least of order $\log n$, where n is the length of the sequence). As soon as one is simultaneously interested in occurrences of different rare words in a sequence, the asymptotic joint distribution of the different counts is of interest; the novelty in this paper is to provide multidimensional results and to give conditions for asymptotic independence. The multidimensional approximation and in particular the asymptotic statistical independence for counts of multiple words is very useful to study statistical properties of any function of these counts. (Note that asymptotic statistical independence does not necessarily have a biological interpretation.)

Instead of using the Chen-Stein method for Poisson process approximation as stated in Arratia et al. (1990) and refined in Barbour et al. (1992b), a more direct approach could have been the Chen-Stein method for compound Poisson approximation, developed by Barbour et al. (1992a), Roos (1994) and Barbour and Utev (1997), which has been applied in this context to approximate the count of single words with simple self-overlapping structure in a two-letter alphabet [Roos and Stark (1996)], but it is not adapted for a multidimensional approximation to multiple words.

For non-rare words (short words related to the length of the sequence), a Gaussian approximation is more appropriate and corresponding results have been shown: Lundstrom (1990) was the first to derive a multidimensional Gaussian approximation (using the δ -method) for a m -tuple of counts; see Waterman (1995, Chapter 12) for an exposition. Prum et al. (1995) or Schbath et al. (1995) give an explicit formula for the asymptotic covariance matrix. These results can be used to construct a Gaussian statistic based on the count of a word family. Recently, Tanushev (1996) proved the multidimensional Gaussian approximation for an m -tuple of renewal counts.

The general case where the letters are modeled using a stationary Markov chain is treated in Reinert and Schbath (1998); Reinert and Schbath (1998) also give a Poisson process approximation. The purpose of this paper is to treat the independent case only, meaning that the letters are assumed to be independent and identically distributed, as under this additional assumption the arguments and bounds in Reinert and Schbath (1998) simplify considerably.

To approximate the counts of occurrences of words, the “declumping” approach is used – first the number of clumps of occurrences is counted, and then the sizes of the clumps are determined. In Section 2, an occurrence of a word and an occurrence of a clump in a sequence are defined, as well as the number of occurrences of a word and the number of clumps of a word in a sequence. Moreover the decomposition of the count of a word with respect to the number of clumps is introduced; this decomposition is fundamental to proving the compound Poisson approximation via the Chen-Stein method given in Arratia

et al. (1990). Next, in Section 3 the compound Poisson process approximations for counts of m words with not necessarily identical lengths are presented. As an illustration (Section 4), the count of some stem-loop motifs are studied, like ATGGC_NNNNGCCAT (\mathbf{N} denotes any letter in the four-letter DNA alphabet), in a model for the λ -phage genome. We give the expected counts, the error bounds and the asymptotic distributions. As we will see, the error bounds are very small, and thus the method provides a useful tool to approximate a collection of counts.

2. Preliminary notation and the Chen-Stein method. Consider a sequence of i.i.d. letters $\mathcal{M} = \{X_i\}_{i \in \mathbb{Z}}$ on a finite alphabet \mathcal{A} , where the letters $\{X_i\}_{i \in \mathbb{Z}}$ are chosen independently with probabilities $\mathbb{P}(X_i = x) = \mu(x)$, $x \in \mathcal{A}$; assume that $\mu(x) > 0 \forall x \in \mathcal{A}$. Let $\underline{u} = u_1 u_2 \cdots u_\ell$ be a word of length ℓ on \mathcal{A} . Say that an occurrence of \underline{u} starts at position i in the infinite sequence \mathcal{M} if $X_i X_{i+1} \cdots X_{i+\ell-1} = u_1 u_2 \cdots u_\ell$, and denote the indicator random variable of this event by $\mathbb{I}_i(\underline{u})$. The probability $\mu(\underline{u})$ that \underline{u} starts at a given position in \mathcal{M} is exactly the expectation of $\mathbb{I}_i(\underline{u})$ and is given by

$$\mu(\underline{u}) := \mathbb{E}\mathbb{I}_i(\underline{u}) = \mu(u_1)\mu(u_2) \cdots \mu(u_\ell).$$

In the finite sequence $S = X_1 X_2 \cdots X_n$ of length n , the number $N(\underline{u})$ of occurrences of \underline{u} in S is defined by

$$N(\underline{u}) = \sum_{i=1}^{n-\ell+1} \mathbb{I}_i(\underline{u}),$$

and its expectation is

$$(2.1) \quad \mathbb{E}N(\underline{u}) = (n - \ell + 1)\mu(\underline{u}).$$

2.1. Overlaps and clumps. Occurrences of a word may overlap in S or \mathcal{M} . Through this section the example $S = \text{TAAGAAGAAGAAGAAGT}$ and $\underline{u} = \text{AAGAAGAA}$ is used. In this case, the word \underline{u} occurs in S at positions 2, 5 and 8. The self-overlapping structure of a word can be described via the set of principal periods defined as follows. The lag between two overlapping occurrences of \underline{u} is said to be a *period* of the word \underline{u} [Guibas and Odlyzko (1981), Lothaire (1983)]. A word may have several periods; for any word \underline{u} , the set $\mathcal{P}(\underline{u})$ of the periods of \underline{u} , is defined as

$$\mathcal{P}(\underline{u}) := \{p \in \{1, \dots, \ell - 1\} : u_i = u_{i+p}, \forall i = 1, \dots, \ell - p\}.$$

The word \underline{u} is a non-self-overlapping word if and only if $\mathcal{P}(\underline{u})$ is empty. The most relevant periods (see for example (2.2) below) are the ones which are

not a nontrivial multiple of the minimal period. These periods are said to be *principal*; let $\mathcal{P}'(\underline{u})$ denote the set of the principal periods of \underline{u} . For example $\mathcal{P}(\underline{u}) = \{3, 6, 7\}$ and $\mathcal{P}'(\underline{u}) = \{3, 7\}$ for $\underline{u} = \text{AAGAAGAA}$.

In order to study the occurrences of a word \underline{u} the concept of clumps of a word \underline{u} is introduced. A *clump* of \underline{u} in a sequence is a maximal set of overlapping occurrences of \underline{u} in this sequence; no two clumps of \underline{u} overlap in the sequence. Say that a clump of \underline{u} starts at position i in the infinite sequence \mathcal{M} if an occurrence of \underline{u} starts at position i in \mathcal{M} and if this occurrence is not overlapped by a preceding occurrence of \underline{u} . Denote the corresponding indicator random variable by $\tilde{\mathbb{I}}_i(\underline{u})$; i.e.

$$\tilde{\mathbb{I}}_i(\underline{u}) = \mathbb{I}_i(\underline{u}) \prod_{j=i-\ell+1}^{i-1} (1 - \mathbb{I}_j(\underline{u})).$$

The probability $\tilde{\mu}(\underline{u})$ that a clump of \underline{u} starts at a given position in \mathcal{M} is exactly the expectation of $\tilde{\mathbb{I}}_i(\underline{u})$; Schbath (1995a) proved that

$$(2.2) \quad \tilde{\mu}(\underline{u}) := \mathbb{E}\tilde{\mathbb{I}}_i(\underline{u}) = \mu(\underline{u}) - \sum_{p \in \mathcal{P}'(\underline{u})} \mu(\underline{u}^{(p)}\underline{u}),$$

where $\underline{u}^{(p)} = u_1 u_2 \cdots u_p$ is the word composed of the first p letters of \underline{u} .

Here is the sketch of the proof for equation (2.2). A clump of \underline{u} starts at position i in the sequence if and only if there is an occurrence of \underline{u} starting at position i and there are none of the $\underline{u}^{(p)}$ starting at $i-p$ where $p \in \mathcal{P}(\underline{u})$. In fact, it suffices to exclude the occurrences of all the $\underline{u}^{(p)}$ at $i-p$ where p is only a principal period of \underline{u} . Thus

$$\begin{aligned} \tilde{\mathbb{I}}_i(\underline{u}) &= \mathbb{I}_i(\underline{u}) \mathbb{I} \left\{ \cap_{p \in \mathcal{P}'(\underline{u})} \{\text{no occurrence of } \underline{u}^{(p)} \text{ starts at } i-p\} \right\} \\ &= \mathbb{I}_i(\underline{u}) (1 - \mathbb{I} \left\{ \cup_{p \in \mathcal{P}'(\underline{u})} \{\text{an occurrence of } \underline{u}^{(p)} \text{ starts at } i-p\} \right\}) \end{aligned}$$

One can then show that the events $\{\text{an occurrence of } \underline{u}^{(p)} \text{ starts at } i-p\}$ for $p \in \mathcal{P}'(\underline{u})$ are disjoint, meaning that any two of them cannot occur simultaneously [Schbath (1995b)]. This leads to

$$\begin{aligned} \tilde{\mathbb{I}}_i(\underline{u}) &= \mathbb{I}_i(\underline{u}) \left(1 - \sum_{p \in \mathcal{P}'(\underline{u})} \mathbb{I}_{i-p}(\underline{u}^{(p)}) \right) \\ &= \mathbb{I}_i(\underline{u}) - \sum_{p \in \mathcal{P}'(\underline{u})} \mathbb{I}_{i-p}(\underline{u}^{(p)}\underline{u}); \end{aligned}$$

equation (2.2) then easily follows.

Now define $\tilde{N}(\underline{u})$ as the count

$$\tilde{N}(\underline{u}) := \sum_{i=1}^{n-\ell+1} \tilde{\mathbb{I}}_i(\underline{u}),$$

so that $\tilde{N}(\underline{u})$ represents the number of clumps of \underline{u} in the infinite sequence \mathcal{M} but starting in S . Its expectation is

$$\mathbb{E}\tilde{N}(\underline{u}) = (n - \ell + 1)\tilde{\mu}(\underline{u}).$$

For $1 \leq i \leq \ell - 1$ the definition of $\tilde{\mathbb{I}}_i(\underline{u})$ involves in particular the letters X_j with $i - \ell + 1 \leq j \leq 0$. In practice, only the sequence $S = X_1 X_2 \dots X_n$ is observable, and the observable number of clumps, denoted by $\tilde{N}^*(\underline{u})$, may be different from $\tilde{N}(\underline{u})$. In the above example, in S there is a unique clump of \underline{u} starting at position 2 and ending at position 15, and $X_1 = T$ ensures that this observable clump is a real one in the infinite sequence. This might not be the case if the first letter X_1 was a G. The quantity of interest is $\tilde{N}^*(\underline{u})$, the observable number of clumps, but here we will work with the count $\tilde{N}(\underline{u})$ instead, since it is easier and the boundary effect can be controlled. Indeed, $\mathbb{P}(\tilde{N}^*(\underline{u}) \neq \tilde{N}(\underline{u}))$ is an upper bound on the total variation distance between $\tilde{N}^*(\underline{u})$ and $\tilde{N}(\underline{u})$ [see, e.g., Barbour et al. (1992b)], and

$$\begin{aligned} \mathbb{P}(\tilde{N}^*(\underline{u}) \neq \tilde{N}(\underline{u})) &\leq \sum_{i=1}^{\ell-1} \mathbb{P}\left(\mathbb{I}_i(\underline{u}) = 1, \mathbb{I}_j(\underline{u}) = 1 \text{ for some } j = i - \ell + 1, \dots, i - 1\right) \\ &= \sum_{i=1}^{\ell-1} \mathbb{P}\left(\mathbb{I}_i(\underline{u}) = 1, \mathbb{I}_{i-p}(\underline{u}) = 1 \text{ for some } p \in \mathcal{P}(\underline{u})\right) \\ &= \sum_{i=1}^{\ell-1} \mathbb{P}\left(\mathbb{I}_i(\underline{u}) = 1, \mathbb{I}_{i-p}(\underline{u}) = 1 \text{ for some } p \in \mathcal{P}'(\underline{u})\right) \\ &= \sum_{i=1}^{\ell-1} \mathbb{P}\left(\mathbb{I}_{i-p}(\underline{u}^{(p)}\underline{u}) = 1 \text{ for some } p \in \mathcal{P}'(\underline{u})\right) \\ &\leq (\ell - 1) \sum_{p \in \mathcal{P}'(\underline{u})} \mu(\underline{u}^{(p)}\underline{u}) = (\ell - 1)(\mu(\underline{u}) - \tilde{\mu}(\underline{u})). \end{aligned}$$

As we think of $\mu(\underline{u})$ as being such that $n\mu(\underline{u})$ is bounded as n tends to infinity (the rare word condition), the above probability is very small for large

n . Therefore, $\tilde{N}^*(\underline{u})$ can be approximated by $\tilde{N}(\underline{u})$; as approximate $\tilde{N}(\underline{u})$ is approximated, the approximation for $\tilde{N}^*(\underline{u})$ follows.

Remark 1. If \underline{u} is not self-overlapping, meaning that \underline{u} has no period, then $\tilde{\mu}(\underline{u}) = \mu(\underline{u})$ and $\tilde{N}^*(\underline{u}) = \tilde{N}(\underline{u}) = N(\underline{u})$.

2.2. *Clumps of different sizes.* Now we distinguish clumps of different sizes. The size of a clump of \underline{u} is the maximal number of overlapping occurrences of \underline{u} contained in the clump. In the above example, the unique clump is of size 3. The structure of a clump can be complex, depending on the self-overlapping structure of the underlying word. Let $\mathcal{C}_k(\underline{u})$ be the set of all the concatenated words composed of exactly k overlapping occurrences of \underline{u} . For example,

$$\mathcal{C}_1(\underline{u}) = \{\text{AAGAAGAA}\},$$

$$\mathcal{C}_2(\underline{u}) = \{\text{AAGAAGAAGAA, AAGAAGAAAGAAGAA}\} \text{ and}$$

$$\begin{aligned} \mathcal{C}_3(\underline{u}) = & \{\text{AAGAAGAAGAAGAA, AAGAAGAAGAAAGAAGAA, AAGAAGAAAGAAGAAGAA,} \\ & \text{AAGAAGAAAGAAGAAAGAA}\} \end{aligned}$$

for $\underline{u} = \text{AAGAAGAA}$. Note that the length of two k -clumps may differ a lot, and that the size of $\mathcal{C}_k(\underline{u})$ increases exponentially with k whenever \underline{u} has more than one principal period. In fact,

$$|\mathcal{C}_k(\underline{u})| = |\mathcal{P}'(\underline{u})|^{k-1}$$

[see Schbath (1995b)].

Say that a k -clump of \underline{u} starts at position i in the infinite sequence \mathcal{M} if and only if a clump of \underline{u} starts at position i and this clump is composed of exactly k overlapping occurrences of \underline{u} . Denote the corresponding indicator random variable by $\tilde{\mathbb{I}}_{i,k}(\underline{u})$; Schbath (1995a) proved that its expectation is

$$\tilde{\mu}_k(\underline{u}) := \mathbb{E}\tilde{\mathbb{I}}_{i,k}(\underline{u}) = (1 - A(\underline{u}))^2 A(\underline{u})^{k-1} \mu(\underline{u}),$$

where

$$(2.3) \quad A(\underline{u}) = \sum_{p \in \mathcal{P}'(\underline{u})} \prod_{j=1}^p \mu(u_{j+1}) = \sum_{p \in \mathcal{P}'(\underline{u})} \frac{\mu(\underline{u}^{(p+1)})}{\mu(\underline{u}_1)}.$$

The derivation of (2.3) is similar to the one of (2.2). Note that

$$(2.4) \quad \sum_{k \geq 1} \tilde{\mu}_k(\underline{u}) = \tilde{\mu}(\underline{u}),$$

$$(2.5) \quad \sum_{k \geq 1} k \tilde{\mu}_k(\underline{u}) = \mu(\underline{u}).$$

Moreover, define $\tilde{N}_k(\underline{u})$ as the count

$$(2.6) \quad \tilde{N}_k(\underline{u}) := \sum_{i=1}^{n-\ell+1} \tilde{\mathbb{I}}_{i,k}(\underline{u}),$$

so that $\tilde{N}_k(\underline{u})$ represents the number of k -clumps of \underline{u} in the infinite sequence \mathcal{M} but starting in S . Again, because of the boundary effects, the count $\check{N}(\underline{u})$ defined by

$$(2.7) \quad \check{N}(\underline{u}) := \sum_{k \geq 1} k \tilde{N}_k(\underline{u})$$

is not equal to the count $N(\underline{u})$ of \underline{u} in the finite sequence S , but their difference is negligible. As they can differ only if a clump in \mathcal{M} overlaps positions 1 or n , the same techniques as for the number of clumps give that the total variation distance between $N(\underline{u})$ and $\check{N}(\underline{u})$ is bounded by

$$(2.8) \quad \mathbb{P}(N(\underline{u}) \neq \check{N}(\underline{u})) \leq 2(\ell - 1) \sum_{p \in \mathcal{P}'(\underline{u})} \mu(u^{(p)}\underline{u}) = 2(\ell - 1)(\mu(\underline{u}) - \tilde{\mu}(\underline{u})).$$

The counts $N(\underline{u})$ and $\check{N}(\underline{u})$ have the same expectation because of (2.5). Now focus on $\check{N}(\underline{u})$ to apply the Chen-Stein method.

Remark 2. If \underline{u} is not self-overlapping, meaning that \underline{u} has no period, then $\tilde{\mu}_1(\underline{u}) = \mu(\underline{u})$, $\tilde{\mu}_k(\underline{u}) = 0 \forall k \geq 2$, and $\check{N}(\underline{u}) = \tilde{N}_1(\underline{u}) = N(\underline{u})$.

2.3. The Chen-Stein method. The Chen-Stein method is a powerful tool for deriving Poisson approximations and compound Poisson approximations in terms of bounds on the total variation distance. For any two random processes \underline{Y} and \underline{Z} with values in the same space E , the total variation distance between their probability distributions is defined by

$$\begin{aligned} d_{\text{TV}}(\mathcal{L}(\underline{Y}), \mathcal{L}(\underline{Z})) &= \sup_{B \subset E} |\mathbb{P}(\underline{Y} \in B) - \mathbb{P}(\underline{Z} \in B)| \\ &= \sup_{h: E \rightarrow [0,1]} |\mathbb{E}h(\underline{Y}) - \mathbb{E}h(\underline{Z})|, \end{aligned}$$

where B and h are assumed to be measurable. The Chen-Stein method for Poisson approximation has been developed by Chen (1975); a friendly exposition is in Arratia et al. (1989, 1990); an exhaustive description with many examples can be found in Barbour et al. (1992b). We will use Theorem 1 in Arratia et al. (1990) with an improved bound by Barbour et al. (1992b) (Theorem 1.A and Theorem 10.A).

