executive summary, 'The highlights of the genome', of the main findings from the draft sequence. It illustrates just how small is the apparent signal to noise ratio in the genome and explores its hierarchy of redundancy and duplication. It looks briefly at the controversy of the horizontal transfer of bacterial genes, and indicates how the draft sequence has been and will be used to track down the genes that are involved in disease.

A chapter of editorialised quotes, headlines and comment from the press is followed by a good section exploring some of the Ethical Legal and Social Issues that the human genome brings into current relevance. These include forensic testing and the proof of innocence, insurance and privacy, somatic and germline gene therapy, eugenics, tailored medicine, patenting and ownership. This chapter ends with a plea for more widespread, deeper and detailed education in the principles of genetics, statistics and science in general, so that the information emerging from the biotechnology revolution is put to use, not misuse.

The final editorial chapter looks to a post-genomic future. It has been commonplace to compare the Human Genome Project with the space programme. They differ in this important sense. Putting a dozen men on the moon between Neil Armstrong's giant step on 20th July 1969 and the departure of Eugene Cernan (who he?) on 10th December 1972 was in many ways (including the literal) a high point of human technological endeavour. It's really been a bit downhill and aimless since then. Looking back on June 2000 or February 2001 from 30 years in the future, I doubt that we will have similar feelings of anti-climax about bioinformatics, biotechnology and genome science. I am sure that the achievements of sequencing a composite human genome will look, if not trivial, at most a large early step in a continuing upward path. It is for this reason that it is worth commemorating the sense of achievement that we all feel now because, by the time we retire, it will be hard to credit how much effort it took to achieve so little.

Andrew Lloyd
Trinity College, Dublin

Phylogenetic Trees Made Easy: A How-To Manual for Molecular Biologists

Barry G. Hall
Sinauer Association Inc., Maryland, USA; ISBN 0 87893 311 5 (paperback); 179 pp.; US$27.95; 2001

Phylogenetics has come of age as a discipline in recent years. Increasingly sophisticated models of molecular sequence evolution are being developed. New statistical methodology is being applied to improve both the estimation of the phylogenetic tree, including the choice of the most appropriate evolutionary model, and the testing of the reliability of the tree. In addition, the use of the Maximum Likelihood framework means trees can be compared, allowing hypotheses such as host–parasite coevolution or the existence of the molecular clock to be tested. Phylogenetic analysis is routinely used in many areas of biology to study: the relationships among species; the evolution of a multigene family members (including detection of orthologues); the detection of recombination among viral strains; and the evolutionary process itself.

Getting started in phylogenetic data analysis can be a daunting task. There are now four main approaches and, in addition, parameter choices that allow very many variants of each approach. The number of software packages and programs has begun to grow at an increasing rate (see Joe Felsenstein's list of phylogeny programs at http://evolution.genetics.washington.edu). With

decisions to be made concerning software, method and parameter values, there is a lurking feeling that some of these choices may be crucial. So where does a newbie phylogenetic analyst start?

Barry Hall's aim is to guide an absolute beginner through all the steps necessary to construct a phylogenetic tree from a set of sequences. The book also sets out to provide some more detailed background for readers with modest experience. Data files and some additional software are available via the web.

The book consists of a short introduction (6 pages), six sections of varying lengths (from 5 to 62 pages) and two appendices. Section 1 (entitled 'Tutorial: create a tree!') is the longest, and deals with a range of tasks, including obtaining related sequences by BLAST search, creating a (protein) multiple alignment with CLUSTALX, and producing a 'best tree' from a protein sequence alignment by the Neighbor Joining distance method. The book uses the almost-comprehensive PAUP* phylogenetic package (version 4.0) for most analyses plus MrBayes (for Bayesian trees) and TREE PUZZLE programs (for protein Maximum Likelihood trees). PAUP* is a modestly priced commercial package (menu-driven for Mac; command-line for DOS and Unix). The freely available PHYLIP package is briefly mentioned as a possible alternative comprehensive package.

The initial data preparation tasks prior to multiple alignment are given in complete detail. Quite rightly, the need for a high-quality multiple alignment is stressed (garbage in, garbage out) and over 17 pages are devoted to the use of CLUSTALX (including optimal choice of amino acid substitution matrices). It should have been mentioned, however, that poorly aligned regions should be omitted from the phylogenetic analysis (for example, the first 15 amino acids of the protein alignment example appear to be misaligned).

