Biomathematics & Statistics Scotland

David A. Elston

Biomathematics & Statistics Scotland (BioSS, www.bioss.ac.uk) delivers research, consultancy and training in statistics, mathematical modelling and bioinformatics. BioSS plays a distinctive role in the Scottish research community, bridging the gap between the development of mathematical methods and their application in addressing important scientific problems. Our staff work in four broad application areas: plant science; animal health and welfare; ecology & environmental science; and human health & nutrition, and we collaborate widely with scientists from many organisations in Scotland and overseas. BioSS's programme of applied strategic research addresses generic issues encountered in these application areas and is managed in three broad themes: statistical bioinformatics; process & systems modelling; and statistical methodology.

BioSS was formed as the Scottish Agricultural Statistics Service in April 1987. This new organisation combined the staff of the AFRC Unit of Statistics in Edinburgh with statistical staff working in agricultural research institutes throughout Scotland. The critical mass of this new structure allowed individuals to specialise in emerging methodologies and application areas. A strong corporate identity was quickly established by the Director, Rob Kempton, leading to a willing sharing of expertise and a mutual desire to improve scientific research through cutting-edge quantitative techniques.

Although BioSS interacts equally with all of the Scottish Government's Main Research Providers, it has always been considered too small to be free-standing and so was established as a distinct unit within SCRI. This arrangement, overseen by a Strategic Planning Group whose membership represents all of BioSS's key stakeholders, has withstood the test of time: we look back proudly on our achievements as part of SCRI and look forward optimistically to our future as part of The James Hutton Institute.

Phylogenetic trees and molecular evolution

Frank Wright

In 1993 the SCRI annual report contained an article by F. Wright and R.A. Kempton entitled 'Phylogenetic trees and molecular evolution'. This marked the beginning of the collaboration of BioSS with SCRI in analysing molecular sequence data to infer relationships among species and aspects of the evolutionary process. The BioSS involvement widened from providing consultancy, to developing statistical methods to analyse multiple alignment data and then to developing software to make the use of these sophisticated methods easier for biologists.

Over this period, the methods to infer phylogenetic trees have increased markedly in statistical sophistication. In the early 1990s Maximum Parsimony and methods based on genetic distance matrices (for example, Neighbour Joining) were common. In those days, the slowness of modern statistical approaches was a major issue. However, in the last five years we have seen a dramatic increase in the use of fast Maximum Likelihood methods and in the use of Bayesian methods which can also make use of computer cluster technology. The availability of increased computing power has also allowed the analyses of large numbers (e.g. thousands) of gene sequences.

To simplify the use of modern statistical and evolutionary phylogenetic methods and to make use of computing power, we developed the TOPALi package for the analysis of nucleotide and protein sequence multiple alignment data. TOPALi 2.0, released in 2009, includes access to methods for model selection (an important step prior to phylogenetic analysis) and phylogenetioc tree estimation including the well-known PhyML and MrBayes programs. TOPALi has a rich graphical interface. The figure below is the output from a TOPALi model selection analysis applied to a DNA dataset showing the optimum model of nucleotide substitution.

More recently BioSS has also provided consultancy in specialised phylogenetic analysis methods for RNA sequence data which take into account known RNA secondary structure (i.e. position of stem and loop regions) and also in evolutionary analysis methods to detect evidence of natural selection.

With the availability of genomic sequence data, we are now in the era of Comparative Genomics, and the emphasis is moving from Phylogenetics to Phylogenomics. The analysis of thousands of loci has created new challenges for the design of phylogenetic analysis protocols. Meeting these challenges will involve close collaboration between statisticians and bioinformatics programmers to automate the analyses and to visualise the output.