Theorem 1 (Arratia et al. (1990), Barbour et al. (1992b)). Let I be an index set. For each $\alpha \in I$, let Y_α be a Bernoulli random variable with $p_\alpha = \mathbb{P}(Y_\alpha = 1) > 0$. Suppose that, for each $\alpha \in I$, we have chosen $B_\alpha \subset I$ with $\alpha \in B_\alpha$. Let Z_α , $\alpha \in I$, be independent Poisson variables with mean p_α . The total variation distance between the Bernoulli process $\underline{Y} = (Y_\alpha, \alpha \in I)$ and the Poisson process $\underline{Z} = (Z_\alpha, \alpha \in I)$ satisfies

$$d_{TV}(\mathcal{L}(\underline{Y}), \mathcal{L}(\underline{Z})) \leq b_1 + b_2 + b_3,$$

where

$$(2.9) \quad b_1 = \sum_{\alpha \in I} \sum_{\beta \in B_\alpha} \mathbb{E} Y_\alpha \mathbb{E} Y_\beta$$

$$(2.10) \quad b_2 = \sum_{\alpha \in I} \sum_{\beta \in B_\alpha, \beta \neq \alpha} \mathbb{E}(Y_\alpha Y_\beta) \\ b_3 = \sum_{\alpha \in I} \mathbb{E} |\mathbb{E} \{Y_\alpha - p_\alpha | \sigma(Y_\beta, \beta \notin B_\alpha)\}|.$$

We think of B_α as a neighborhood of strong dependence of Y_α . Intuitively, b_1 describes the contribution related to the size of the neighborhood and the weights of the random variables in that neighborhood; if all Y_α had the same probability of success, then b_1 would be directly proportional to the neighborhood size. The term b_2 accounts for the strength of the dependence inside the neighborhood; as it depends on the second moments, it can be viewed as a “second order interaction” term. Finally, b_3 is related to the strength of dependence of Y_α with random variables outside its neighborhood. In particular, $b_3 = 0$ if Y_α is independent of $\sigma(Y_\beta, \beta \notin B_\alpha)$.

One consequence of this theorem is that for any indicator of an event, i.e. for any measurable functional h from E to $[0, 1]$, there is an error bound of the form $|\mathbb{E}h(\underline{Y}) - \mathbb{E}h(\underline{Z})| \leq d_{TV}(\mathcal{L}(\underline{Y}), \mathcal{L}(\underline{Z}))$. Thus, if $T(\underline{Y})$ is a test statistic then, for all $t \in \mathbb{R}$,

$$|\mathbb{P}(T(\underline{Y}) \geq t) - \mathbb{P}(T(\underline{Z}) \geq t)| \leq b_1 + b_2 + b_3,$$

which can be used to construct confidence intervals and to find p-values for tests based on this statistic.

3. Occurrences of m words of different lengths.

3.1. *Notation.* Now consider m different words $\underline{u}_1, \underline{u}_2, \dots, \underline{u}_m$ of length $\ell_1, \ell_2, \dots, \ell_m$, respectively;

$$\underline{u}_r = u_{r,1} u_{r,2} \cdots u_{r,\ell_r}, \quad \forall r \in \{1, \dots, m\}.$$

Assume that

(A) $\forall r \neq r'$, \underline{u}_r is not a substring of any composed word in $\mathcal{C}_2(\underline{u}_{r'})$.

Clumps of \underline{u}_r and clumps of $\underline{u}_{r'}$ may overlap in the sequence. Assumption (A) guarantees that a clump of \underline{u}_r and a clump of $\underline{u}_{r'}$ can overlap on at most $\max\{\ell_r, \ell_{r'}\} - 1$ letters. Heuristically, if \underline{u}_r and $\underline{u}_{r'}$ do not satisfy Assumption (A), the approximation of their counts by independent Poisson variables should not be valid.

In order to describe the possible overlaps between two words \underline{u}_r and $\underline{u}_{r'}$, define

$$\mathcal{P}(\underline{u}_r, \underline{u}_{r'}) := \{p \in \{1, \dots, \ell_r - 1\} : u_{r',i} = u_{r,i+p}, \text{ for all } i = 1, \dots, \ell_r - p\}.$$

Under Assumption (A), if $p \in \mathcal{P}(\underline{u}_r, \underline{u}_{r'})$ then $p > \ell_r - \ell_{r'}$ and the last $(\ell_r - p)$ letters of \underline{u}_r are equal to the first $(\ell_r - p)$ letters of $\underline{u}_{r'}$ ($\underline{u}_{r'}$ can overlap \underline{u}_r from the right). Note the lack of symmetry; for example, for $\underline{u} = \text{AAGAAGAA}$ and $\underline{v} = \text{AAGAATCA}$, it follows that $\mathcal{P}(\underline{u}, \underline{v}) = \{3, 6, 7\}$ and $\mathcal{P}(\underline{v}, \underline{u}) = \{7\}$.

For a bound on the error in the compound Poisson approximation, the following quantities, defined for all r and $r' \in \{1, \dots, m\}$, are needed.

$$(3.11) M(\underline{u}_r, \underline{u}_{r'}) = \mathbf{1}(r \neq r') \sum_{p \in \mathcal{P}(\underline{u}_r, \underline{u}_{r'})} \frac{1}{\mu(\underline{u}_{r'}^{(\ell_r-p)})}$$

$$(3.12) R(\underline{u}_r, \underline{u}_{r'}) = (n - \ell_r + 1)\{(\ell_r - 1)\mu(\underline{u}_r)\tilde{\mu}(\underline{u}_{r'}) + (\ell_{r'} - 1)\tilde{\mu}(\underline{u}_r)\mu(\underline{u}_{r'}) \\ + (2\ell_r + 2\ell_{r'} - 3)\tilde{\mu}(\underline{u}_r)\tilde{\mu}(\underline{u}_{r'})\}$$

$$(3.13) T(\underline{u}_r, \underline{u}_{r'}) = (n - \ell_r + 1)\mu(\underline{u}_r)\mu(\underline{u}_{r'})\{2(\ell_r + \ell_{r'} - 2) + M(\underline{u}_r, \underline{u}_{r'}) \\ + M(\underline{u}_{r'}, \underline{u}_r)\}.$$

The quantity $M(\underline{u}_r, \underline{u}_{r'})$ can be seen as a measure of the overlapping structure between \underline{u}_r and $\underline{u}_{r'}$. If \underline{u}_r and $\underline{u}_{r'}$ cannot overlap, $M(\underline{u}_r, \underline{u}_{r'})$ equals zero; otherwise, the more they can overlap from the right, the larger is $M(\underline{u}_r, \underline{u}_{r'})$. The quantities R and T correspond to the quantities R and T in Reinert and Schbath (1998). It will turn out that R is used to describe the “neighborhood size term” b_1 , whereas T describes the “second order interaction term” b_2 , when applying Theorem 1.

Moreover introduce the set of possible words of length $\ell - 1$ preceding a clump of \underline{u} :

$$\mathcal{G}(\underline{u}) = \{\underline{g} = g_1 \cdots g_{\ell-1} : \text{for all } p \in \mathcal{P}(\underline{u}), g_{\ell-p} \cdots g_{\ell-1} \neq \underline{u}^{(p)}\}.$$

Similarly, $\mathcal{D}(\underline{u})$ is the set of words allowed after a clump of \underline{u} ;

$$\mathcal{D}(\underline{u}) = \{\underline{d} = d_1 \cdots d_{\ell-1} : \forall p \in \mathcal{P}(\underline{u}), d_1 \cdots d_p \neq u_{\ell-p+1} \cdots u_\ell\}.$$

Recall that $\mathcal{C}_k(\underline{u})$ is the set of all the concatenated words composed of exactly k overlapping occurrences of \underline{u} . Thus, a k -clump of \underline{u} starts at position i in the infinite sequence \mathcal{M} if and only if one of the words \underline{gCd} , where $\underline{g} \in \mathcal{G}(\underline{u})$, $\underline{C} \in \mathcal{C}_k(\underline{u})$ and $\underline{d} \in \mathcal{D}(\underline{u})$, occurs at position $i - \ell + 1$. From Schbath (1995a), no two different \underline{C} and \underline{C}' in $\mathcal{C}_k(\underline{u})$ can occur simultaneously at position i . Therefore,

$$\tilde{\mathbb{I}}_{i,k}(\underline{u}_r) = \sum_{\underline{g} \in \mathcal{G}(\underline{u}_r), \underline{C} \in \mathcal{C}_k(\underline{u}_r), \underline{d} \in \mathcal{D}(\underline{u}_r)} \mathbb{I}_{i-\ell_r+1}(\underline{gCd}).$$

Note that

$$(3.14) \quad \sum_{k \geq 1} \sum_{\underline{g} \in \mathcal{G}(\underline{u}), \underline{C} \in \mathcal{C}_k(\underline{u})} \mu(\underline{gC}) = \sum_{k \geq 1} \sum_{k^* \geq k} \tilde{\mu}_{k^*}(\underline{u}) = \mu(\underline{u}).$$

3.2. Results. Enumerate the elements of $\mathcal{C}_k(\underline{u}_r)$ from C_1 to $C_{|\mathcal{P}'(\underline{u}_r)|^{k-1}}$. Select the index set

$$I = \{(i, r, k, c) : 1 \leq i \leq n, r = 1, \dots, m, k = 1, 2, \dots, c = 1, \dots, |\mathcal{P}'(\underline{u}_r)|^{k-1}\}.$$

For each $(i, r, k, c) \in I$, define the Bernoulli process $\tilde{\mathbb{Y}} = (\tilde{Y}_{(i,r,k,c)})_{(i,r,k,c) \in I}$ by

$$\tilde{Y}_{(i,r,k,c)} = \sum_{\underline{g} \in \mathcal{G}(\underline{u}_r), \underline{d} \in \mathcal{D}(\underline{u}_r)} \mathbb{I}_{i-\ell_r+1}(\underline{gCd}).$$

Thus $\tilde{Y}_{(i,r,k,c)}$ equals 1 if and only if a specific clump \underline{C}_c of size k of word \underline{u}_r occurs at position i in the sequence. Furthermore define the Poisson process $\underline{\mathbb{Z}} = (Z_{i,r,k,c})_{(i,r,k,c) \in I}$ by having independent components and each component being Poisson distributed with mean $\mathbb{E}\tilde{Y}_{(i,r,k,c)}$. We think of $Z_{i,r,k,c}$ approximating the indicator random variable $\tilde{Y}_{(i,r,k,c)}$.

For $(i, r, k, c) \in I$, choose as neighborhood

$$B_{i,r,k,c} := \{(j, r', k', c') \in I : -|\underline{C}_{c'}| - \ell_{r'} - \ell_r + 3 \leq j - i \leq |\underline{C}_c| + \ell_r + \ell_{r'} - 3\}.$$

Theorem 2. *Under Assumption (A) and with the notation (3.12) and (3.13), we have*

$$d_{TV} \left(\mathcal{L}(\tilde{\mathbb{Y}}), \mathcal{L}(\underline{\mathbb{Z}}) \right) \leq \sum_{1 \leq r, r' \leq m} \{R(\underline{u}_r, \underline{u}_{r'}) + T(\underline{u}_r, \underline{u}_{r'})\}.$$

Let $(Z_k^{(r)})_{k \geq 1, r \in \{1, \dots, m\}}$ be independent Poisson variables with expectation $\mathbb{E}Z_k^{(r)} = (n - \ell_r + 1)\tilde{\mu}_k(\underline{u}_r)$. From Theorem 2 and Equation (2.6), the following corollary is easily obtained.

Corollary 1. *Under Assumption (A) and with the notation (3.12) and (3.13), we have*

$$\begin{aligned} d_{TV}\left(\mathcal{L}((\tilde{N}_k(\underline{u}_r))_{k \geq 1, r \in \{1, \dots, m\}}), \mathcal{L}((Z_k^{(r)})_{k \geq 1, r \in \{1, \dots, m\}})\right) \\ \leq \sum_{1 \leq r, r' \leq m} \{R(\underline{u}_r, \underline{u}_{r'}) + T(\underline{u}_r, \underline{u}_{r'})\}. \end{aligned}$$

Now let $CP^{(r)}$ denote the compound Poisson distribution of $\sum_{k \geq 1} k Z_k^{(r)}$. From Theorem 2, Equation (2.7) and Inequality (2.8), we have the following two corollaries.

Corollary 2. *Under Assumption (A) and with the notation (3.12) and (3.13) we have*

$$\begin{aligned} (i) \quad d_{TV}\left(\mathcal{L}(\check{N}(\underline{u}_1), \dots, \check{N}(\underline{u}_m)), CP^{(1)} \otimes \dots \otimes CP^{(m)}\right) \\ \leq \sum_{1 \leq r, r' \leq m} \{R(\underline{u}_r, \underline{u}_{r'}) + T(\underline{u}_r, \underline{u}_{r'})\}, \\ (ii) \quad d_{TV}\left(\mathcal{L}(N(\underline{u}_1), \dots, N(\underline{u}_m)), CP^{(1)} \otimes \dots \otimes CP^{(m)}\right) \\ \leq \sum_{1 \leq r, r' \leq m} \{R(\underline{u}_r, \underline{u}_{r'}) + T(\underline{u}_r, \underline{u}_{r'})\} + 2 \sum_{r=1}^m (\ell_r - 1)(\mu(\underline{u}_r) - \tilde{\mu}(\underline{u}_r)). \end{aligned}$$

Let $(Z_k)_{k \geq 1}$ be independent Poisson variables with expectation

$$\mathbb{E}Z_k = \sum_{r=1}^m (n - \ell_r + 1)\tilde{\mu}_k(\underline{u}_r) ,$$

and let CP denote the compound Poisson distribution of $\sum_{k \geq 1} k Z_k$.

Corollary 3. *Under Assumption (A) and with the notation (3.12) and (3.13), we have*

$$\begin{aligned} d_{TV}\left(\mathcal{L}\left(\sum_{r=1}^m N(\underline{u}_r)\right), CP\right) &\leq \sum_{1 \leq r, r' \leq m} \{R(\underline{u}_r, \underline{u}_{r'}) + T(\underline{u}_r, \underline{u}_{r'})\} \\ &+ 2 \sum_{r=1}^m (\ell_r - 1)(\mu(\underline{u}_r) - \tilde{\mu}(\underline{u}_r)). \end{aligned}$$

Note that the compound Poisson distributions defined in Corollaries 2 and 3 reduce to simple Poisson distributions if every \underline{u}_r , $r = 1, \dots, m$, is a non self-overlapping word.

Poisson approximations should be good for rare events. Here, rare words mean that $\mathbb{E}N(\underline{u}_r)$ is bounded away from 0 and ∞ for $r = 1, \dots, m$; we use the notation $\mathbb{E}N(\underline{u}_r) \asymp 1$. For a fixed alphabet, this asymptotic framework is equivalent to $\mu(\underline{u}_r) \asymp \frac{1}{n}$, and $\ell_r \asymp \log n$ because

$$\mu(u_1)(\min_{x \in \mathcal{A}} \mu(x))^{\ell_r - 1} \leq \mu(\underline{u}_r) \leq \mu(u_1)(\max_{x \in \mathcal{A}} \mu(x))^{\ell_r - 1}.$$

If $\mathbb{E}N(\underline{u}_r) \asymp 1$, then $T(\underline{u}_r, \underline{u}_{r'}) \asymp n^{-1} \log n + (M(\underline{u}_r, \underline{u}_{r'}) + M(\underline{u}_{r'}, \underline{u}_r))n^{-1}$, and

$$d_{\text{TV}} \left(\mathcal{L}(\widetilde{\mathbb{Y}}), \mathcal{L}(\mathbb{Z}) \right) \leq O(n^{-1} \log n) + O \left(\sum_{r \neq r'} (M(\underline{u}_r, \underline{u}_{r'}) + M(\underline{u}_{r'}, \underline{u}_r))n^{-1} \right).$$

Therefore, if \underline{u}_r and $\underline{u}_{r'}$ cannot overlap too much for $r \neq r'$ (this is measured by $M(\underline{u}_r, \underline{u}_{r'}) + M(\underline{u}_{r'}, \underline{u}_r)$), the error bound is very small for large n . In the extreme case where $\underline{u}_1 = \text{AAA} \cdots \text{AAA}$ and $\underline{u}_2 = \text{TAA} \cdots \text{AAA}$, both of length ℓ , we have $M(\underline{u}_2, \underline{u}_1) = \sum_{s=1}^{\ell-1} \mu(\text{A})^{-s} \asymp \mu(\text{A})^{-\log n}$, so the error bound may fail to converge to zero as n tends to infinity. This confirms the intuition that, because of considerable overlaps, the occurrences of \underline{u}_1 and \underline{u}_2 should not be independent even asymptotically.

3.3. Proof of Theorem 2. Our task consists now in bounding b_1 and b_2 given in (2.9) and (2.10). For any $\underline{C} \in \mathcal{C}_k(\underline{u}_r)$ it follows that $|\underline{C}| \leq k(\ell_r - 1)$. This motivates the introduction of the set

$$(3.15) \quad B_{i,k,k',r,r'} := \{j \in \{1, \dots, n\} : i - (k' + 1)(\ell_{r'} - 1) + \ell_r - 1 \leq j \leq i + (k + 1)(\ell_r - 1) + \ell_{r'} - 1\};$$

for each fixed i, r, r', k, k', c, c' , thus $\{j : (j, r', k', c') \in B_{i,r,k,c}\} \subset B_{i,k,k',r,r'}$.

Bounding b_1 : We have

$$\begin{aligned} b_1 &= \sum_{(i,r,k,c) \in I} \sum_{(j,r',k',c') \in B_{i,r,k,c}} \mathbb{E} \tilde{Y}_{i,r,k,c} \mathbb{E} \tilde{Y}_{j,r',k',c'} \\ &\leq \sum_{r,r'=1}^m \sum_{k,k' \geq 1} \sum_{i=1}^{n-\ell_r+1} \sum_{j \in B_{i,k,k',r,r'}} \tilde{\mu}_k(\underline{u}_r) \tilde{\mu}_{k'}(\underline{u}_{r'}) \\ &\leq \sum_{r,r'=1}^m (n - \ell_r + 1) \end{aligned}$$

$$\times \sum_{k,k' \geq 1} \tilde{\mu}_k(\underline{u}_r) \tilde{\mu}_{k'}(\underline{u}_{r'}) ((k+2)(\ell_r - 1) + (k'+2)(\ell_{r'} - 1) + 1),$$

where $B_{i,k,k',r,r'}$ is defined in (3.15). Now use (2.4) and (2.5) to obtain

$$\begin{aligned} b'_2 &\leq \sum_{r,r'=1}^m (n - \ell_r + 1) \{ (\ell_r - 1) \mu(\underline{u}_r) \tilde{\mu}(\underline{u}_{r'}) + (\ell_{r'} - 1) \tilde{\mu}(\underline{u}_r) \mu(\underline{u}_{r'}) \\ &\quad + (2\ell_r + 2\ell_{r'} - 3) \tilde{\mu}(\underline{u}_r) \tilde{\mu}(\underline{u}_{r'}) \} \\ &= \sum_{r,r'=1}^m R(\underline{u}_r, \underline{u}_{r'}). \end{aligned}$$

Bounding b_2 : To bound b_2 write

$$b_2 = \sum_{(i,r,k,c) \in I} \sum_{(j,r',k',c') \in B_{i,r,k,c} \setminus \{(i,r,k,c)\}} \mathbb{E} \tilde{Y}_{i,r,k,c} \tilde{Y}_{j,r',k',c'} = \sum_{r,r'=1}^m b'_2(\underline{u}_r, \underline{u}_{r'}),$$

with

$$b'_2(\underline{u}_r, \underline{u}_{r'}) = \sum_{i=1}^{n-\ell_r+1} \sum_{k,k' \geq 1} \sum_{j: (j,r',k',c') \in B_{i,r,k,c} \setminus \{(i,r,k,c)\}} \mathbb{E} \tilde{Y}_{i,r,k,c} \tilde{Y}_{j,r',k',c'}.$$

Now distinguish two cases: The first one when the k -clump starting at i and the k' -clump starting at j overlap in \mathcal{M} , and the second one when the clumps do not overlap but the “enlarged clumps” – including the $(\ell_r - 1)$ preceding letters and the $(\ell_{r'} - 1)$ following letters of the clumps – overlap. Write $\underline{C} = \underline{C}_c$, and $\underline{C}' = \underline{C}'_{c'}$, for convenience.