What is surprising is that the protein sequence alignment is used in the subsequent distance-based phylogenetic analysis, given that PAUP* does not yet have a full set of protein analysis methods. PAUP* has a limited choice of protein distances (and no protein Maximum Likelihood method). A protein parsimony analysis would have been a better choice at this point in the book. It would be better to use the DNA alignment (the decision to use protein sequences is especially surprisingly as the author provides a program to create a DNA sequence alignment guided by the protein alignment). The choice of protein distance used is not discussed and the analysis appears to be based on a simple distance that treats all amino acid replacements as equally likely. On first reading, I assumed that this was simply an approximate tree to allow the reader to get started. However, there is no subsequent discussion on distance options in PAUP* that are used with either protein or DNA sequence data. My concern is that a beginner may conclude that this is the definitive way to carry out a Neighbor Joining analysis on a protein-coding gene. Section 1 finishes with very good coverage of tree interpretation, rooting and bootstrap analysis to determine the reliability of clusters.

Section 2 (45 pages) starts with a nine page discussion on which is the best method to use. The author stresses the similarities in accuracy among the four approaches rather than the differences. While he has a preference for the Bayesian approach, he makes no strong recommendation, and notes that it is often reasonable to use all four approaches. In terms of distance methods, the little-used UPGMA method is discussed, whereas Fitch–Margoliash and Minimum Evolution methods are not. Surprisingly, the use of weighting to improve parsimony, and of parameter estimation to improve distance methods and Maximum Likelihood are not discussed in the context of comparing methods.

Creating Parsimony, Maximum Likelihood (exact and approximate) and Bayesian trees is then covered. The

discussion of Maximum Likelihood and Bayesian trees (using the MrBayes package) is very good, although the excerpts from the MrBayes documentation will be hard for a beginner to follow. Nucleotide models and rate heterogeneity models are covered reasonably well. Unlike the Neighbor Joining and Parsimony analyses, PAUP* options are used to optimise the Maximum Likelihood analysis. The approximate Maximum Likelihood method (TREE PUZZLE program) is introduced for protein Maximum Likelihood analysis as PAUP* does not currently have this option. The TREE PUZZLE analysis is not optimised, with the rate heterogeneity option set at the default value (uniform, ie all sites evolving at the same rate). In addition, no mention is made of the approximate nature of the 'bootstrap' values produced by TREE PUZZLE.

The remaining sections are shorter and cover a range of smaller topics including data format conversion and presenting and printing trees.

'Phylogenetic Trees Made Easy' is a useful book for beginners in phylogenetics, but it is uneven in several respects, in particular in its coverage of the four approaches. Distance methods should have been given better coverage in the practical examples, both in the choice of the type of distance used and in the tree building method. In addition, only uncorrected parsimony is shown, with no mention of the problems with this approach when there are long branches in the underlying tree. With many scientists beginning to produce alignments containing large numbers of sequences (eg much greater than 50), better coverage of fast methods (ie distance and parsimony) would have been useful. In contrast, (exact) Maximum Likelihood analyses and Bayesian analyses are demonstrated well with the appropriate parameters optimised.

The nucleotide models and rate heterogeneity models used for distances (essentially maximum likelihood method applied to a two-sequence alignment), maximum likelihood and Bayesian analysis could all have been presented as a unified framework. This would have been helpful in the discussion of increasingly sophisticated models used by these three approaches as the book progressed. With regard to alignment, the clear explanation of many of CLUSTALX's features should lead to improved care in the production of multiple alignments.

Several areas of phylogenetics were probably omitted due to lack of space, but I would have liked some brief mention to alert the reader of their existence, eg statistical tests to compare trees, tests to choose the most suitable model of nucleotide substitution, and distance methods that were unaffected by significant variation in base composition among sequences. I would have preferred a slightly larger book with more real data analysis and more extensive details of the PAUP* package. I feel, however, that molecular biologists will find Barry Hall's book helpful in getting them started in phylogenetic analysis, and to use Maximum Likelihood and Bayesian methods well.

Frank Wright
Biomathematics and Statistics Scotland