Statistical Methods in Medical Research 2009; **18**: 435–436

Editorial

Special Issue on Statistical Bioinformatics

Bioinformatics is the meeting ground of molecular biology, computer science, mathematics and statistics. From its origins in computational biology, the subject has grown tremendously in the last 10 years, propelled by ever-expanding genomic and proteomic databases, themselves the result of massive investment and technological advances in DNA sequencing, gene expression measurement and macromolecular structure determination. New technologies continue to push back the frontiers of science, providing many computational, mathematical and statistical challenges.

Statisticians have played an important part in this scientific revolution, nowhere more so than in the study of gene expression. Genes within each cell's DNA provide the templates for proteins, which are the workhorses of many of the structural, biochemical and signalling processes of the body. Control of gene activity is mediated by other molecules, including proteins, in response to the organism's needs to feed, grow, heal and reproduce. Gene expression studies measure the level of gene activity, and microarray platforms allow this to be done in tens of thousands of genes simultaneously. Microarray data are often noisy and exhibit many artefacts. Inferring the truth underlying such data has been the subject of many statistical papers. In this issue, Andy Lynch and co-authors introduce the field, then go on to describe a state-of-the-art microarray platform with some statistical approaches to exploiting its potential.

DNA sequence underlies almost all of biology. Availability of abundant DNA sequence data is a new phenomenon in the biological sciences. The human genome sequence was largely completed in 2003, and since then many more animal and some plane genomes have been sequenced. Having the sequence is one thing, understanding it is something else. Looking for patterns in DNA, for example strings which occur multiply within a genome or recur in many genomes, is one way to gain a purchase on the problem. Such patterns rarely recur perfectly; typically they possess substitutions, insertions or deletions, reflecting their independent evolutionary paths from shared ancestry. Aligning sequences to reveal these modifications has been a major preoccupation in bioinformatics, spawning many approaches and algorithms. Purely algorithmic approaches, however, tend to flounder in the presence of complex evolutionary patterns. Here, statistical modelling and methods can provide a more principled approach, as discussed in this issue by István Miklós and co-authors.

Evolution of DNA sequence is most often studied at the level of genes. However, larger scale evolutionary events involving whole sets of genes, wherein large segments of DNA sequence are removed, duplicated, inverted or translocated between chromosomes, can produce intricate patterns in DNA sequence, and may give rise to large

10.1177/0962280209349831

gene families containing tens, hundreds or even thousands of related genes within one species, extending across an entire phylum of the tree of life. The study of these larger scale evolutionary processes is becoming increasingly possible through the availability of whole-genome sequence in distantly related organisms. Tom Nye, in this issue, describes such processes and explores the potential for statistical modelling of the evolutionary history of multi-gene families.

Proteins are chains of amino acids, the sequence of a chain being encoded in the DNA sequence of its gene. Assisted by the development of bioinformatic tools, burgeoning DNA sequence databanks have led to a concomitant increase in protein sequence data. However, possession of a protein's amino-acid sequence does not by itself tell us much about its three-dimensional structure or biological function, and proteins depend absolutely on their structure to function correctly. While technological advancement continues to accelerate the acquisition of high-quality DNA sequence, development of high-throughput laboratory processes for determining the three-dimensional structure of proteins has been much more difficult. In principle, the sequence of a protein dictates its three-dimensional form; so methods to predict structure from sequence would be of enormous value. This holy grail of structural biology has yet to be discovered, despite decades of research. In this issue, Thomas Hamelryk reviews this field, and describes some new statistical approaches.

These four papers represent a cross-section of current research in statistical bioinformatics, but provide no more than a glimpse of the breadth of work currently underway in the field of genomics at large, much of it offering great possibilities for statistical involvement and innovation. With increasing integration of the fields of statistical genetics and statistical bioinformatics, the subject should more properly be termed *statistical genomics*. Its development will continue apace with the continual advancement in technology, providing ample scope for open-minded statisticians seeking fresh challenges.

Wally R. Gilks
Professor of Statistical Bioinformatics,
University of Leeds, UK
and Head of Statistics,
Rothamsted Research,
Harpenden, UK
E-mail: wally.gilks@bbsrc.ac.uk