1. First consider the case when \underline{C} and \underline{C}' overlap in the sequence, that is,

$$j \in \{i + |\underline{C}| - \ell_r, \dots, i + |\underline{C}| - 1\} \cup \{i - |\underline{C}'| + 1, \dots, i - |\underline{C}'| + \ell_{r'}\}.$$

Let $b'_{21}(\underline{u}_r, \underline{u}_{r'})$ denote the quantity corresponding to this case. Since two clumps of \underline{u}_r cannot overlap in the sequence, the composed words \underline{C} and \underline{C}' starting at i and j cannot overlap. Therefore we may restrict ourselves to the case $r \neq r'$. First focus on $j > i$. If \underline{C} and \underline{C}' overlap, Assumption (A) ensures that only the last occurrence of \underline{u}_r in \underline{C} overlaps with the first occurrence of $\underline{u}_{r'}$ in \underline{C}' . The last occurrence of \underline{u}_r in \underline{C} starts at position $i + |\underline{C}| - \ell_r$. An occurrence of $\underline{u}_{r'}$ starting at position j may overlap \underline{u}_r at $i + |\underline{C}| - \ell_r$ only if $j = i + |\underline{C}| - \ell_r + p$

with $p \in \mathcal{P}(\underline{u}_r, \underline{u}_{r'})$. Therefore,

$$\begin{aligned}
& \sum_{i=1}^{n-\ell_r+1} \sum_{k, k' \geq 1} \sum_{\substack{\underline{g} \in \mathcal{G}(\underline{u}_r) \\ \underline{d} \in \mathcal{D}(\underline{u}_r) \\ \underline{C} \in \mathcal{C}_k(\underline{u}_r)}} \sum_{\substack{\underline{g}' \in \mathcal{G}(\underline{u}_{r'}) \\ \underline{d}' \in \mathcal{D}(\underline{u}_{r'}) \\ \underline{C}' \in \mathcal{C}_{k'}(\underline{u}_{r'})}} \sum_{j=i+|\underline{C}|-l_r}^{i+|\underline{C}|-1} \mathbb{E} \mathbb{I}_{i-\ell_r+1}(\underline{g} \underline{C} \underline{d}) \mathbb{I}_{j-\ell_{r'}+1}(\underline{g}' \underline{C}' \underline{d}') \\
& \leq \sum_{i=1}^{n-\ell_r+1} \sum_{\substack{k \geq 1 \\ k' \geq 1}} \sum_{\substack{\mathcal{G}(\underline{u}_r) \\ \mathcal{C}_k(\underline{u}_r)}} \sum_{\substack{\mathcal{G}(\underline{u}_{r'}) \\ \mathcal{D}(\underline{u}_{r'}) \\ \mathcal{C}_{k'}(\underline{u}_{r'})}} \sum_{p \in \mathcal{P}(\underline{u}_r, \underline{u}_{r'})} \mathbb{E} \mathbb{I}_{i-\ell_r+1}(\underline{g} \underline{C}) \mathbb{I}_{i+|\underline{C}|-l_r-\ell_{r'}+p+1}(\underline{g}' \underline{C}' \underline{d}') \\
& \leq \sum_{i=1}^{n-\ell_r+1} \sum_{k \geq 1} \sum_{\substack{\underline{g} \in \mathcal{G}(\underline{u}_r) \\ \underline{C} \in \mathcal{C}_k(\underline{u}_r)}} \sum_{p \in \mathcal{P}(\underline{u}_r, \underline{u}_{r'})} \mathbb{E} \mathbb{I}_{i-\ell_r+1}(\underline{g} \underline{C}) \mathbb{I}_{i+|\underline{C}|-l_r+p}(\underline{u}_{r'}) ;
\end{aligned}$$

the last inequality comes from summing over $\underline{d}', k', \underline{g}', \underline{C}', \underline{d}$, and then using that $\widetilde{\mathbb{I}}_{i+|\underline{C}|-l_r+p}(\underline{u}_{r'}) \leq \mathbb{I}_{i+|\underline{C}|-l_r+p}(\underline{u}_{r'})$. Now, $\underline{g} \underline{C}$ starting at $i - \ell_r + 1$ and $\underline{u}_{r'}$ starting at $i + |\underline{C}| - l_r + p$ overlap at most on $(\ell_r - 1)$ letters; thus

$$\mathbb{E} \mathbb{I}_{i-\ell_r+1}(\underline{g} \underline{C}) \mathbb{I}_{i+|\underline{C}|-l_r+p}(\underline{u}_{r'}) \leq \mu(\underline{g} \underline{C}) \frac{\mu(\underline{u}_{r'})}{\mu(\underline{u}_{r'})^{(\ell_r-p)}}.$$

Finally, using (3.14) and applying the same reasoning to $j < i$ yields

$$(3.16) \quad b'_{21}(\underline{u}_r, \underline{u}_{r'}) \leq (n - \ell_r + 1) \mu(\underline{u}_r) \mu(\underline{u}_{r'}) (M(\underline{u}_r, \underline{u}_{r'}) + M(\underline{u}_{r'}, \underline{u}_r))$$

where M is given by (3.11).

2. $\underline{g} \underline{C} \underline{d}$ and $\underline{g}' \underline{C}' \underline{d}'$ overlap in the sequence (at most on $\ell_r + \ell_{r'} - 2$ letters), but \underline{C} and \underline{C}' do not overlap, that is,

$$j \in \{i - |\underline{C}'| - \ell_{r'} - \ell_r + 3, \dots, i - |\underline{C}'|\} \cup \{i + |\underline{C}|, \dots, i + |\underline{C}| + \ell_r + \ell_{r'} - 3\}.$$

Denote the corresponding quantity by $b'_{22}(\underline{u}_r, \underline{u}_{r'})$. For the case that $j > i$ it follows

$$\sum_{i=1}^{n-\ell_r+1} \sum_{k, k' \geq 1} \sum_{\substack{\underline{g} \in \mathcal{G}(\underline{u}_r), \underline{d} \in \mathcal{D}(\underline{u}_r) \\ \underline{C} \in \mathcal{C}_k(\underline{u}_r)}} \sum_{j=i+|\underline{C}|}^{i+|\underline{C}|+\ell_r+\ell_{r'}-3} \mathbb{E} \mathbb{I}_{i-\ell_r+1}(\underline{g} \underline{C} \underline{d}) \widetilde{\mathbb{I}}_{j, k'}(\underline{u}_{r'})$$

$$\begin{aligned}
&\leq \sum_{i=1}^{n-\ell+1} \sum_{k,k' \geq 1} \sum_{\substack{\underline{g} \in \mathcal{G}(\underline{u}_r) \\ \underline{C} \in \mathcal{C}_k(\underline{u}_r)}} \sum_{j=i+|\underline{C}|}^{i+|\underline{C}|+\ell_r+\ell_{r'}-3} \mathbb{E} \mathbb{I}_{i-\ell_r+1}(\underline{g} \underline{C}) \mathbb{I}_{j,k'}(\underline{u}_{r'}) \\
&= \sum_{i=1}^{n-\ell_r+1} \sum_{k \geq 1} \sum_{\substack{\underline{g} \in \mathcal{G}(\underline{u}) \\ \underline{C} \in \mathcal{C}_k(\underline{u})}} \sum_{j=i+|\underline{C}|}^{i+|\underline{C}|+\ell_r+\ell_{r'}-3} \mu(\underline{g} \underline{C}) \mu(\underline{u}_{r'}).
\end{aligned}$$

The case $j < i$ is treated analogously. Summing and using (3.14) leads to

$$(3.17) \quad b'_{22}(\underline{u}_r, \underline{u}_{r'}) \leq 2(n - \ell_r + 1)(\ell_r + \ell_{r'} - 2)\mu(\underline{u}_r)\mu(\underline{u}_{r'}).$$

Combining (3.17), and (3.16) gives $b'_2(\underline{u}, \underline{v}) \leq T(\underline{u}, \underline{v})$. ■

4. Application. To illustrate the goodness of the approximation for collections of words in DNA sequences, consider motifs on the four-letter DNA alphabet $\mathcal{A} = \{A, C, G, T\}$ having the structure

$$(4.18) \quad a_1 a_2 \cdots a_r (\text{N})_s \overline{a_r} \cdots \overline{a_2} \overline{a_1},$$

where the integers r and s are fixed, and $a_i \in \mathcal{A}$ for $i = 1, \dots, r$. Here N represents any letter in the alphabet \mathcal{A} , $(\text{N})_s$ denotes s consecutive, possibly different letters N , and for each $i = 1, \dots, r$, $\overline{a_i}$ is the complementary letter of a_i (A is the complementary letter of T and vice versa, and C is the complementary letter of G and vice versa). For example, the motif AGGCNNNGCCT is such a motif, involving the collection of the sixteen words $(\text{AGGC}ab\text{GCCT})_{a,b \in \mathcal{A}}$ of length 10.

These motifs are particularly interesting because of their possible stem-loop structure; the prefix $a_1 a_2 \cdots a_r$ could form a stem with the suffix $\overline{a_r} \cdots \overline{a_2} \overline{a_1}$, leading to a loop of length s as shown by Figure 1. Such a structure may lead to errors when the polymerase replicates the genome. This phenomenon can also occur with RNA sequences.

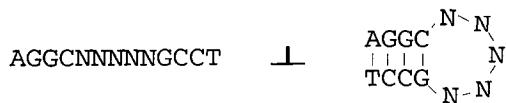


FIG. 1. *Stem-loop structure*

To assess its extent, the number of occurrences of these motifs in a DNA sequence are approximated. The number of occurrences of the motif AGGCNNNGCCT,

for instance, is denoted by $N(\text{AGGCNNGCCT})$ and can be easily obtained by summing the numbers of occurrences of each word $\text{AGGC}ab\text{GCCT}$, $a, b \in \mathcal{A}$:

$$N(\text{AGGCNNGCCT}) = \sum_{a,b \in \mathcal{A}} N(\text{AGGC}ab\text{GCCT}).$$

The expected count of AGGCNNGCCT in a sequence is

$$\mathbb{E}N(\text{AGGCNNGCCT}) = \sum_{a,b \in \mathcal{A}} \mathbb{E}N(\text{AGGC}ab\text{GCCT}),$$

where the right-hand terms are given in (2.1). Using Corollary 3, we calculate the error bound for approximating the count $N(\text{AGGCNNGCCT})$ by a compound Poisson variable $\sum_{k \geq 1} kZ_k$ such that the Z_k 's are independent Poisson variables with expectation

$$\mathbb{E}Z_k = \sum_{a,b \in \mathcal{A}} (1 - A(\text{AGGCNNGCCT}))^2 A(\text{AGGCNNGCCT})^{k-1} \mathbb{E}N(\text{AGGCNNGCCT}),$$

where A is given in (2.3).

In the application below, a sequence of length $n = 48502$ on the alphabet \mathcal{A} is considered, with the letter probabilities

$$(4.19) \quad \mu(A) = .2544 \quad \mu(C) = .2342 \quad \mu(G) = .2643 \quad \mu(T) = .2471.$$

These values correspond approximately to the genome of the bacteriophage *Lambda*.

Table 1 gives the expected counts in the sequence of the motifs AGGCCCT , ATGCGCAT , ATGGCGCCAT , and ATTGGCGCCAAT ; the first two nonzero digits are given. Note that inserting N 's in the word does not change their expected count, so that the expected count of $\text{AGGC}(N)_3\text{GCCT}$, for example, equals the expected count of AGGCCCT , which is 0.72. All the above motifs have a very small expected count, which is in agreement with the rare word condition. Naturally, the expected counts decrease with increasing sequence length.

For a stem of fixed length, increasing the size s of the loop does not change the order of the bound substantially; it only gently increases. For computational reasons we restrict our study to loop sizes less or equal than 3, even though relevant biological loop sizes are slightly larger. This result is particularly interesting since increasing s means enormously increasing the number of words in our family, which is of course penalizing.

Comparing the results for the motifs $\text{AGGC}(N)_s\text{GCCT}$ and $\text{ATGC}(N)_s\text{GCAT}$, the bounds for $\text{ATGC}(N)_s\text{GCAT}$ are larger because of its more complicated overlapping

TABLE 1

Expected counts of some stem-loop motifs, in a sequence of length 48,502 of i.i.d. letters generated by (4.19)

	expected count
AGGCGCCT	0.72
ATGCCGCAT	0.73
ATGGCGCCAT	4.5e-02
ATTGGCGCCAAT	2.8e-03

structure. Consider, for instance, the motif ATGCNNNGCAT; Figure 2 describes the different possible overlaps that can occur between two words belonging to this motif. Let u and v be two words belonging to the ATGCNNNGCAT family; the set of periods $\mathcal{P}(u, v)$ is then necessarily equal to either $\{5, 9\}$ (16 pairs) or $\{9\}$ ($4^6 - 16$ pairs). Moreover, the set $\mathcal{P}(u)$ is equal to $\{9\}$ for all words in the motif.

A T G C a b c G C A T	A T G C a' b' c' G C A T
↑	
9	
A T G C a A T G C A T	A T G C A T b G C A T
↑	
5	

FIG. 2. *Self-overlaps of the family ATGCNNNGCAT*

In contrast, $\mathcal{P}(u) = \emptyset$ for all words u in the AGGCNNNGCCT family. Figure 3 describes the possible overlaps between two words of the AGGCNNNGCCT family. Among all the pairs (u, v) in the AGGCNNNGCCT family, only 16 pairs have a nonempty period set $\mathcal{P}(u, v)$, which in this case equals $\{5\}$.

A G G C a A G G C C T	
A G G C C T b G C C T	
↑	
5	

FIG. 3. *Self-overlaps of the family AGGCNNNGCCT*

A motif composed of words with many overlaps between themselves will pro-

duce large quantities M and hence large quantities T (given by (3.13)). The overlapping structure of a motif has an important influence on the error bound through the terms T . This explains the larger bounds for $\text{ATGC}(N)_s\text{GCAT}$.

However, Table 2 shows that adding just one letter to the stem of the motifs $\text{ATGC}(N)_s\text{GCAT}$, in such a way that the high overlapping structure of the motif is preserved, for instance yielding $\text{ATGGC}(N)_s\text{GCCAT}$, is sufficient to reduce the global error bound considerably to the order of 10^{-6} . Adding another letter leads to an error bound of order 10^{-8} for $\text{ATTGGC}(N)_s\text{GCCAAT}$. This illustrates that the approximation improves with increasing word length, i.e. with decreasing expected count of the motif (see Table 1).

Now focus on the weight of each of the three terms appearing in the global error bound given by Corollary 3. The three terms correspond, respectively, to the bounds of b_1 , b_2 , and the boundary effect calculated in Section 3. Table 2 gives the bound for each term and for the motifs $\text{AGGC}(N)_s\text{GCCT}$, $\text{ATGC}(N)_s\text{GCAT}$, $\text{ATGGC}(N)_s\text{GCCAT}$, and $\text{ATTGGC}(N)_s\text{GCCAAT}$, s varying from 1 to 3. It is obvious that the boundary effect is negligible and decreases smoothly as s increases. On the other hand, b_1 and b_2 are the main terms, and their bounds are about of the same order. These bounds increase very slightly while increasing the loop size s , leading to a small increase of the global error bound. Note that there is no boundary effect for $\text{AGGC}(N)_s\text{GCCT}$, $s = 1, \dots, 3$; the explanation is that these motifs are composed of non-self-overlapping words (see (2.8)).

TABLE 2

Weight of the different terms involved in the global error bound for the compound Poisson approximation of some stem-loop motif counts, in a sequence of length 48,502 of i.i.d. letters generated by (4.19)

	b_1	b_2	boundary effect	global bound
AGGCNGCCT	4.4e-04	3.4e-04	0	7.8e-04
AGGCNNGCCT	5.4e-04	4.3e-04	0	9.8e-04
AGGCNNNGCCT	5.4e-04	7.7e-04	0	1.3e-03
ATGCNGCAT	4.5e-04	6.2e-04	1.4e-08	1.0e-03
ATGCNNGCAT	5.1e-04	7.3e-04	7.0e-08	1.2e-03
ATGCNNNGCAT	5.6e-04	1.1e-03	1.1e-09	1.7e-03
ATGGCNGCCAT	2.1e-06	2.7e-06	6.9e-11	4.8e-06
ATGGCNGCCAT	2.3e-06	3.1e-06	1.9e-11	5.5e-06
ATGGCNNNGCCAT	2.5e-06	4.7e-06	5.2e-12	7.3e-06
ATTGGCNGCCAGT	1.0e-08	1.2e-08	3.3e-13	2.2e-08
ATTGGCNGNGCCAGT	1.1e-08	1.3e-08	9.0e-14	2.4e-08
ATTGGCNGNNNGCCAGT	1.1e-08	1.4e-08	2.4e-14	2.6e-08

In Reinert and Schbath (1998), corresponding bounds are calculated for the case that the sequence is not composed of i.i.d. letters but generated by a stationary Markov case. In comparison, the bounds for the independent model

are orders of magnitude smaller.

Acknowledgment: The first author would like to thank the organizers for the opportunity to present this paper. Moreover we would like to thank Michael Waterman and Terry Speed for very helpful discussions. Finally the anonymous referees deserve thanks for many useful suggestions.

REFERENCES

- ARRATIA, R., GOLDSTEIN, L. and GORDON, L. 1989. Two moments suffice for Poisson approximations: the Chen-Stein method. *Ann. Prob.* **17** 9–25.
- ARRATIA, R., GOLDSTEIN, L. and GORDON, L. 1990. Poisson approximation and the Chen-Stein method. *Statistical Science* **5** 403–434.
- BARBOUR, A. D., CHEN, L. H. Y. and LOH, W.-L. 1992a. Compound Poisson approximation for nonnegative random variables via Stein's method. *Ann. Prob.* **20** 1843–1866.
- BARBOUR, A. D., HOLST, L. and JANSON, S. 1992b. *Poisson approximation*. Oxford - University Press.
- BARBOUR, A. D. and UTEV, U. 1997. Compound Poisson approximation in total variation. Preprint.
- CHEN, L. H. Y. 1975. Poisson approximation for dependent trials. *Ann. Prob.* **3** 534–545.
- GESKE, M. X., GODBOLE, A. P., SCHAFFNER, A. A., SKOLNICK, A. M. and WALLSTROM, G. L. 1995. Compound Poisson approximations for word patterns under Markovian hypotheses. *J. Appl. Prob.* **32** 877–892.
- GUIBAS, L. J. and ODLYZKO, A. M. 1981. Periods in strings. *J. Combinatorial Theory A* **30** 19–42.
- LOTHAIRE, M. 1983. *Combinatorics on words*. Addison-Wesley.
- LUNDSTROM, R. 1990. Stochastic models and statistical methods for DNA sequence data. Ph.D. Thesis, Department of Mathematics, University of Utah.
- PRUM, B., RODOLPHE, F. and TURCKHEIM, É. DE 1995. Finding words with unexpected frequencies in DNA sequences. *J. R. Statist. Soc. B* **57** 205–220.
- REINERT, G. and Schbath, S. 1998. Compound Poisson and Poisson process approximations for occurrences of multiple words in Markov chains. Preprint.
- ROOS, M. 1994. Stein's method for compound Poisson approximation: the local approach. *Ann. Appl. Prob.* **4** 1177–1187.
- ROOS, M. and STARK, D. 1996. Compound Poisson approximation of the number of visits to a small set in a Markov chain. Preprint.
- SCHBATH, S. 1995a. Compound Poisson approximation of word counts in DNA sequences. *ESAIM: Probability and Statistics* **1** 1–16. (<http://www.emath.fr/ps/>).
- SCHBATH, S. 1995b. *Étude asymptotique du nombre d'occurrences d'un mot dans une chaîne de Markov et application à la recherche de mots de fréquence exceptionnelle dans les séquences d'ADN*. PhD thesis, Université René Descartes, Paris V.
- SCHBATH, S., PRUM, B. and de TURCKHEIM, E. 1995. Exceptional motifs in different Markov chain models for a statistical analysis of DNA sequences. *J. Comp. Biol.* **2** 417–437.
- TANUSHEV, M. 1996. Central limit theorem for several patterns in a Markov chain sequence of letters. Preprint.
- WATERMAN, M. S. 1995. *Introduction to computational biology*. Chapman & Hall.

