ICM 1998, Berlin, Aug. 18–27

Abstracts of Plenary and Invited Lectures

Section:

12. Probability and Statistics

1991 MS Classification: 60J27
Siegmund, David, Stanford University, USA
Genetic linkage analysis: an irregular statistical problem

The goal of gene mapping, or linkage analysis, is to locate genes that affect particular traits, e. g., genes that affect human susceptibility to particular diseases or genes that affect productivity of agriculturally important species. Recent changes in technology that provide large numbers of informative genetic markers at known locations throughout a genome suggest new statistical problems concerned with the design and analysis of gene mapping experiments. An artificially simplified, but illuminating genetic model leads to the following class of statistical problems. Observations are available on a doubly indexed set of random variables $Z(c, i\Delta)$, where $c = 1, \cdots, 23$ indexes the set of human chromosomes of genetic lengths ℓ_c and $i\Delta, 0 \leq i\Delta \leq \ell_c$ are the locations of markers spaced at intermarker distance Δ along each chromosome. For different values of c the random variables are independent. For each fixed c, $Z(c, t)$ is a Gaussian process in t, which satisfies

$$\mathrm{Cov}[Z(c, s), Z(c, t)] = \exp(-\beta|t - s|). \tag{1}$$

For most or perhaps all values of c

$$E[Z(c, t)] = 0 \text{ for all } t, \tag{2}$$

while for some c', $0 < \tau < \ell_{c'}$ and $\xi > 0$

$$E[Z(c', t)] = \xi \exp(-\beta|t - \tau|). \tag{3}$$

The values of c', τ, and ξ are all unknown. Thus the data consist of a large number of zero mean Ornstein-Uhlenbeck processes observed at equally spaced "time" points. A small number of these processes are superimposed on a double exponential drift defining a "peak" of an unknown height ξ at an unknown location τ. The problem is to decide which chromosomes, if any, harbor such a location τ and estimate the location by a confidence region.

The purpose of this paper is (i) to explain briefly the genetic background of the preceding problem, (ii) propose a framework for its solution that is useful as a point of departure for discussing more realistic versions of the problem, and (iii) describe some alternative models designed to capture the

complicating features arising in practice. Special consideration is given to (a) the problem of multiple comparisons that arises through examining the large number of variables $Z(c, i\Delta)$ in searching for the relatively few values of c', t where the expected value is substantially different from 0, (b) the power to identify correctly the values c' as a function of the principal parameters τ, Δ, ξ as well as other parameters that enter into more realistic models, and (c) estimation of τ by confidence regions. Some of these problems can be understood in terms of recent literature on change-point problems, to which they are closely related. Particular attention will be given to complex and quantitative traits, which one expects to involve multiple, possibly interacting, genes, and to the relative efficiency of different strategies.