MATHEMATICS MATTERS

Stats in Your Genes

Sequencing the human genome was a fantastic achievement, but it was only the beginning. Now, statisticians are coming up with new methods to sift through large amounts of a genetic data and identify the differences in DNA that can lead to diseases.

In 2003 the Human Genome Project published the first complete version of our genetic code, opening the doors to a new revolution in medicine. Understanding how certain genes are linked to cancer, diabetes or many other diseases offers the hope of new cures and treatments, but there is still much more research to be done, and all of this work is heavily reliant on statistics and computing. Sequencing the 3.2 billion letters of the human genome took 13 years, but this was actually two years ahead of schedule, thanks in part to advances in statistical and computing techniques. Now, as biological experiments produce larger and larger amounts of genetic data, the role of statisticians in genomics is increasingly important.

Much of this data is now gathered from genome-wide association studies, which take DNA samples from individuals with and without a certain disease and identify the differences in their genomes. The genetic code of individuals within both sets is compared, and if individuals with the disease have particular features that differ from those without the disease, there is potentially a link between those regions and the disease.

These features can be identified using microarrays, which contain millions of DNA probes on a single slide no larger than a postal stamp. Such arrays are used to identify features in large numbers of DNA samples, generating masses of raw data that must be cleaned up with statistical techniques before it can be used. The result is a read-out of the genetic differences, along with an assessment of the accuracy of the data. Microarray technology has evolved at an incredibly fast pace, with new arrays entering the market every few months in an attempt to bring down the cost of individual readings, and every time the technology is revised, these statistical techniques also have to be re-engineered.

> "Finding genetic differences helps identify individuals who are at risk from certain diseases."

Once the data has been collected and cleaned up, further statistical techniques are used to draw conclusions, but this is not simply "finding the gene for x" as it is often presented in the media. The simplest variations within the genetic code are called single-nucleotide polymorphisms (SNPs), positions in the DNA that differ from person to person. Statisticians use a method known as regression in order to determine whether a SNP is linked to a particular disease but traditional methods of regression are not up to the task, so researchers such as Simon Tavaré at the University of Cambridge have developed more advanced statistical tools. Tavaré and colleagues have come up with a regression technique known as sparse partitioning that efficiently identifies the important SNP positions and the interactions between them.

Finding these genetic differences enables the development of genetic tests to identify individuals who are at risk from certain diseases, helping them to get treatment faster. Such tests already exist for a range of diseases, including Alzheimer's and some forms of cancer, and more are currently being developed.

Looking ahead, one objective is to provide personalised medicine to patients, so that they receive optimised treatments according to their genetic profiles. For example, it is thought that half of women with breast cancer are given unnecessary treatment that won't help to cure their disease, simply because it is not currently possible to identify which individuals will respond positively to particular treatments.

In order to achieve this, it is important that biologists work with statisticians to ensure clinical trials are conducted in the best

Institute of mathematics & its applications
EPSRC Pioneering research and skills

possible way. Clinical trials are often run by dividing patients into two groups: one group gets a placebo while the other does not. In a more complicated trial, each group might be switched to the other treatment in a second round, so each patient receives one active treatment and one placebo throughout the course of the trial. Doctors then identify the effects of the treatment by comparing each individual's change in response.

One problem with this method, identified by Stephen Senn at the University of Glasgow, is that patients might not respond consistently to treatment. For example, in a trial with 100 patients, 70 might respond to treatment while 30 do not. The common interpretation says that the treatment always works for 70% of patients and never works for the other 30%. An equally statistically valid interpretation is that the treatment works for all patients, but only 70% of the time.

Determining which interpretation is true is vital if we are to develop personalised medicine, because genetic differences are only a factor in the first case. Solving this issue requires trials in which patients repeatedly switch between treatment and placebo, so that more detailed comparisons can be made using a technique known as random effects modelling. This separates out the variability of a trial into different sources, such as natural variation between patients or even between the same patient at different points in time, allowing statisticians to pin down a patient's individual response.

What ever the future holds, it is clear that sequencing the human genome was a fantastic scientific achievement made possible by advanced statistical work. It unlocked the tantalising possibility of personalised medicine, but much more work is needed for this to be realised. As DNA sequencing and related technologies develop over the next few years, the amount of data produced is expected to increase one hundredfold or more. Sorting through this data is an immense task, so biologists and statisticians will need to continue working closely together.

TECHNICAL SUPPLEMENT

Sparse partitioning regression

Statisticians have used different forms of regression for over 200 years. The earliest methods only work for a small handful of data, but modern techniques are far more powerful and are constantly being improved. Simon Tavaré's latest method, known as sparse partitioning, can handle large datasets with interacting variables, making it useful for analysing whole genome association studies.

For example, an association study for a particular disease involves looking through the genetic data to find correlations. If people with the disease all show one SNP variety at a particular point in the genome, and people without have the other SNP variety, there is likely an association between the genome position and the disease. In practice the pattern won't be so clear cut, so regression is needed to seek out the truly influential SNP positions.

Since the large majority of SNP positions on the genome won't be relevant to the particular disease being studied, there is a very small probability that any particular position chosen at random will be of interest. Tavaré and colleagues have taken advantage of this fact to identify the important interacting SNP positions.

The new technique is also more flexible than previous ones. Many traditional regression methods are additive, meaning that the effects of individual SNP mutations are just stacked together, ignoring the possibility that the effect of one mutation might influence the effect of another. Other methods that do allow interactions must also place restrictions on the dataset, some of which may not be suitable in all situations. Tavaré's method seeks to do away with these restrictions by partitioning mutations into interacting groups, then using Bayesian methods to identify those groups with the greatest influence in predicting disease.

References

Senn, S. (2004) Individual response to treatment: is it a valid assumption? *British Medical Journal,* 329(7472), 966-968. DOI: 10.1136/bmj.329.7472.966

Speed, D. & Tavaré, S. (2010) Sparse partitioning: nonlinear regression with binary or tertiary predictors, with application to association studies. *Annals of Applied Statistics* (in press).

EPSRC Grants

Reference: EP/E018173/1
Title: Simplicity, Complexity and Modelling

The IMA would like to thank Professor Simon Tavaré, University of Cambridge and Professor Stephen Senn, University of Glasgow for their help in the preparation of this document