DERIVING INTERATOMIC DISTANCE BOUNDS FROM CHEMICAL STRUCTURE

BY MICHAEL W. TROSSET¹ AND GEORGE N. PHILLIPS, JR.²

College of William and Mary and Rice University

Structural molecular biology is concerned with determining 3-dimensional representations of molecules. Various computational challenges arise in making such determinations, several of which have attracted some attention in the statistics and numerical optimization communities. One such problem is that of determining a 3-dimensional structure that is consistent with bounds on a molecule's interatomic distances; one source of such bounds is the molecule's chemical structure. Because realistic examples are not readily available to computational scientists hoping to test their algorithms, we provide a detailed description of how plausible bounds can be obtained.

1. Introduction. Knowledge of 3-dimensional molecular structure can be of enormous value in a variety of scientific endeavors. In this report, we assume the importance of such knowledge and focus on one class of mathematical problems that are sometimes solved to obtain it.

The problem that motivates this report is that of calculating 3-dimensional Cartesian coordinates of atoms from information about interatomic distances. Such information usually assumes the form of lower and upper bounds on the distances. This report is concerned with the derivation of such bounds from a molecule's chemical structure, which we assume to be known. It is addressed to computational scientists who are interested in problems that involve determining 3-dimensional molecular structures that are consistent with specified bounds on interatomic distances. These researchers require sample problems on which to test new algorithms. The methods described herein provide an alternative to (1) inventing structures with no chemical plausibility and (2) mastering specialized techniques that require considerable expertise in structural molecular biology.

In Section 2 we introduce some technical notation and provide some relevant background material. In Section 3 we provide detailed descriptions of several useful calculations for inferring lower and upper bounds on interatomic distances from chemical structure. In Section 4 we apply the techniques of Section 3 to

This research was supported by the W. M. Keck Center for Computational Biology under Medical Informatics Training Grant 1T1507093 from the National Library of Medicine.

¹Also supported by Grant DMS-9622749 from the National Science Foundation.

²Also supported by Grant C-1142 from the Robert A. Welch Foundation.

AMS 1991 *subject classifications*. Primary 92E10, 92-08; secondary 62H99, 90C90.

Key words and phrases. Multidimensional scaling, distance geometry, computational biology.

an analogue of the antibiotic trichogin A IV. This is a small ($n = 38$ atoms) molecule on which we have been testing prototype algorithms. In Section 5 we conclude by identifying broader contexts in which having bounds on interatomic distances may be useful.

This report documents details of our research that have not been recorded elsewhere. It also provides other researchers with a plausible set of bounds on which to test their algorithms and—more importantly—with a methodology for generating other plausible data sets. Most of all, we hope that this report will provide computational scientists with a better understanding of the relation between certain mathematical problems and the chemical considerations that motivate them.

2. Background. The ultimate goal of structural molecular biology is to determine the unique 3-dimensional structure into which a given molecule folds. A natural place to begin this determination is with the structural information that is contained in the chemistry of the molecule. One way to represent such information is by lower and upper bounds, implied by the chemical structure, on the interatomic distances. The purpose of this report is to indicate the nature of such bounds; in this section, we establish a context for this representation by sketching how Cartesian coordinates can be extracted from bounds on interpoint distances.

A symmetric matrix $\Delta = (\delta_{ij})$ is a dissimilarity matrix if and only if $\delta_{ij} \geq 0$ and $\delta_{ii} = 0$. A dissimilarity matrix is a p -dimensional Euclidean distance matrix if and only if there exist $x_1, \dots, x_n \in \mathbb{R}^p$ such that $\delta_{ij} = \|x_i - x_j\|$. We denote the closed cone of p -dimensional Euclidean distance matrices by $\mathcal{D}_n(p)$. If $\Delta \in \mathcal{D}_n(p)$, then simple procedures for determining $x_1, \dots, x_n \in \mathbb{R}^p$ such that $\delta_{ij} = \|x_i - x_j\|$ are well-known.

Now suppose that the molecule in question has n atoms. Let $L = (\ell_{ij})$ and $U = (u_{ij})$ be $n \times n$ dissimilarity matrices that contain the specified lower and upper bounds on the interatomic distances. The rectangle $[L, U]$ is defined to be the set of dissimilarity matrices that satisfy the specified bounds, i.e. $\Delta \in [L, U]$ if and only if $\delta_{ij} \in [\ell_{ij}, u_{ij}]$. Glunt, Hayden and Raydan (1993) called this rectangle the *data box*. If $\Delta \in \mathcal{D}_n(3) \cap [L, U]$ and $x_1, \dots, x_n \in \mathbb{R}^3$ is such that $\delta_{ij} = \|x_i - x_j\|$, then x_1, \dots, x_n represent the Cartesian coordinates of a possible 3-dimensional structure of the molecule.

To find a dissimilarity matrix that is (approximately) contained in $\mathcal{D}_n(3) \cap [L, U]$, let ρ denote an error criterion for measuring the discrepancy between a given distance matrix and a given dissimilarity matrix. Then the problem of inferring a possible 3-dimensional structure from the specified interatomic

distance bounds can be formulated as the following optimization problem:

$$(2.1) \quad \text{minimize } \rho(D, \Delta) \text{ subject to } D \in \mathcal{D}_n(3) \text{ and } \Delta \in [L, U].$$

Both the Data Box Algorithm proposed by Glunt, Hayden and Raydan (1993) and the embedding approach described by Trosset (1998) can be derived from special cases of this general formulation.

Problem (2.1) is an example of a problem in *distance geometry*. Other formulations are also possible; a recent example is Moré and Wu (1997). In statistics, techniques for inferring a p -dimensional configuration of points from information about interpoint distances are collectively known as *multidimensional scaling* (MDS). In analogy to Problem (2.1), these techniques can be conceived as algorithms for minimizing some measure of discrepancy between a set of distance matrices and a set of dissimilarity matrices. De Leeuw and Heiser (1982) and Trosset (1997) have surveyed a variety of MDS procedures from this perspective. Trosset (1998) described the relation between Problem (2.1) and nonmetric MDS.

This report is concerned with the reasoning by which a data box is derived from the chemical structure of a molecule. There are two reasons for wanting to make the data box as small as possible. First, we would like to eliminate as many distance matrices as possible in order to narrow the search for the correct interatomic distance matrix. Second, we would like to eliminate as many dissimilarity matrices as possible in order to facilitate solution of Problem (2.1). We now consider how to accomplish these objectives.

3. Bound derivation. We assume that the chemical structure of the molecule is known *a priori*, i.e. we assume knowledge of the atomic bonds within the molecule. To simplify our description of how bounds on interatomic distances can be inferred from such knowledge, we introduce some *ad hoc* notation and terminology.

Suppose that a particular pair of atoms has been specified. We denote each of these atoms by an upper case Roman letter, e.g. C for carbon, N for nitrogen, O for oxygen. If these atoms are bonded together, then they comprise a “1-2” pair, e.g. C-C. If they are not bonded together, but are each bonded to a common atom, which we denote by a lower case Roman letter, then they comprise a “1-3” pair, e.g. C-c-C. In similar fashion, we define and denote “1-4” pairs, “1-5” pairs, etc.

Atoms bond at distances that are approximately fixed by nature. These distances depend on identifiable chemical characteristics, e.g. the types of atoms (carbon, nitrogen, oxygen, etc.), the nature of the bond (covalent, double, etc.), and the type of structure within which the bond occurs (benzyl ring, ester,

peptide, etc.). Evidently, some knowledge of chemistry is required to identify comparable categories. Within such categories, there is a certain amount of apparently random variation in bond lengths for which lower and upper bounds can be obtained by inspecting data banks of structures whose bond lengths are known. Hence, it is essentially an empirical exercise to obtain fairly stringent upper and lower bounds on 1-2 distances.

It is also the case that atoms bond at angles that are approximately fixed by nature. Again, these angles depend on identifiable chemical characteristics. A guiding principle is that the bonds to a common atom will arrange themselves to maximize the minimum angle between any two bonds. For example, if an atom is bonded to four other atoms, then it is easily calculated that the minimum angle is maximized if each angle equals

$$\arccos(-1/3) \doteq 1.91 \doteq 109.5^\circ.$$

Again, within the appropriate categories, there is a certain amount of apparently random variation in bond angles for which lower and upper bounds can be obtained by inspecting data banks of known structures.

If bond lengths and angles are both approximately fixed by nature, then 1-3 pairs are approximately rigid structures and 1-3 distances are approximately fixed. If we denote the bond lengths by a and b and the bond angle by τ , then the 1-3 distance x is given by

$$(3.2) \quad x^2 = a^2 - 2ab \cos(\tau) + b^2.$$

Given lower and upper bounds on a , b and τ , we can use equation (3.2) to derive lower and upper bounds on x . In practice, however, it seems preferable to infer these bounds directly by inspecting the appropriate 1-3 distances in data banks of known structures.

In fact, whenever the molecule contains an approximately rigid structure we eschew trigonometric calculation and directly infer bounds on all of the interatomic distances within that structure by inspecting known cases. For example, benzyl rings are approximately planar, hexagonal structures. The approximate rigidity of a benzyl ring can be described by specifying suitably stringent lower and upper bounds on the 1-2, 1-3 and 1-4 distances. All of these bounds can be obtained by inspecting examples of benzyl rings for which interatomic distances have been determined.

If rigidity cannot be assumed, then trigonometric calculation may be of considerable value. Consider the case of 1-4 distances. Let a , b and c denote the bond lengths; let τ_1 denote the angle between the a and b bonds; and let

τ_2 denote the angle between the b and c bonds. Let d denote the 1-3 distance defined by

$$d^2 = b^2 - 2bc \cos(\tau_2) + c^2$$

and let

$$\phi = \arcsin\left(\frac{c}{d} \sin(\tau_2)\right).$$

Then the 1-4 distance x lies between the lower (+) and upper (-) bounds defined by

$$(3.3) \quad x^2 = a^2 + 2ad \cos(\pi - \tau_1 \pm \phi) + d^2.$$

Given lower and upper bounds on a , b , c , τ_1 and τ_2 , we can use equation (3.3) to derive lower and upper bounds on x .

Equations analogous to (3.3) can be derived for 1-5 distances, 1-6 distances, etc. Such equations, however, are of diminishing utility because the gap between the lower and upper bounds rapidly widens as the number of intervening bonds increases. It is easier (and often sharper) to impose lower bounds that approximate the repulsive forces that keep unbonded atoms apart. Reasonable upper bounds can be deduced by applying the triangle inequality to the upper bounds for the 1-2, 1-3 and 1-4 distances, an elementary example of *bound smoothing*. See Sections 5.3 (Triangle Inequality Bound Smoothing) and 5.4 (Tetrangle Inequality Bound Smoothing) of Crippen and Havel (1988) for an introduction to bound smoothing.

4. Example. We now apply the methods of Section 3 to a small molecule described by Crisma et al. (1994). This molecule, an analogue of the antibiotic trichogin A IV, contains 7 oxygen atoms, 4 nitrogen atoms, and 27 carbon atoms. Its chemical structure is diagrammed in Figure 1. Our goal is to infer plausible lower and upper bounds on the $703 = 38 \cdot 37/2$ interatomic distances associated with these $n = 38$ atoms.

The 3-dimensional structure of the molecule in Figure 1 is known. One coordinatization, measured in angstroms, is presented in Table 1. These coordinates were obtained from the Brookhaven Protein Data Bank (Bernstein et al., 1977), which can be accessed electronically at the web site <http://www.pdb.bnl.gov/>.

So that our example be self-contained, we will exploit replication within the molecule itself instead of replication in an external database of known molecules. This will cause us to underestimate the natural variability of the component structures, but the resulting bounds will be sufficiently plausible to be illustrative. The purpose of this report is to explicate the logic of how bounds are

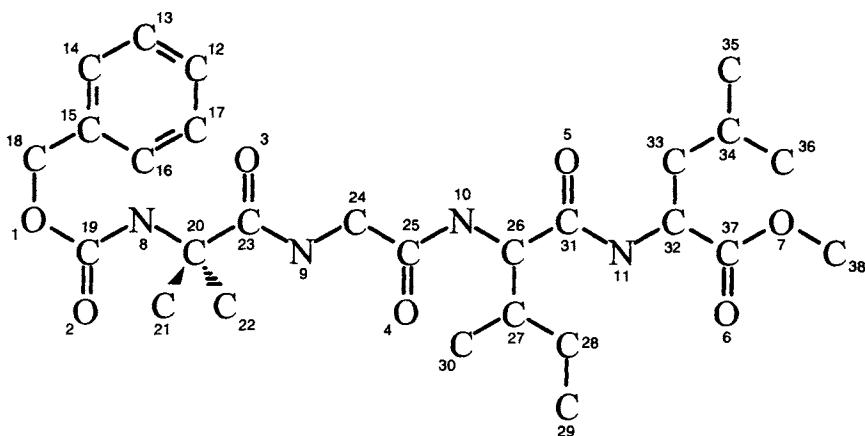


FIG. 1. *Chemical structure of an analogue of the antibiotic molecule trichogin A IV.*

derived and to reveal some of the qualitative features of a plausible data box, not to construct the most realistic data box imaginable for the molecule in question. We believe that using even crude data boxes constructed along the lines indicated in this report will advance the present state of numerical experimentation.

We illustrate the derivation of lower and upper bounds using a 10-atom fragment of the molecule in Figure 1. The fragment that we consider comprises a benzyl ring (atoms 12–17) and an ester (atoms 18, 1, 19, 2).

We begin by examining the benzyl ring. The ring contains six 1-2 pairs, six 1-3 pairs, and three 1-4 pairs. The corresponding distances and bounds that include them are reported in Table 2. The 1-2 distances are listed clockwise from the 17-16 pair; the 1-3 distances are listed clockwise from the 17-16 pair; and the 1-4 distances are listed clockwise from the 17-14 pair.

Atom 18 is tetrahedral, i.e. it has bonds to four other atoms. (The bonds to two hydrogen atoms are not depicted in Figure 1.) Although the 18-15 bond is not replicated in the molecule, there are fifteen C-C bonds in which one (or both) of the atoms is tetrahedral. These 1-2 distances range from a minimum of 1.4687 to a maximum of 1.5456, so we adopt bounds of 1.46 and 1.55.

Because the benzyl ring is (approximately) rigid, each distance between atom 18 and an atom in the ring is (approximately) fixed. For the 18-16 and 18-14 distances we observe that

$$2.47 \leq 2.4769, 2.5212 \leq 2.53;$$

and for the 18-17 and 18-13 distances we observe that

$$3.76 \leq 3.7774, 3.7646 \leq 3.78.$$

TABLE 1
Atomic coordinates of an analogue of the antibiotic molecule trichogin A IV.

Atom	Coordinates			Atom	Coordinates		
1 O	1.629	1.310	4.250	20 C	3.277	4.529	4.289
2 O	2.633	2.357	2.508	21 C	2.422	5.328	3.320
3 O	5.199	5.284	3.098	22 C	3.418	5.277	5.619
4 O	7.777	2.147	1.545	23 C	4.682	4.358	3.714
5 O	5.088	2.230	-1.834	24 C	6.656	2.963	3.451
6 O	2.699	5.605	-1.867	25 C	6.701	2.307	2.089
7 O	3.967	5.011	-0.154	26 C	5.407	1.189	0.302
8 N	2.638	3.244	4.608	27 C	4.771	-0.186	0.520
9 N	5.327	3.216	3.970	28 C	5.660	-1.039	1.421
10 N	5.537	1.894	1.579	29 C	5.081	-2.325	1.831
11 N	3.443	2.477	-0.327	30 C	4.525	-0.901	-0.828
12 C	0.121	-3.130	5.716	31 C	4.643	2.031	-0.701
13 C	1.418	-2.846	5.531	32 C	2.627	3.347	-1.139
14 C	1.777	-1.783	4.797	33 C	1.163	3.290	-0.715
15 C	0.873	-0.950	4.212	34 C	0.576	1.866	-0.753
16 C	-0.453	-1.221	4.392	35 C	-0.919	2.004	-0.393
17 C	-0.830	-2.361	5.137	36 C	0.784	1.116	-2.012
18 C	1.272	0.224	3.357	37 C	3.094	4.785	-1.106
19 C	2.320	2.319	3.694	38 C	4.481	6.369	-0.016

The 18-12 pair is not replicated, but we might guess that

$$4.24 \leq 4.2590 \leq 4.28$$

are plausible bounds for our purposes.

Now we consider the ester C-O-C=O. This nonrigid structure appears twice: 18-1-19-2 and 38-7-37-6. Distances and bounds for the 1-2 and 1-3 pairs of atoms in this structure are reported in Table 3. Because the structure is not rigid, the 1-4 pair is omitted from Table 3. and bounds are obtained from equation (3.3). The bond angles observed for the C-O-C ester fragment are 115.7° and 117.6°,

TABLE 2
Distances between atoms in the benzyl ring.

	1-2 pairs	1-3 pairs	1-4 pairs
1.4131	2.3972	2.6919	
1.3653	2.3351	2.7274	
1.3614	2.3731	2.7532	
1.3407	2.3241		
1.3406	2.3332		
1.3531	2.3931		
Lower bound	1.34	2.32	2.69
Upper bound	1.42	2.40	2.76

so we adopt bounds of 115° and 118° . The bond angles observed for the O-C=O ester fragment are 123.7° and 124.6° , so we adopt bounds of 123° and 125° . Using $a = 1.45$, $\tau_1 = 115^\circ$, $b = 1.31$, $\tau_2 = 123^\circ$ and $c = 1.18$, we obtain a lower bound on the 1-4 distance of 2.58. Using $a = 1.46$, $\tau_1 = 118^\circ$, $b = 1.35$, $\tau_2 = 125^\circ$ and $c = 1.23$, we obtain an upper bound on the 1-4 distance of 3.58.

TABLE 3
Distances between atoms in the C-O-C=O esters.

	1-2 pairs			1-3 pairs	
	C-O-c=o	c-O-C=o	c-o-C=O	C-o-C=o	c-O-c=o
1.4506	1.3434	1.2272	2.3666	2.2669	
1.4586	1.3113	1.1864	2.3708	2.2125	
Lower bound	1.45	1.31	1.18	2.36	2.21
Upper bound	1.46	1.35	1.23	2.38	2.27

Because atom 18 is a tetrahedral carbon, we can infer bounds on the angle between the 18-15 and 18-1 bonds by examining the twenty-two instances of such angles that occur in the molecule. They range from a minimum of 106.2° to a maximum of 115.8° , so we adopt bounds of 106° and 116° . Combining these bounds with the bounds previously obtained for the 18-15 and 18-1 distances, we can exploit equation (3.2) to obtain lower and upper bounds on the 15-1 distance. Using $a = 1.46$, $\tau = 106^\circ$ and $b = 1.45$, we obtain a lower bound of 2.32. Using $a = 1.55$, $\tau = 116^\circ$ and $b = 1.46$, we obtain an upper bound of 2.56. Similarly, we can exploit equation (3.3) to obtain lower and upper bounds on the 15-19 distance. Using $a = 1.46$, $\tau_1 = 106^\circ$, $b = 1.45$, $\tau_2 = 115^\circ$ and $c = 1.31$, we obtain a lower bound of 2.41. Using $a = 1.55$, $\tau_1 = 116^\circ$, $b = 1.46$, $\tau_2 = 118^\circ$ and $c = 1.35$, we obtain an upper bound of 3.80.

The angle between the 15-16 and 15-18 bonds is 119.1° and the angle between the 15-16 and 15-18 bonds is 123.0° , so we adopt bounds on these angles of 119° and 124° . We can now exploit equation (3.3) to obtain lower and upper bounds on the 16-1 and 14-1 distances. Using $a = 1.34$, $\tau_1 = 119^\circ$, $b = 1.46$, $\tau_2 = 106^\circ$ and $c = 1.45$, we obtain a lower bound of 2.51. Using $a = 1.42$, $\tau_1 = 124^\circ$, $b = 1.55$, $\tau_2 = 116^\circ$ and $c = 1.46$, we obtain an upper bound of 3.89.

The remaining lower bounds in the fragment approximate the repulsive forces that keep unbonded atoms apart. For N-O pairs, we suggest a lower bound of 2.80; for other pairs, we suggest a lower bound of 3.20.

Finally, the remaining upper bounds in the fragment are obtained by applying the triangle inequality to the upper bounds that we have already derived. This yields upper bounds of 5.32 for the 12-1 distance, 6.66 for the 12-19 distance, 7.59 for the 12-2 distance, 4.96 for the 17-1 and 13-1 distances, 6.16 for the 17-19 and 13-19 distances, 7.23 for the 17-2 and 13-2 distances, 4.91 for the

16-19 and 14-19 distances, 6.11 for the 16-2 and 14-2 distances, and 4.83 for the 15-2 distance.

We have derived lower and upper bounds for each of the $45 = 10 \cdot 9/2$ interatomic distances associated with a 10-atom fragment of a 38-atom molecule. The same techniques can be applied to the entire molecule. This methodology allows us to approximate plausible *a priori* lower and upper bounds on each of the molecule's 703 interatomic distances. A file containing a set of such bounds is available from the first author.

5. Discussion. The bounds derived in Section 4 define a rectangular feasible region whose features typify the feasible regions $[L, U]$ for Problem (2.1) that might actually arise in practice. As such, our derivations not only provide a useful example on which to test numerical algorithms, but also contribute to our understanding of what is involved in inferring 3-dimensional structure from information about interatomic distances.

We note that most of the distance information that can be derived from chemical structure pertains to atoms that are separated by a small number of bonds. It should be emphasized that this information will rarely—if ever—suffice to determine a unique 3-dimensional structure. Hence, solving Problem (2.1) will not produce the unique 3-dimensional structure that the molecule actually assumes, only one of many structures that are consistent with the specified bounds on the interatomic distances. Stated differently, we do not know how to infer Table 1 from Figure 1.

Of course, as remarked in Section 2, the ultimate goal of structural molecular biology is to find the solution of Problem (2.1) that corresponds to the actual 3-dimensional structure of the specified molecule. To do so, it is necessary to consider additional information about the molecule. We conclude this report by sketching some of the problems that arise in this manner.

5.1. NMR spectroscopy. Distances between nearby hydrogen atoms can at times be measured experimentally by Nuclear Magnetic Resonance (NMR) spectroscopy. Bounds on the measurement error resulting from this procedure can be interpreted as bounds on the interatomic distances themselves. This observation was the motivation for the Data Box Algorithm proposed by Glunt, Hayden, and Raydan (1993).

If we combine the distance bounds derived from chemical structure with the distance bounds determined by NMR spectroscopy, then we must again solve Problem (2.1). However, incorporating additional bounds restricts the original feasible region. Usually, the new feasible region is sufficiently restricted that, by solving Problem (2.1) repeatedly, one can obtain an ensemble of possible molec-

ular structures with common features of interest. General reviews of distance geometry in the context of using NMR spectroscopy to determine the structures of protein molecules have been provided by Havel and Wüthrich (1985), Braun (1987), Crippen and Havel (1988), Kuntz, Thomason and Oshiro (1989), Havel (1991), Brünger and Nilges (1993), and Havel, Hyberts and Naifeld (1997).

5.2. Protein folding. An alternative possibility is to introduce a function that quantifies Nature's preferences for certain molecular structures, thereby allowing us to anticipate which solutions of Problem (2.1) are most likely to occur. This possibility motivates us to consider the protein folding problem, which is to predict the 3-dimensional structure of a protein molecule from its amino acid sequence.

A self-contained introduction to the protein folding problem, together with many references, was recently provided by Neumaier [15]. Realistic models of protein folding, e.g. the CHARMM [3] potential, include terms corresponding to both bonded and unbonded interactions, the latter comprising both the long-range, slowly decaying electrostatic (Coulomb) interaction and the short-range, fast-decaying van der Waals interaction. Thus, a second possibility is to search for a unique 3-dimensional structure by minimizing a theoretical objective function subject to constraints imposed by the data box derived for Problem (2.1). It should be noted, however, that finding global solutions of such problems is an extremely difficult task.

5.3. X-ray crystallography. If the specified molecule can be crystallized, then a third possibility is to utilize data about its diffraction pattern obtained from a crystallography experiment. A general introduction to X-ray crystallography was provided by Glusker and Trueblood (1985); a more recent survey of the computational challenges associated with the phasing, model building, and refinement stages was provided by Brünger and Nilges (1993).

Mathematically, this possibility is similar to the preceding one in that the objective function for protein folding is replaced with a criterion for measuring the fit between the theoretical diffraction pattern of a 3-dimensional structure and the observed diffraction pattern of the molecule in question. Again, it should be noted that finding global solutions of such problems is an extremely difficult task.

Finally, we observe that there may be an interesting role for probability and statistics to play in solving the above problems. In each approach that we have described, it is necessary to sample from the data box. This is necessary in the first case in order to obtain a meaningful ensemble of possible structures; it is necessary in all three cases in order to search for global solutions for optimization

problems that are typically plagued with myriad nonglobal solutions. To the extent that additional "prior" information can be represented in the form of a probability distribution from which dissimilarity matrices in the data box are drawn, it may be possible to accelerate the search for meaningful 3-dimensional structures.

REFERENCES

- BERNSTEIN, F. C., KOETZLE, T. F., WILLIAMS, G. J. B., MEYER, E., BRYCE, M. D., ROGERS, J. R., KENNARD, O., SHIKANOUCHI, T. and TASUMI, M. (1977). The protein data bank: A computer-based archival file for macromolecular structures. *Journal of Molecular Biology* **112** 535–542.
- BRAUN, W. (1987). Distance geometry and related methods for protein structure determination from NMR data. *Quarterly Reviews of Biophysics* **19** 115–157.
- BROOKS, B. R., BRUCCOLERI, R., OLAFSON, B., STATES, D., SWAMINATHAN, S. and KARPLUS, M. (1983). CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry* **4** 187–217.
- BRÜNGER, A. T. and NILGES, M. (1993). Computational challenges for macromolecular structure determination by X-ray crystallography and solution NMR spectroscopy. *Quarterly Reviews of Biophysics* **26** 49–125.
- CRIPPEN, G. M. and HAVEL, T. F. (1988). *Distance Geometry and Molecular Conformation*. John Wiley & Sons, New York.
- CRISMA, M., VALLE, G., MONACO, V., FORMAGGIO, F. and TONILO, C. (1994). N alpha-benzyloxycarbonyl-alpha-aminoisobutyryl-glycyl-l-isoleucyl-l-leucine methyl ester monohydrate. *Acta Crystallographica* **50** 563–565.
- DE LEEUW, J. and HEISER, W. (1982). Theory of multidimensional scaling. In Krishnaiah, P. R. and Kanal, I. N., editors, *Handbook of Statistics*, volume 2, chapter 13, pages 285–316. North-Holland Publishing Company, Amsterdam.
- GLUNT, W., HAYDEN, T. L. and RAYDAN, M. (1993). Molecular conformations from distance matrices. *Journal of Computational Chemistry* **14** 114–120.
- GLUSKER, J. P. and TRUEBLOOD, K. N. (1985). *Crystal Structure Analysis: A Primer*. Oxford University Press, New York.
- HAVEL, T. F. (1991). An evaluation of computational strategies for use in the determination of protein structure from distance constraints obtained by nuclear magnetic resonance. *Progress in Biophysics and Molecular Biology* **56** 43–78.
- HAVEL, T. F., HYBERTS, S. and NAIFIELD, I. (1997). Recent advances in molecular distance geometry. In Hofestödt, R., editor, *Bioinformatics*, pages 62–71. Springer, Berlin. *Lecture Notes in Computer Science*, Number 1278.
- HAVEL, T. F. and WÜTHRICH, K. (1985). An evaluation of the combined use of nuclear magnetic resonance and distance geometry for the determination of protein conformations in solution. *Journal of Molecular Biology* **182** 281–294.
- KUNTZ, I. D., THOMASON, J. F. and OSHIRO, C. M. (1989). Distance geometry. In Oppenheimer, N. J. and James, T. L., editors, *Nuclear Magnetic Resonance, Part B: Structure and Mechanism*, pages 159–204. Academic Press, New York. Volume 177 of *Methods in Enzymology*.
- MORÉ, J. J. and WU, Z. (1997). Global continuation for distance geometry problems. *SIAM Journal on Optimization* **7** 814–836.
- NEUMAIER, A. (1997). Molecular modeling of proteins and mathematical prediction of protein structure. *SIAM Review* **39** 407–460.
- TROSSET, M. W. (1997). Numerical algorithms for multidimensional scaling. In Klar, R. and Opitz, P., editors, *Classification and Knowledge Organization*, pages 80–92, Berlin. Springer. Proceedings of the 20th annual conference of the Gesellschaft für Klassifikation e.V., held March

6–8, 1996, in Freiburg, Germany.

TROSSET, M. W. (1998). Applications of multidimensional scaling to molecular conformation. *Computing Science and Statistics* **29**. To appear.

MICHAEL W. TROSSET
DEPARTMENT OF MATHEMATICS
COLLEGE OF WILLIAM AND MARY
P.O. Box 8795
WILLIAMSBURG, VA 23187-8795
TROSSET@MATH.WM.EDU

GEORGE N. PHILLIPS, JR.
DEPARTMENT OF BIOCHEMISTRY AND CELL BIOLOGY
RICE UNIVERSITY—MS 140
HOUSTON, TX 77005-1892
GEORGE@BIOC.RICE.EDU

PROTEIN FOLD CLASS PREDICTION IS A NEW FIELD FOR STATISTICAL CLASSIFICATION AND REGRESSION

BY LUTZ EDLER AND JANET GRASSMANN

*Biostatistics Unit, Research Program on Genome Research and Bioinformatics,
German Cancer Research Center, Heidelberg, Germany and BRAIN²,
Germany*

Protein structure classification and prediction is introduced and elaborated for the application of standard and new statistical classification, discrimination and regression methods. With the sequence to structure to function paradigm in the background, methods of secondary and tertiary structure prediction will be reviewed and super-secondary classes and of fold classes will be defined. We apply two branches of statistical classification - methods based on posterior probabilities and methods based on class conditional probabilities - and we will explore the role of artificial neural networks for the protein structure prediction. The procedures will be applied to a set of 268 previously described protein sequences for their statistical performance in the prediction of the four super-secondary classes and also in the prediction of 42 fold structure classes.

1. Introduction. Knowledge of the three-dimensional (3D) structure of a protein is essential for describing and understanding its function and for its use in molecular modeling [Fasman (1989)]. The impact of the structural knowledge for medical interventions and the understanding of diseases and their evolution has been clearly demonstrated [Branden and Tooze (1991), Giersch and King (1990)]. Knowledge of the 3D structure of hemoglobin [e.g. Perutz (1978) Dickerson and Geis (1983)] enabled researchers to increase its oxygen capacity. This was the first and crucial step of a development which resulted in a synthetic hemoglobin substitute with consequences for blood transfusion [Mickler and Longnecker (1992)]. On the other hand, sickle cell anemia is caused by a single mutation in the amino acid sequence of hemoglobin, a change from *Glu* to *Val* on the surface of the globin fold [see Branden and Tooze (1991) p. 39] which causes movements of the α -helices relative to each other and makes the cell membrane more permeable to potassium ions. The disease is lethal for homozygotes, but increases the resistance to malaria in heterozygotes by killing the parasite through the drop of the potassium ion concentration. Therefore, the determination of structure is useful in different aspects: altering an existing protein's function (protein engineering), creating a protein *de novo* (protein de-

AMS 1991 subject classifications. Primary 92A10; secondary 62H30.

Key words and phrases. Protein structure, fold class, classification and prediction, discriminants, regression, neural networks, cross-validation.

sign), or understanding the evolution of diseases. Understanding and predicting how sequence information translates into 3D structure and folding of the then biologically active protein (functional properties) has become one of the most challenging problems in current molecular biology [Sternberg (1996)]. A solution of the protein folding problem would have great implications on interpreting sequence data as those created by the Human Genome Project. It would improve gene function analysis with implications on understanding hereditary genetics and diseases and it would provide clues for drug design and biological engines with considerable commercial consequences (see e.g. Cambridge Healthtech Institute: <http://www.healthtech.com/>).

The three-dimensional structure of a protein is determined physically by the 3D coordinates of all atoms of the protein. It is mostly obtained by x-ray crystallography and for smaller proteins also by NMR (nuclear magnetic resonance), see Branden and Tooze (1991). This determination has been achieved at present only for a small percentage of known proteins. On the other hand, the extraordinary improvement of the efficiency of modern sequencing techniques creates a large gap between the number of sequenced proteins and the number of structurally 'explained' proteins. Figure 1 shows the sharp increase of the number of entries of protein sequences and protein domains in the SWISSPROT data base compared to the much tardier increase of proteins of known 3D structure in the PDB data base [Bernstein et al. (1977), Bairoch and Apweiler (1997), Benson et al. (1997)]. Since the Human Genome Project will generate an enormous amount of protein sequences more over the next few years [Rowen et al. (1997)] this gap will increase rapidly. The need to bridge the gap has called for biochemical and biophysical methods for the determination of the 3D structure which circumvent x-ray and NMR and use the basic sequence and physical properties of the building blocks of proteins. The protein fold problem poses itself then as the question [Richards (1991)]: How to predict the 3D structure of a protein from its amino acid sequence? This question had been around in protein research since the seminal proposition of Anfinsen (1961) to predict the conformation of a protein on the basis of its linear amino acid sequence. From there originates the hypothesis that the sequence of amino acids of a protein is necessary and sufficient for the determination of the 3D structure and, consequently, for its function. Almost 40 years later this problem is still in the center of theoretical and practical biotechnological protein research. Although unsolved in its original sense, the question continues to elicit important research results of structural biology and molecular modeling.

Anfinsen's hypothesis suggests that the amino acid sequence together with physical and chemical principles should suffice to determine the forces responsible for the folding and determination of the ultimate 3D structure. One approach

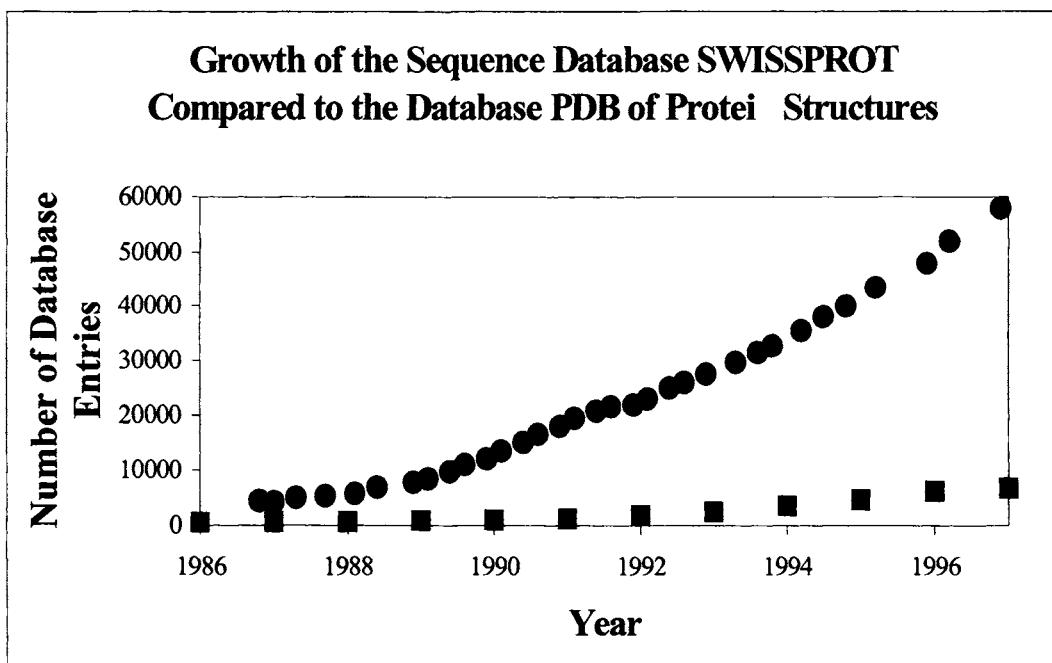


FIG. 1. Increase of the number of entries of sequences of proteins and protein domain in SWISSPROT data base and of the sequences of known 3D structure in the PDB data bank.

has been the *ab initio* calculations which combine biophysical, biochemical, and quantum mechanical calculations with molecular modeling in order to determine the native protein and its 3D structure from the denatured and unfolded protein [Dinur and Hagler (1991), Defay and Cohen (1995)] by simulating dynamically nature's folding pathways [Creighton (1984, 1990)]. An increasing number of proteins could be constructed this way (see e.g. recent issues of the journal *Protein Engineering*) but the method is still far from being applicable to large proteins or to be used generally.

A second approach formulates the protein folding problem in terms of mathematical physics. This point of view has been reviewed excellently by Neumaier (1997) who provides also a survey of most of the relevant literature on chemical structure and local geometry of a protein and molecular mechanics. Focussing on mathematical models and molecular dynamics, quantitative 3D prediction applies (global) optimization of the potential-energy functions, directly making use of the physical forces between the atoms in the poly-peptide sequence of the protein. This method is computationally intensive and can succeed only if it's able to find the global energy minimum which determines the ultimate 3D structure. The limitations of this method are in the computational complexity combined with the intrinsic problem to avoid local energy minima. A

third approach evolved recently from machine learning and artificial intelligence and claims to predict structural classes of proteins from the basic amino acid sequence and features derived from that [Holley and Karplus (1989), Lambert and Sheraga (1989), Friedrichs et al. (1991), Taylor (1992)]. Although success was rather limited [Schulz (1988)] this approach is appealing because of its new information theoretical access to the problem. It has attracted computer scientists, biomolecular modelers, bioengineers and biophysicists to use especially artificial neural networks (ANN) for the classification and prediction. The goals were lower in this case: not the quantitative prediction of the ultimate 3D structure but the prediction of the qualitative appertaining to a limited number of folding categories was searched. In principle, all three approaches could not reach their goal because of the problem's complexity which translates almost immediately into computing complexity as e.g. when biophysical methodology is combined with modern computer algorithms or when all possible configurations are screened for those with low energy [Neumaier (1997)]. Therefore, the search for better prediction methods is ongoing [see Defay and Cohen (1996) and the endeavors of the Ansilomar Conference from 1994].

So far, this type of protein prediction was without much interaction with statistics, although standard statistical methods such as regression, discriminant analysis, and cluster analysis have been applied to a variety of prediction problems with considerable success. To our knowledge, traditional statistical techniques have not been applied systematically to the protein folding problem. The protein structure prediction problem is usually complicated. This fact may have deterred researchers from using standard statistical methods to predict protein structure. However, if statistical methods could be applied to this problem it would be very interesting to compare their performance with that of machine learning and artificial neural networks. More attention could then be given to the assumptions of the procedures, the sampling of the data and the realism of the error probabilities and the prediction accuracy. By this work we want to add statistical classification and statistical methodology of pattern recognition to the toolbox for predicting protein structure directly from the sequence information, and we want to initiate the exchange and transfer of methods between the disciplines of protein research and applied statistics.

In principle, biophysical methodology should be able to define the unique structure of a protein from the atomic structure of the amino acids in the sequence. At present, it fails with the complexity of the calculations. Therefore, it may be a reasonable strategy at this stage to develop methods that exploit both types of information, the biophysical information from molecular mechanics and the statistical information from classification and regression on sequences. We can not be sure that this combination will lead soon to a breakthrough for this

knotty problem but it will require innovative collaboration between molecular biologists and statisticians. Interaction of the two disciplines could push the problem one step further to its solution. The following methods and ideas represent a statistical contribution to such a collaboration. In other words, the aim of this paper is to introduce statisticians to the subject, to provide some background of the problem and to improve the interaction between statisticians and computer scientists for approaching the statistical problem of classification and prediction of protein fold class.

Below, we will provide a short introduction to protein structure prediction and to some of the results achieved. We will introduce in the next section the biological problem of fold class prediction and review shortly previous methods of fold classification and prediction. This will comprise the data structure of the amino acid sequence and the definition of the statistical classification and decision problem such that it becomes amenable to regression, discrimination and classification methods. Section 2.2 provides a short review of the methods used for secondary structure prediction and their achievements. Based on the secondary structure we define the four supersecondary classes which have to be predicted by the statistical methods we will introduce later. Section 2.3 introduces tertiary structure and fold prediction. Emphasis is given to the statistical content of the methods used previously by protein researchers. In Section 2.4 we will present standard and new regression and discriminant methods to be used competitively for prediction including the neural nets. These methods are applied in our case study of a data set of 268 protein sequences [Grassmann et al. (1998)] described in 2.5 which we investigated recently in collaboration with colleagues from the biophysics discipline in our research program. Selected results of the fitting of the models to these protein data will be given in Section 3 and discussed in Section 4.

2. Methods.

2.1. *Proteins and their classification.* From a chemical point of view, a protein is a polymer or polypeptide consisting of a long chain of amino acids linked by peptide bonds to a one-dimensional directed polypeptide chain, see Brandon and Tooze (1991) for an illustrative introduction. Important for the 3D geometry is that fact that each amino acid in the sequence contains a central carbon atom (C_α atom) and that each amino acid is characterized by its side chain (residue) attached to the C_α atom. Interatomic forces bend and twist the protein into a characteristic 3D folded state. The sequence of the C_α atoms represents the so-called backbone of the protein. Their three-dimensional coordinates represent the genuine 3D structure. For the local geometry and all other

details see Neumaier (1997).

Nature has provided twenty amino acids; see Brandon and Tooze (1991) or any standard monograph on molecular biology for a listing and characterization. A *protein of length N* is then formally represented as an ordered sequence

$$P = (s_1, \dots, s_N)$$

with elements s_i from the finite set $A = \{A_1, \dots, A_{20}\}$. The length of proteins varies considerably between the tens and a few thousands. An average sized protein has a length of 100-200 amino acids. For combinatorial reasons the number of possible proteins is therefore huge: Given an averaged sized protein of 150 amino acids, the number of possible sequences would be $20^{150} \approx 10^{200}$. At present, the number of proteins existing in living nature is not known. Based on a number of assumptions, Zhang (1997) estimated the number of human proteins roughly to $5 - 10 \times 10^5$. Similarities of the shape and functional similarities of proteins motivated researchers to define *structural classes* for proteins, say classes C_1, \dots, C_K . The largest set of structural classes would be obtained by the complete description of a protein by all C_α atoms and their 3D coordinates. Although, the determination of the full set of 3D coordinates of the C_α atoms is an important task performed in crystallography [Zanotti (1992)] and NMR spectroscopy [Torda and van Gunsteren (1992)] broader classifications lead to structural families. Holm and Sander (1994) describe e.g. 270 fold classes for 838 families with a class occupancy ranging between 1 and 73. Using some type of sampling statistics and empirical data, Wang (1998) estimates a total number of 1150 protein superfamilies and about 650 protein folds to exist in nature. That means we are dealing roughly with a number of sequences of the order of magnitude of perhaps $10^5 - 10^6$ to be classified into about $10^2 - 10^3$ classes, given all proteins have been once sequenced.

The definition of structural classes of proteins started with the identification of local structure in the primary amino acid sequence. There are only three types of so-called *secondary structures*: α -helices, β -sheets and coils (γ). For illustrative details see Branden and Tooze (1991) and Fetrow et al. (1997). Secondary elements can combine with each other to form **motifs** or so-called **super-secondary structures** which finally assemble globally and form the tertiary structure. Classification and prediction of secondary structure is considered also as intermediate step to tertiary structure [Stolorz et al. (1992)]. A very simple global classification is obtained by characterizing the protein by the presence or absence of α -helices and β -sheets. This results in four *super-secondary classes (SSC)*: only α , only β , one part α plus one part β ($\alpha + \beta$), and α and β alternating (α/β). Based on topological similarity of the backbone, a definition of 38 fold classes has been proposed by Pascarella and Argos (1992)

and recently enlarged to 42 fold classes by Reczko et al. (1994). That set of classes was later enlarged to 45 classes [Reczko and Bohr (1994)] and further to 49 classes by Reczko et al. (1997).

The raw information given by the protein sequence $P = (s_i, i = 1, \dots, N)$ with elements from $A = \{A_1, \dots, A_{20}\}$ is usually reduced and restructured for protein classification and prediction such that each protein is represented by an element of a suitable predictor space (feature space) X . A very simple example is the space of the frequency distributions of amino acids, the 20-dimensional unit cube $X = [0, 1]^{20}$ where each protein is represented by a vector (f_1, \dots, f_{20}) of the relative frequencies of the 20 amino acids in its sequence. Other feature spaces have been constructed by some type of ‘reading’ information. There, a moving window $x_j^{(a)} = (s_j, s_{j+1}, \dots, s_{j+a-1})$ say of length a , is gliding along the protein sequence (Figure 2). The sample $x_j^{(a)}, j = 1, \dots, N$ represents the sequence $P = (s_1, \dots, s_N)$ and a suitable feature space X is e.g. the space of the frequency distribution of all a -tuples with elements of A . Special cases are the dipeptides obtained for $a = 2$ which give raise to a 400 dimensional space of dipeptide frequencies. We consider in the following the complete protein as sampling unit. In some cases, sequences of sub-domains of proteins or motifs were treated as independent samples even if they originated from the same protein. Two *feature spaces* X will be used in our analysis: Firstly, the space of the amino acid frequencies (AAF)

$$X_1 : x = (f_1, \dots, f_n)$$

where f_i denotes the relative frequency of the amino acid A_i in the sequence. Secondly, the space of the dipeptide matrices (DPF)

$$X_2 : x = (f_{1,1}, \dots, f_{1,20}, f_{2,1}, \dots, f_{2,20}, \dots, f_{20,1}, \dots, f_{20,20})$$

where $f_{i,k}$ denotes the relative frequency of the amino acid pairs (A_i, A_k) in an ordered sequence of residues (the case of a moving window of length $a = 2$).

A *classification / prediction rule* R is a rule which maps the feature information $x \in X$ of each primary sequence $P = (s_1, \dots, s_N)$ into a finite set of structural classes $C = C_1, \dots, C_K$. For convenience, the structural classes C_k are represented by the consecutive numbers $\{1, \dots, K\}$. From a statistical point of view the protein structure prediction is nothing more than a prediction of an element of a finite set of structural classes based on information $x \in X$ where X is an Euclidean Space, e.g. a subset of $\mathbb{R}^n, n > 1$. However, the space X is not straightforwardly given in practice. There are many options to define X , see above, such that it represents relevant information and is still dimensionally tractable. When using moving windows $x_j^{(a)}$, the window size a can be chosen

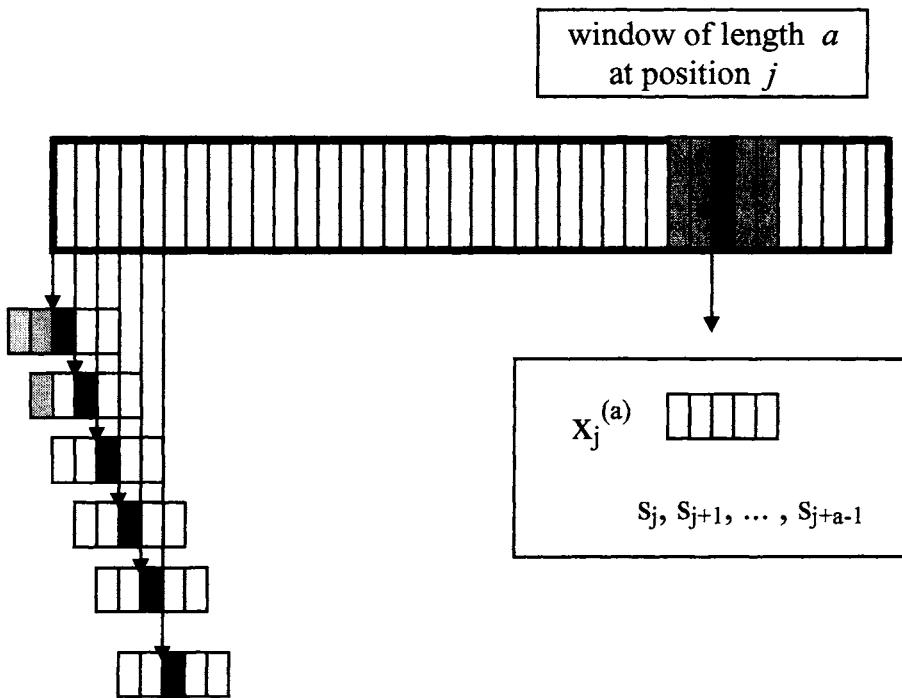


FIG. 2. *Moving window information read from the protein sequence. A moving window of length 5 glides in this example from the left to the right. At the start (and at the end, respectively) the segments have to be augmented by spacers.*

depending how much neighborhood information is thought to be useful. Given only a limited number of observations available in practice, one has to account for the sparseness of data in X .

2.2. Secondary structure prediction. Because of its direct connection with the SSC classification used below and because of the wealth of previously obtained results we address here secondary structure prediction for further illustration. The secondary structure of the amino acid sequence is defined as a local property and induces an one-to-one mapping

$$R_{SEC} : P = (s_1, \dots, s_N) \longrightarrow Q = (r_1, \dots, r_N)$$

from the set of all possible sequences $\{(s_1, \dots, s_n) : s_i \in A, i = 1, \dots, n, n = 1, 2, \dots\}$ to the set $\{(r_1, \dots, r_n) : r_i \in \{\alpha, \beta, \gamma\}, i = 1, \dots, n, n = 1, 2, \dots\}$. Each element of P is mapped onto exactly one element from $\{\alpha, \beta, \gamma\}$. The rule R_{SEC} assigns to a protein its secondary structure Q as an estimate $\hat{Q} = (\hat{r}_1, \dots, \hat{r}_N)$ depending on a sample of n known pairs $(P_j, Q_j) j = 1, \dots, n$, of proteins and structures.

Let us use this framework to consider the definition of a classification error. The naive *per protein error rate* is the ratio w/N of the number w of incorrectly assigned secondary classes of the residues of that protein P of length N . This error is further differentiated with respect to the three secondary types. This yields a misclassification table or confusion matrix [Ripley (1996) Chap. 2.7]:

		True Class			
		α	β	γ	
Assigned Class	α	$w_{\alpha\alpha}$	$w_{\alpha\beta}$	$w_{\alpha\gamma}$	\hat{N}_α
	β	$w_{\beta\alpha}$	$w_{\beta\beta}$	$w_{\beta\gamma}$	\hat{N}_β
	γ	$w_{\gamma\alpha}$	$w_{\gamma\beta}$	$w_{\gamma\gamma}$	\hat{N}_γ
		N_α	N_β	N_γ	N

The diagonal contains the number of correctly and the off-diagonal contains the numbers of incorrectly classified amino acids. Therefore,

$$e_i = (N_i - w_{ii})/N_i$$

is the *structure specific error rate*, $i = 1, 2, 3$, and

$$e = \left(N - \sum_{i=1}^3 w_{ii} \right) / N$$

the *total error rate*, usually denoted by Q_3 in secondary structure prediction. If classification or prediction is performed for a set of proteins one calculates an *overall structure specific error rate* and *overall total error rate* by pooling all residues of all the proteins. Another measure of discordance is the so-called *Matthew correlation* [Matthew (1975)] defined as

$$MC_i = \frac{w_{ii} \left(\sum_{j,k \neq i} w_{jk} \right) - \left(\sum_{j \neq i} w_{ij} \right) \left(\sum_{j \neq i} w_{ji} \right)}{\hat{N}_i \left(\sum_{j \neq i} \hat{N}_j \right) N_i \left(\sum_{j \neq i} N_j \right)}$$

for $i = 1, 2, 3$. MC_i is up to the factors N_i the square root of the chi-square statistic for the classification into the i -th secondary structure category if the data are organized as a four fold table with the numbers w_{ii} of correct classifications into category i and the numbers $\sum_{j=k; j,k \neq i} w_{jk}$ of correct classification into the non- i category.

Other error estimates are obtained by splitting the data into a training set and a test set and calculating the *training error rate* and the *test error rate*, or

TABLE 1
Previous results on secondary structure prediction.

The overall percentage of prediction (Q3 accuracy) is given and where available the sample sizes of the training and the test set in brackets [$n(\text{training}), n(\text{test})$]. In some cases the method could not be described in short terms and is missing (-). If more than one result is published, only the best is reported. In two cases where only the individual predictions for the α -helices and the β -sheet was published those are reported in parentheses. jack = jackknife procedure, $CV(x)$ = x -fold cross validation. For references see <http://www.dkfz-heidelberg.de/biostatistics/protein/protlist.html>.

AUTHOR	METHOD	%
Lim (1974)	physico-chemical characteristics	59
Chou & Fasman (1978)	preference index	57
Garnier et al. (1978)	GOR, maximum likelihood	56
Gibrat et al. (1987)	GOR II	63
Zvelebil et al. (1987)	GOR + evolutionary conservation	66.1 [-,11]
Levin et al. (1986)	KNN + homology	62.2
Biou et al. (1988)	GOR Combined,	65.5 [67,-; jack]
Levin & Garnier (1988)	KNN	63.0
Holley & Karplus (1989)	ANN	63.2 [48,14]
Qian & Seinowski (1988)	ANN	64.3 [91,15]
Rooman & Wodack (1988)	-	62
Kneller et al. (1990)	ANN + a priori information	65
King & Sternberg (1990)	symbolic machine learning	60 [43,18]
Muskal & Kim (1992)	tandem ANN - α	95.0 [105,15]
Salzberg & Cost (1992)	tandem ANN - β	95.4
Stolorz P et al. (1992)	machine learning	71 [100,28]
Zhang et al. (1992)	tandem FNN	63.5 [91,14]
Sasagawa & Tajima (1992)	ANN, nearest neighbor hybrid	66.4 [107,-]
Asai et al. (1993)	ANN	56.2 [33,29]
Leng et al. (1993)	Hidden Markov methods	66.0 [120,-;jack]
Yi & Lander (1993)	two level method	69.3
Rost & Sander (1993)	KNN + ANN + scoring	68
Rost & Sander (1994)	ANN + Jury	70.8 [130,-;CV(7)]
Ellis & Milius (1994)	-	72.5
Wako & Blundell (1994)	GOR	62.2 [239,-]
Geourjon & Deleage (1994)	-	77 [14,-]
Solovyev & Salamov (1994)	self-optimized, binary, similarity	69
Barlow (1995)	LDA + multiple alignment	68.2 [126,-;jack]
Salamov & Solovyev (1995)	hierarchical mixture of ANNs	63 [91,14]
Chandonia & Karplus (1996)	KNN + scoring table	72.2 [126,-]
DiFrancesco et al. (1996)	FNN+ sequence profiles	72.9 [318,-;CV(32)]
Frishman & Argos (1997)	Logistic regression	71.5 [115,-;CV(7)]
Ito et al. (1997)	local pairwise alignment	74.8 [125,-;jack]
Kawabata & Doi (1997)	3D-1D compatibility/pseudo energy	69.3 [325,77]
Fiser et al. (1997)	BW-mod GOR + mult. alignment	68.2 [126,-;CV(7)]
Levin (1997)	Deleage method - α	68.6 [80,-]
Rychelwski & Godzik (1997)	Deleage method - β	65.0
	KNN	72.8 [372,111]
	(update of Levin & Garnier, 1988)	
	segmented similarity after alignment	72.4 [256,256]

by calculating a *cross-validation error* (CV) or the *jackknife error* [Efron and Tibshirani (1993)], see also Grassmann et al. (1998).

Table 1 gives an overview of the development of the **secondary structure prediction**. Since its beginning, prediction accuracy has improved from less than 60% to more than 70%. Among the methods not using intrinsic biochemical information were mostly the nearest neighbor methods and artificial neural networks (ANNs). From a statistical point of view to mention is the quadratic logistic regression applied by DiFrancesco et al. (1996). They obtained a prediction error of 27.6% (with 7 fold CV). Interestingly, this rate was reduced to 21.5% when they incorporated techniques from bioinformatics as e.g. the relative frequencies of residues at each position after multiple alignment of homologue sequences, a variability score describing conserved residue patterns, or insertions and deletions. For details of the statistical modeling and the performance of the maximum likelihood estimation method see Di Francesco et al. (1996).

2.3. Tertiary structure prediction. Prediction of 3D structure is extremely complicated and has been confined to only a small number of shorter sequences. An illustrative view of the present state of tertiary structure prediction is obtained from reports on the recent contest of the Asilomar Conference in 1994 [Defay and Cohen (1995)]. Out of 33 proteins 14 were examined successfully in 12 laboratories. Fold prediction from primary sequence information was performed by 9 research groups performing 23 predictions on 11 sequences and obtaining 4 totally correct predictions. Hubbard and Park (1995) classify in another exercise 9 out of 27 sequences. They apply methods based on evolutionary information contained in multiple sequence alignments and hidden Markov models using various computer algorithms and alignment scores. In contrast to these predictions aiming at the 3D structure we will consider in our statistical classification two sets of classes:

a) the *four super-secondary classes (SSC)* defined above:

$$C = \{\text{only } \alpha, \text{ only } \beta, \text{ one part } \alpha \text{ plus one part } \beta (\alpha + \beta), \text{ and } \alpha \text{ and } \beta \text{ alternating } (\alpha/\beta) \}$$

b) the 42 classes of Reczko et al. (1994) based on topological similarity of the backbone and presented by the numbering of the classes (www[2]):

$$C = \{1, 2, \dots, 42\}$$

The SSC were chosen for our study for reasons of convenience and because of the possibility of comparison with the work of Reczko et al. (1994). SSC has also been investigated by Geourjon and Deleage (1994) and Efimov (1994), also Barton (1995). Supersecondary structure beyond these four classes has been investigated by Sun and coworkers. They used a vector projection method [Sun

et al. (1996)] and later also a feed forward neural net with one hidden layer of about half of the number of input units [Sun et al. (1997)] to predict a set of 11 standard motifs in 56 non-redundant proteins selected from a set of 240 sequences from the PDB. The motifs were defined as potential building blocks for tertiary structure and are characterized by a well defined 3D structure related to the backbone [Sun and Jiang (1996)]. Prediction accuracy obtained with the neural net for the 11 super-secondary classes of motifs ranged between 68% and 80% [Sun et al. (1997)] and was similar to the accuracy obtained in secondary structure prediction (Table 1). The vector projection method [Sun et al. (1996)] was tailored to those motifs and yielded an accuracy between 83% and 96%. This result may encourage stepping from primary via secondary to tertiary structure prediction. Reczko and Bohr (1994) actually tried this approach to some extent by combining the 42 fold classes with the SSC and they could improve their previous accuracy of 71% for the SSC prediction to about 91%.

2.4. Statistical methods. The problem the classification and prediction of secondary or tertiary structure can be formulated statistically in the framework of statistical decision theory. For this purpose we refer to Chapter 2 in Ripley (1996). General statistical decision theory and especially the background of Bayesian methods are found in the monograph of Berger (1985). Stolorz et al. (1992) introduced and discussed Bayesian analysis for secondary structure prediction. Grassmann et al. (1998) identified statistical methods of classification and discrimination as possible tools for fold prediction and applied them straightforwardly. They distinguish two cases: (i) methods based on the *posteriori class probability* and (ii) methods based on the *class conditional probability*. In both cases, an input vector x is assigned to its structural class k by a decision rule $d(x)$. The input vector $x = (x_1, \dots, x_p)$ is an element of the feature space X associated with the sampling units (i. e. the protein sequences), the structural class k is an element of C associated with the protein fold classification. Random elements of X and C are denoted by X and C , respectively.

(i) posteriori class probability

Case (i) builds the decision on the posteriori class probability of class k given x

$$p(k|x) = P(C = k|X = x)$$

A sequence x is assigned to that class k for which $p(k|x)$ is maximum. This assignment minimizes the total risk if a standard loss function is assumed [Ripley (1996) Chap. 2.1]. Such, the decision rule $d(x)$ is given as

$$d(x) = \{k \in C : p(k|x) = \max_j p(j|x)\}$$

This decision rule is directly related to regression which enables the application of logistic regression and in its sequel the use of the feed forward neural networks. Set

$$(2.1) \quad f_k(x) = E[Y_k|x] = P(C = k|x) = p(k|x)$$

where $Y_k, k = 1, \dots, K$ are “dummy” variables coding the class variable C as follows

$$Y_k = 1 \quad \text{if } C = k \quad \text{and} \quad Y_k = 0 \quad \text{if } C \neq k.$$

An example is the multiple logistic model

$$(2.2) \quad f_k(x) = p(k|x) = \frac{\exp(\eta_k(x))}{\sum_{m=1}^K \exp(\eta_m(x))}$$

with the linear predictor $\eta_k = \beta \cdot x$ [Ripley (1996) Chap. 3.5]. This generalizes the well known correspondence of Fisher’s linear discrimination and linear regression. Maximum likelihood methods are used to fit the model and to estimate $f_k(x) = p(k|x)$. The multiple logistic regression (2.1) and (2.2) is equivalent to the single-layer feedforward neural network (FNN) which uses as input the feature vector $x = (x_1, \dots, x_p) \in X$ and has the K output units Y_1, \dots, Y_K . $\eta_k(x) = w_k^T \cdot x$ represents the output function with weight vectors, see e.g. Grassmann and Edler (1996) and Schuhmacher et al. (1994). Below, we will apply this FNN and also the feedforward network FNN(H) with a layer of H hidden units [Ripley (1996) Chap. 5]. To minimize the error between the current state net output and the target output we use as error function the Kullback-Leibler distance which is equivalent to the use of the likelihood function. Weight decay regularizes the FNN. Notice, the number of hidden units plays an important role for the structure of the non-linearity and the dimension of the parameter space. Two further regression methods based on the posterior class probability are applied below as described in Grassmann et al. (1998). These are the *additive model* of Hastie and Tibshirani (1990) known as the so-called BRUTO method [Hastie et al. (1994)], and the *projection pursuit regression* (PPR) of Friedman and Stützle (1981).

(ii) class conditional probability Case (ii) builds the decision on the *class conditional probability* of the feature x given the class k

$$p(x|k) = P(X = x|C = k)$$

Using Bayes formula

$$p(k|x) = \frac{p(x|k) \cdot p(k)}{\sum_y p(x|k) \cdot p(k)} = \frac{p(x|k) \cdot p(k)}{p(x)}$$

the decision rule $d(x)$ can obviously be rewritten as

$$(2.3) \quad d(x) = \{k \in C : p(x|k) \cdot p(k) = \max_j p(x|j) \cdot p(j)\}$$

The prior probabilities $p(k)$ are either assumed to be known or they are estimated by the relative class frequencies $\hat{p}(k)$. The conditional densities $p(x|k)$ can be estimated either by assuming a parametric model $p(x|k, \theta)$ or by non-parametric methods (kernel or nearest neighbor methods), see Ripley (1996; Chap. 2 and 6). Assuming for $p(x|k)$ the multidimensional normal density as a parametric family we obtain the Linear Discriminant Analysis (LDA). Assuming different variance-covariance matrices for the different classes yields the Quadratic Discriminant Analysis (QDA). If the number of classes is large and if the number of available sequences with fold structure information is limited, the full QDA requires the estimation of a too large number of unknown parameters. Therefore QDA is restricted to the so-called QDA-MONO where only the diagonal elements (variances) differ between the classes.

Finally we use the *K-Nearest Neighbor Classification* (KNN) which assigns an object with feature vector x to the majority class of its neighbors. Decision rule 2.3 is applied with an estimate of the class conditional probability of the form

$$\hat{p}(x|k) = \frac{B_k}{n_k \cdot A(H, x)}$$

where B_k denotes the number of the K nearest neighbors of x that belong to class k , n_k is the total number of objects in class k , and $A(K, x)$ is the content of the smallest hypersphere containing the K nearest points to x [Ripley (1996) Chap. 6.2]. The methods described above were computationally realized by S-Plus software (e.g., *lda*, *qda*, *fda*, *knn*).

2.5. Data. The data used for illustration of the statistical classification and prediction described above originate from Reczko et al. (1994). They considered 268 proteins including a few sub-domains which had been classified into the four SSC and into 42 fold categories related to the Pascarella and Argos (1992) classification. Figure 3 exhibits the frequency of the SSC in the data set of the 268 protein sequences. This sample was subdivided into a training set of 143

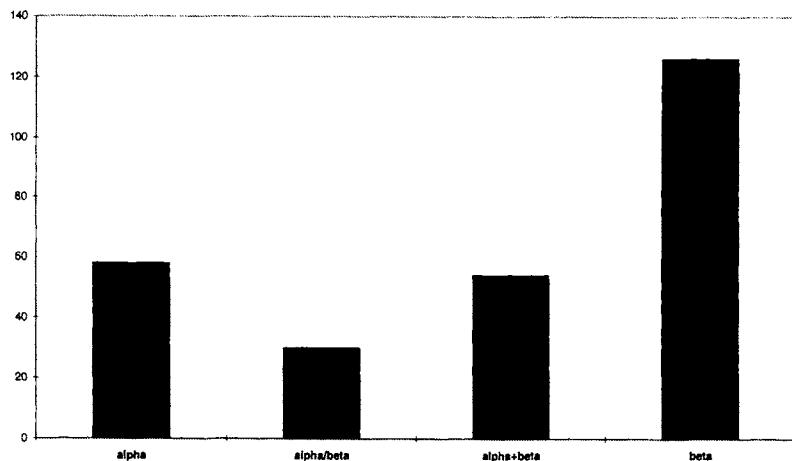


FIG. 3. Occupation frequency of the four supersecondary classes (SSC) of the data set of 268 sequences.

sequences and a test set of 125 sequences [same as used by Reczko et al. (1994)] by random sampling stratified according to the 42 classes such that each of the 42 fold categories was occupied at least by one sequence in the training set and in the test set and otherwise balanced at best, but putting ‘excess sequences’ into the training set, see [www\[2\]](#) for the set of sequences, the partition and the classification into the SSC and the 42 fold classes. Tentatively, we used a second partition with sizes of the training set and the test set in the ratio of about 2:1. In this case 90 proteins were randomly sampled into the test set and 178 remained in the training set.

3. Results of the prediction analysis. In this section we will illustrate the application of the methods described above through prediction based on primary sequence information. We used two simple feature spaces: the amino acid frequencies X_1 and the dipeptide matrices X_2 . Since the number of 143 sequences in the training set was smaller than 400, the dimension of X_2 , a principal component analysis (PCA) was applied in order to reduce the dimension. Deliberately, we set the cut off point to 90% explained variation and obtained so 74 remaining components. This defined a third feature space X_3 of dimension 74. For the use of PCA in protein classification see also Ferran et al. (1993). Table 2 outlines the classification task depending on the chosen feature informa-

TABLE 2
Outline of the classification task.

Classification was considered for two sets of classes: the supersecondary classes (SSC) and the 42 fold classes based on the backbone topology of Pascarella and Argos (1992). Feature information was available as the amino acid frequencies (AAF), the dipeptid frequencies (DPF), and the 74 first principal component values of the DPF: (DPF-74PC). The notation $p \rightarrow K$ informs on the dimension p of the feature space and the number of classes K . The DPF column was not realised in this evaluation because of the too large dimension of the feature space X in relation to the sample size.

Classification	Feature Information		
	AAF	DPF	DPF-74PC
SSC	$20 \rightarrow 4$	$(400 \rightarrow 4)$	$74 \rightarrow 4$
42-CAT	$20 \rightarrow 42$	$(400 \rightarrow 42)$	$74 \rightarrow 42$

tion (AAF or DPF) and the chosen classification (SSC or 42-CAT, the 42 fold classes). The error rate of the reclassification of the training sequences gives rise to the *apparent prediction error* (APE) which was determined to judge overfitting. As objective measures for the prediction error we calculated the *test prediction error rate* (TPE) and the *cross validation error rate* using a 10-fold cross validation (CV-10). We will focus here on prediction based on the AAFs of the SSCs ('20 → 4') and of the 42 fold classes ('20 → 42').

3.1. *Results for the case 20 → 4.* Figure 4 summarizes the error rates for the prediction of the SSCs through the AAFs ('20 → 4' case). Table 3 provides the error rates numerically. The FNN reached perfect apparent prediction on the training set with 7 and 9 to 14 hidden units in our standard splitting of 143 training and 125 test sequences, see Table 3 upper part. The test error rate (TPE) and the cross validation error rate (CV) decreased with few minor exceptions when the number H of hidden units is increased from 0 (logistic regression) to 10. Increasing H forced the error rates to a plateau. Increasing H further lead to numerically unstable estimates. Best prediction accuracy was 77.6% (22.4% CV(10), 23.2% TPE) obtained with an FNN(10). Projection pursuit regression (PPR) could almost reach this accuracy, see Table 3 middle. With a comparable number of terms ($H' = 12$), PPR could even beat the FNN(10) in terms of TPE (22.0%). PPR became less predictive with a higher number of terms. The discriminant analysis methods performed reasonably good, except the full QDA, which is obviously over-parameterized. QDA showed a perfect prediction on the training set, but it became a disaster on the test set also in terms of the CV error. Remarkably, QDA-MONO yielded one of the best results in '20 → 4' with 20.1% CV error. The additive model (BRUTO) performed almost identical to the LDA. It is also seen that the discriminant based methods show almost

no over-optimism in the apparent error rate (LDA: APE = 30.1 versus CV = 29.5, QDA-MONO : APE = 18.2 versus CV = 20.1). We investigated another splitting of the data into a larger training set and a smaller test set of a ratio of about 2:1 but we obtained worse results (see error rates in parentheses in Table 3). The better performance of QDA-MONO compared to LDA is exhibited in the discriminant plot in Figure 5. The three classes only α (1), α and β alternating (α/β) (2), one part α plus one part β ($\alpha + \beta$) (3), and only β (4) are further separated in the QDA-MONO discriminant plot and especially the two mixed types are better discriminated.

Because of the identity of the data sets we can now compare our results obtained by standard statistical methods directly with those of Reczko et al. (1994) who had used a cascade correlation network (a partially recurrent neural net allowing for varying topologies). Our accuracy is higher than theirs reported as 71% (TPE = 29%).

3.2. Results for the case $20 \rightarrow 42$. When we predicted the 42 fold classes from the 20 amino acid frequencies (' $20 \rightarrow 42$ '), see Table 4, the prediction accuracy became worse, as could be expected because of the larger number of classes. LDA exhibited almost as good results as the FNN(12) in terms of the test error rate. PPR showed results similar as the FNN as long as the number of hidden units H and number of terms H' was small. When increasing H and H', PPR became inferior to FNN. The result of QDA was comparable to that of LDA. The KNN was not calculated in this case.

3.3. Results for the case $400/74 \rightarrow 4/42$. The cases ' $400 \rightarrow 4$ ' and ' $74 \rightarrow 4$ ', described in detail in Grassmann et al. (1998), are summarized here for comparison with the cases above. The few successful technically working applications of the FFNs in the case ' $400 \rightarrow 4$ ' were not reliable because of the too high dimension of the feature space compared with the number of observations. Therefore, we restricted the analysis to the feature space X_3 of the first 74 principal components. In the case of ' $74 \rightarrow 4$ ' the FNN(10) provided in terms of CV the highest prediction rate of 77.6%. LDA gave 73.1% and QDA-MONO and BRUTO only 62.7% and 64.6%, respectively. In the case ' $74 \rightarrow 42$ ' the FNN(9) provided the highest prediction rate of 67.5%. LDA yielded an even better result of 69.8%; QDA-MONO 61.6% and BRUTO 64.6%. The prediction accuracy Reczko et al. (1994) obtained by a cascade-correlation network was about 73%. They could improve the classification by enlarging the set of classes to 45 [Reczko and Bohr (1994)] and to 49 [Reczko et al. (1997)]. They report then an accuracy of about 82% for the 42 classes when using another constraint network.

TABLE 3
Prediction of supersecondary classes (SSC) from the amino acid frequencies (AAF).

The apparent error (APE), test error (TPE) and cross-validation error are presented for the neural networks with varying numbers of hidden units: FNN(H) the projection pursuit regression with varying number of terms H' PPR(H', linear discriminant analysis: LDA, Quadratic discriminant analysis: QDA, quadratic discriminant analysis restricted to varying variances: QDA-MONO, generalized additive model: BRUTO, K-th nearest neighbor method: KNN. The set of 286 sequences were split between training and test set in two ways denoted by '143'/'125': (143 training, 125 test) and '178'/'90' : (178 training, 90 test).

	Classification Error in %				
	APE		TPE		CV-10
	training set	test set	test set	cross-validation	
FNN: H hidden units					
0	30.8	(34.3)	33.6	(34.4)	36.9
1	41.3	(40.4)	37.6	(38.9)	42.5
2	25.9	(25.3)	30.4	(42.2)	33.6
3	30.1	(18.0)	32.0	(35.6)	30.2
4	9.1	(11.2)	31.2	(37.8)	31.7
5	2.8	(9.0)	29.6	(28.9)	27.2
6	0.7	(11.2)	32.0	(28.9)	26.9
7	0.0	(11.8)	24.0	(27.8)	27.6
8	2.1	(10.1)	25.6	(30.0)	26.1
9	0.0	(11.2)	26.4	(28.9)	26.1
10	0.0	(11.2)	23.2	(27.8)	22.4
11	0.0	(10.1)	20.8	(26.7)	25.7
12	0.0	(2.8)	21.6	(24.4)	24.6
13	0.0	(3.9)	21.6	(26.7)	23.1
14	0.0	(2.8)	27.2	(24.4)	21.6
15	2.8	(2.8)	20.0	(24.4)	19.8
PPR: H' terms					
1	36.5		42.4		
2	25.2		32.8		
4	16.8		29.6		
8	6.3		24.0		
10	0.7		32.8		28.0
12	0.7		22.0		
13	0.7		20.0		
20	0.0		32.0		27.2
LDA	30.1	(29.2)	36.0	(33.3)	29.5
QDA	0.0	(0.6)	60.0	(72.2)	60.4
QDA-MONO	18.2	(16.9)	28.0	(32.2)	20.1
BRUTO	30.1	(29.8)	36.0	(34.4)	29.5
KNN	0.0	(-)	19.2	(-)	22.8

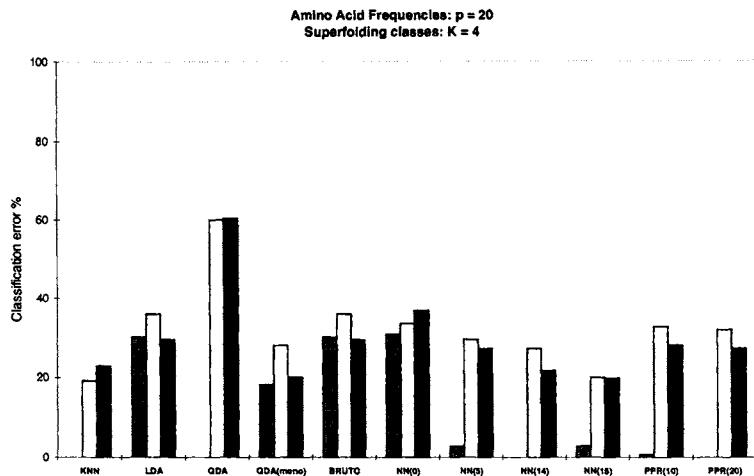


FIG. 4. Histogram of the error rates (apparent, test, and cross-validation error rate) of the classification into the four SSCs using the AAF information.

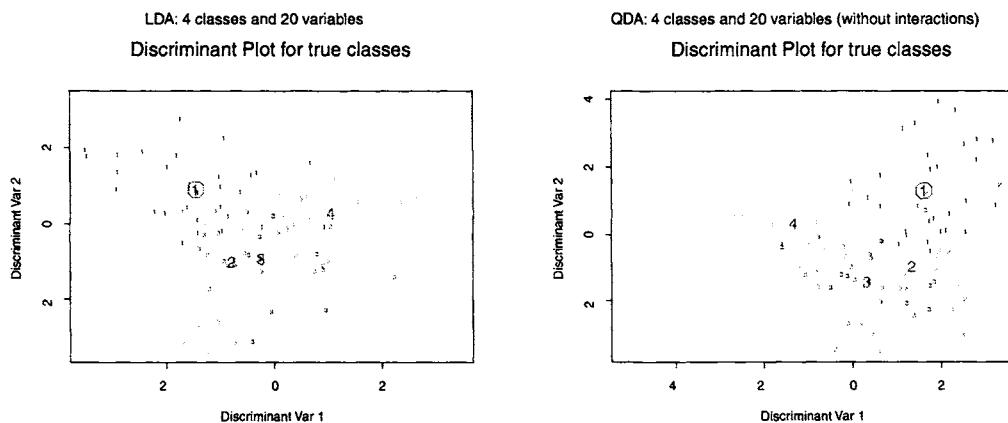


FIG. 5. Discriminant plot of the Linear Discriminant Analysis (LDA), upper part, and the Quadratic Discriminant Analysis without varying interaction (off-diagonal) terms (QDA-MONO), lower part, for the classification 20 → 4 of 125 test sequences into the four supersecondary classes (SSC) α (1), α and β alternating (α/β) (2), one part α plus one part β (α + β) (3) and only β (4) on the basis of the 20 amino acid frequencies (AFF).

TABLE 4

Prediction of the 42 fold classes based on backbone topology of Pascarella and Argos (1992) from the amino acid frequencies (AAF).

The apparent error (APE), test error (TPE) are presented for the neural networks with varying numbers of hidden units: FNN(H) the projection pursuit regression with varying number of terms H': PPR(H'), the linear discriminant analysis: LDA, the quadratic discriminant analysis: QDA, the quadratic discriminant analysis restricted to varying variances: QDA-MONO, and the generalized additive model: BRUTO. The set of 286 sequences was split into 143 training sequences and 125 test sequences.

FNN:	H hidden units	Classification Error in %	
		APE	TPE
		training set	test set
	0: Log. Regr.	0.0	46.4
	1	76.2	74.4
	2	59.4	61.6
	4	28.0	48.0
	6	16.8	32.8
	8	9.2	31.2
	12	0.7	27.2
PPR:	H' terms		
	6	61.5	60.8
	10	48.3	56.0
	15	33.6	48.0
	17	27.3	44.0
	18	23.1	44.8
LDA		2.8	27.5
QDA		0.0	29.7
QDA-MONO		0.0	71.4
BRUTO		2.8	27.5

4. Discussion. The prediction of protein folds from their amino acid sequence is an impressively long-standing challenge in molecular biology and biophysics [Finkelstein (1997)]. After a few blind predictions in the seventies it was realized in the eighties that the efforts are 'not hopeless'. Two large scale blind predictions performed in 1994 [Moult et al. (1995), Prediction Center (1996)] exhibit the difficulty to predict 3D folds systematically if the proteins are not closely related to previously known ones, see Lemer et al. (1995), Defay and Cohen (1995) and Hubbard and Tramontano (1996). Prediction experience of the past years showed that the most successful tools are knowledge based systems in combination with experience and statistical methods [Rost and O'Donoghue (1997)]. To our knowledge, tertiary structure predictors have used almost always rather small data sets. Statistical procedures which exhibit their power on large sizes much better have not been applied systematically. Our results obtained with statistical classification methods show that their application is also 'not

hopeless' and that their combination with biophysical methodology may add quality to future prediction, which is interesting in face of the fastly increasing information from the protein data bases. A statistical approach to the problem, with careful attention to assumptions, variation, sampling, defensible precision estimates, and realistic estimates of error probabilities, could strengthen existing procedures as well as provide new ones. Compared with the present aims of protein prediction [Rost and O'Donoghue (1997)] our results above fall somehow short when considered for improving the direct protein prediction. This is not surprising given the simplicity of the feature spaces used in our analysis and the fact that no physical and chemical properties of the molecular level were included. Further research is needed when a richer - and then also more complex - feature space is used, see e.g. the proposal of class-directed structure where only representative members of classes will be fully structurally characterized [Terwilliger et al. (1997)]. Our investigation focussed in the role of standard and new statistical methods with the goal of identifying a possibly best statistical procedure. We considered especially the neural networks. FNNs are in fact non-linear regression methods subordinated under posterior class probability based classification. A further aim was the application of these methods on a larger data set than it has been used in most previous evaluations of automatic protein fold prediction. From our analyses we conclude that linear discriminant and nearest neighbor methods are potent competitors to the more flexible neural networks. The results obtained from the modern discriminant and a regression methods (e.g., BRUTO, PPR) were mixed. In some cases these methods competed very well in other cases the results we obtained so far were disappointing in the sense that they were not able to yield an improved prediction and had sometimes serious problems to cope with the ill-posedness of the classification problem. Typical for this was the failure of the QDA when the number of input variables became larger than the sample size. Search for more efficient regularization methods is needed to exploit the power of these new statistical methods. Previous predictions especially for secondary structure proved the usefulness of FNNs. This was corroborated in our investigation where the FNNs competed well with other methods. This is not surprising since the FNNs are non-linear regression methods and implement standard statistical tools. However, as universal approximators, neural nets are always in danger of overfitting, which we experienced in our analysis too. Therefore, the bias-variance trade off has to be considered carefully. Automatic smoothing and regularization are statistical methods to be investigated further. A clear disadvantage of all FNNs is the lack of interpretation of the weights and the fact that quite different weights and weight patterns can lead to the same prediction outcome.

The optimal method for assessing the validity of the prediction procedure

would be the use of an independent validation set sampled from the set of all proteins. We used a test set after dividing the sample of 268 sequences into 143 training and 125 test sequences to calculate the test error rates (TPE). Those may be still a little optimistic, but the bias is usually in the order of a few percentages, when compared with the CV error. Similar small biases have been observed by DiFrancesco et al. (1996) for secondary structure prediction. This corroborates the recommendation of using some sort of cross-validation error, at least as long as no independent validation data are available. We calculated the cross-validation error (CV-10 fold) mostly in addition to the TPE. In some cases as e.g., PPR, however, computing of the CV error became very time consuming and was not performed for each architecture. The fact that our error rates obtained by the statistical procedures are around 30% is not too disappointing given that no higher order information from the protein was taken into account. Without that it might be difficult to surpass that margin. Obviously, more protein sequence data and perhaps both, more informative feature spaces and functionally more realistic class definitions are needed. Inclusion of information on distant interaction in the protein sequence and its quantitative presentation could be as helpful as the use of physico-chemical properties of the amino acids. At present it seems that the number of classes and the classification itself is too much tailored to the existing information on 3D structure. The number of existing relevant structural families is estimated to about 200-500 but present fold class prediction is limited to much smaller numbers, as our analysis of 42 classes or the analysis of Sun et al. (1996) of 11 classes.

The prediction methods from above assume independent sampling of a complete and correct sequence. Only then are the statistical model estimates and the accuracy measures valid statistics. However, in practice occur errors during sequencing and in a number of cases interest focuses in parts of proteins only, as e.g. motifs. This creates for any procedure - not only the statistical ones - the errors-in-variables problem and the non-independent sampling problem which both need further investigation.

Usually, protein researchers distinguish two situations: (i) Presence of sequence similarity such that the investigated sequences is similar to a sequence of known 3D structure in the data base. (ii) Absence of sequence similarity such that no similar sequence exists in the 3D data base. Similarity is defined as sequence identity after alignment of at least 25% to 30%, which is at the same time considered sufficient to infer structural similarity [Rost and O'Donoghue (1997), Schneider et al. (1997)]. The distinction between (i) and (ii) established the prediction paradigm: If a protein has been newly sequenced, then search the 3D data base for at least one sufficiently similar sequence. If one is found structure and function is predicted from the knowledge available from that. If

none is found an automated structure prediction is tried by comparing the sequence information of the new sequence with the sequence information available for all proteins whose 3D structure has been clarified so far. In our analysis we did not account thoroughly for the effect of the similarity between the 268 sequences. Details on the sequence similarity is provided in www[2]. However, it is shown in Grassmann et al. (1998) how with our data the prediction accuracy decreases when the dissimilarity increases. Further research on the performance of the statistical procedures is needed on dependency of sequence similarity. The recently provided representative sample from the PDB data bank of 838 proteins of Holm and Sander (1996) could be an excellent data set. There, all sequences have less than 25% sequence identity and fall into 270 different fold classes with class occupancy ranging between 1 and 73, see <ftp://ftp.embl-heidelberg.de/databases/fssp/>.

Acknowledgements The authors thank two anonymous referees for their critical but also very constructive and helpful comments. We thank Sandor Suhai and Martin Reczko for providing us the with the data and some biophysical background and for John Crowley drawing once our attention to neural nets. For stimulating discussions on statistical classification and prediction and help in applying them the second author is very grateful to Trevor Hastie. Part of this research was supported by the German Academic Exchange Service (DAAD, Doktorandenstipendium HSP II/AUFE). Finally, the first author is grateful to AMS for support to attend the Summer Research Conference in Seattle.

REFERENCES

- ANFINSEN, C., HABER, E., SELA, M. and WHITE, F.J. (1961) The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proceedings of the National Academy of Sciences of the USA* **47** 1309-1314.
- BAIROCH, A. and APWEILER, R. (1995) The SWISS-PROT protein sequence data bank and its new supplement TREMBL. *Nucleic Acids Research* **24** 21-25.
- BARTON, G.J. (1995) Protein secondary structure prediction. *Current Opinions in Structural Biology* **5** 372-376.
- BERGER, J.O. (1985) *Statistical Decision Theory and Bayesian Analysis*. Springer, New York.
- BERNSTEIN, F.C., KOETZLE, T.F., WILLIMAS, G.J.B., MEYER, E., BRYCE, M.D., ROGERS, J.R., KENNARD, O., SHIKANOUCHI, T. and TASUMI, N. (1977) The Protein Data bank: a computer based archival file for macromolecular structures. *Journal of Molecular Biology* **112** 535-542.
- BENSON, D.A., BOGUSKI, M.S., LIPMAN, D.J. and OSTELL, J. (1997) GenBank. *Nucleic Acids Research* **25** 1-6.
- BRANDEN, C. and TOOZE, J. (1991) *Introduction to Protein Structure*. Garland Publ., New York.
- CREIGHTON, T.E. (1984) *Proteins. Structure and Molecular principles*. Freeman, New York.
- CREIGHTON, T.E. (1990) Understanding protein folding pathways and mechanisms. In *Protein Folding*, Gierasch, L. M., King, J. (eds), Amer. Assoc. Adv. Sci., Washington, 157-170.
- DEFAY, T. and COHEN, F.E. (1995) Evaluation of current techniques for ab initio protein prediction. *Proteins* **23** 431-445.

- DICKERSON, R.E. and GEIS, I. (1983) *Hemoglobin: Structure, Function, Evolution and Pathology*. Benjamin Cummings, Menlo Park CA.
- DiFRANCESCO, V., GARNIER, J. and MUNSON, P.J. (1996) Improving protein secondary structure prediction with aligned homologous sequences. *Protein Science* **5** 106–113.
- DINUR, U. and HAGLER, A. T. (1991) New approaches to empirical force fields. In *Reviews in Computational Chemistry*, Vol. II, Lipkowitz, K. B., Boyd, D. B. (eds). VCH Pbl., New York, 99–164.
- EFIMOV, A.V. (1994) Super-secondary structures in proteins. In *Protein Structure by Distance Analysis*, Bohr & Brunak (Eds) IOS Press Amsterdam, 187–200.
- EFRON, B. and TIBSHIRANI, R.J. (1993) *An Introduction to the Bootstrap*. Chapman & Hall, Cambridge.
- FASMAN, G. (1989) *Prediction of Protein Structure and the Principles of Protein Conformation*. Plenum, New York.
- FETROW, J.S., PALUMBO, M.J. and BERG, G. (1997) Patterns, structures, and amino acid frequencies in structural building blocks, a protein secondary frequencies in structural building blocks, a protein secondary structure classification scheme. *Proteins* **27**, 249–271.
- FERRAN, E.A., FERRARA, P. and PFLUGFELDER B. (1993) Protein classification using neural networks. In: ISMB-93. Proceedings of the First International Conference on Intelligent Systems for Molecular Biology, Hunter, L., Searls, D., Shavlik, J. (Eds). AAAI Press, Menlo Park, CA, 127–135.
- FINKELSTEIN, A.V. (1997) Protein structure: what is it possible to predict now? *Current Opinion in Structural Biology* **7** 60–71.
- FRIEDMAN, J.H. and STÜTZLE, W. (1981) Projection pursuit regression. *Journal of the American Statistical Association* **76** 817–823.
- FRIEDRICH, M.S., GOLDSTEIN, R.A. and WOLYNES, P.G. (1991) Generalized protein tertiary structure recognition using associative memory Hamiltonians. *Journal of Molecular Biology* **222** 1013–1034.
- GEOURJON, C. and DELEAGE, G. (1994) SOPM: A self-optimized method for protein secondary structure prediction. *Protein Engineering* **7** 157–164.
- GIERASCH, L.M. and KING, J. (1990) *Protein Folding: Deciphering the Second Half of the Genetic Code*. Amer. Assoc. Adv. Sci., Washington.
- GRASSMANN, J. (1996) Artificial neural networks in regression and discrimination. In *Softstat '95, Advances in Statistical Software 51*, Faulbaum, F. and Bandilla, W. (Eds). Lucius & Lucius, Stuttgart, 399–406.
- GRASSMANN, J. and EDLER, L. (1996) Statistical classification methods for protein fold class prediction. In *COMPSTAT. Proceedings in Computational Statistics*, Prat A. (ed). Physica-Verlag, Heidelberg, 277–282.
- GRASSMANN, J., SUHAI, S., REČKO, M. and EDLER, L. (1998) Protein Fold Class Prediction: Statistical Classification versus Artificial Neural Networks. Manuscript submitted.
- HASTIE, T. and TIBSHIRANI, R.J. (1990) *Generalized Additive Models*. Chapman & Hall, Cambridge.
- HASTIE, T., TIBSHIRANI, R.J. and BUJA, A. (1994) Flexible discriminant analysis by optimal scoring. *Journal of the American Statistical Association* **89** 1255–1270.
- HOLLEY, L. and KARPLUS, M. (1989) Protein secondary structure prediction with a neural network. *Proceedings of the National Academy of Sciences of the USA* **86** 152–156.
- HOLM, L. and SANDER, C. (1994) Searching protein structure databases has come of age. *Proteins* **19** 165–173.
- HOLM, L. and SANDER, C. (1996) Alignment of three-dimensional protein structures: network server for database searching. *Methods in Enzymology* **266** 653–62.
- HUBBARD, T. and TRAMONTANO, A. (1996) Update on protein structure prediction: results of the 1995 IRBM workshop. *Folding and Design* **1** R55–R63.
- LAMBERT, M. and SCHERAGA H. (1989) Pattern recognition in the prediction of protein structure. I-III. *Journal of Computational Chemistry* **10** 770–831.

- LEMER, C.M.R., ROOMAN, M.J. and WODAK, S.J. (1995) Protein structure prediction by threading methods: evaluation of current techniques. *Proteins* **23** 337–355.
- MATTHEW, B. W. (1975) Comparison of the Predicted and Observed Secondary Structure of T4 Phage Lysozyme. *Biochimica et Biophysica Acta* **405** 442–451.
- MICKLER, T.A. and LONGNECKER, D. E. (1992) The immunosuppressive aspects of blood transfusion. *I. Intensive Care Medicine* **7** 176–188.
- MOULT, J., JUDSON, R., FIDELIS, K. and PEDERSEN, J.T. (Eds) (1995) Large scale experiment to assess protein structure prediction methods. *Proteins* **23** ii–iv.
- NEUMAIER, A. (1997) Molecular modeling of proteins and mathematical prediction of protein structure. *SIAM Reviews* **39** 407–460.
- PASCARELLA, S. and ARGOS, P. (1992) A data bank merging related protein structures and sequences. *Protein Engineering* **5** 121–137.
- PERUTZ, M.F. (1978) Hemoglobin structure and respiratory transport. *Scientific American* **239** 92–125.
- PREDICTION CENTER (1996) Protein structure prediction center on WWW URL <http://Predictioncenter.llnl.gov/>.
- RECZKO, M. and BOHR, H. (1994) The DEF data base of sequence based protein fold class predictions. *Nucleic Acids Research* **22** 3616–3619.
- RECZKO, M., KARRAS, D. and BOHR, H. (1997) An update of the DEF database of protein fold class prediction. *Nucleic Acids Research* **25** 235.
- RECZKO, M., BOHR, H., SUBRAMAMIAM, S., PAMIDIGHANTAM, S. and HATZIGEORGIOU, A. (1994) Fold class prediction by neural networks. In *Protein Structure by Distance Analysis*, Bohr & Brunak (Eds). IOS Press Amsterdam, 277–285.
- RICHARDS, F.M. (1991) The protein folding problem. *Scientific American* **264** 54–63.
- RIPLEY, B.D. (1996) *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge.
- ROST, B. and O'DONOGHUE, S. (1997) Sisyphus and prediction of protein structure. *CABIOS* **13** 345–356.
- ROST, B. and SANDER, C. (1994) Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins, Structure, Function and Genetics* **19** 55–72.
- ROWEN, L., MAHAIRAS, G. and HOOD, L. (1997) Sequencing the human genome. *Science* **278** 605–607.
- SCHULZ, G.E. (1988) A critical evaluation of methods for prediction of protein secondary structures. *Annual Review of Biophysics and Chemistry* **17** 1–21.
- SCHUMACHER, M., ROSSNER, R. and VACH, W. (1994) Neural networks and logistic regression: Part I. *Computational Statistics and Data Analysis* **21** 661–682.
- SCHUMACHER, M., ROSSNER, R. and VACH, W. (1994) Neural networks and logistic regression: Part II. *Computational Statistics and Data Analysis* **21** 683–701.
- SCHNEIDER, R., de DARUVAR, A. and SANDER, C. (1997) The HSSP database of protein structure-sequence alignments. *Nucleic Acids Research* **25** 226–239.
- STERNBERG, M.J.E. (1996) *Protein Structure Prediction*. Oxford University press, Oxford.
- STOLORZ P, LAPEDES A and XIA Y (1992) Predicting protein secondary structure using neural net and statistical methods. *Journal of Molecular Biology* **225** 363–377.
- SUN, Z. and JIANG, B. (1996) Patterns and conformations of commonly occurring supersecondary structures (basic motifs) in Protein Data Bank *Journal of Protein Chemistry* **15** 675–690.
- SUN, Z., ZHANG, C.-T., Wu, F.-H. and PENG, L.-W. (1996) A vector projection method for predicting supersecondary motifs. *Journal of Protein Chemistry* **15** 721–729.
- SUN, Z., RAO, X.-Q., PENG, L.-W. and Xu, D. (1997) Prediction of protein supersecondary structure based on the artificial neural network. *Protein Engineering* **10** 763–769.
- TAYLOR, W.R. (1992) *Patterns in Protein Sequence and Structure*. Springer, New York.
- WANG, Z.X. (1998) A re-estimation for the total numbers of protein folds and superfamilies. *Protein Engineering* **11** 621–626.

www[1]: <http://www.dkfz-heidelberg.de/biostatistics/protein/proflit.html>

www[2]: <http://www.dkfz-heidelberg.de/biostatistics/protein/gsme97.html>

ZHANG, C-T. (1997) Relations of the numbers of protein sequences, families and folds. *Protein Engineering* **10** 757-761.

GERMAN CANCER RESEARCH CENTER, HEIDELBERG
BIOSTATISTICS UNIT - R0700
PO Box 10 19 49
69009 HEIDELBERG
GERMANY
EDLER@DKFZ-HEIDELBERG.DE

BRAIN²
LEITENWEG 8A
97286 WINTERHAUSEN
GERMANY
JANET.GRASSMANN@BRAINN.DE

