

Some manuscript and provisional names are used. It will be good when this practice disappears from Australian plant publications.

There are no diacritics on authors' names, e.g. 'Ave-Lall' (p. 36), 'A. & D.Love' (p. 186), 'G.Lopez' (p. 288), 'L'Her.' (p. 316). A few typographical errors have slipped through ('who's' instead of 'whose', 'gadens' instead of

'gardens'), likewise there is some faulty punctuation (e.g. 'however' always used with no commas), and there are several discrepancies between the number of a subject and its verb.

Overall, however, this book brings together, successfully, a huge amount of data in a compact, user-friendly way and it deserves to do well. The price is very reasonable.

Bioinformatics

Review by David A Morrison

Department of Parasitology (SWEPAR), National Veterinary Institute,
751 89 Uppsala, Sweden

E-mail: David.Morrison@bvf.slu.se

Instant Notes: Bioinformatics. By David R. Westhead, J. Howard Parish and Richard M. Twyman. 2002. BIOS Scientific Publishers, Oxford. x+257 pp. ISBN 1-85996-272-6 £18.99 (paperback).

Bioinformatics for Dummies. By Jean-Michel Claverie and Cédric Notredame. 2003. Wiley Publishing, Hoboken, NJ. xx+452 pp. ISBN 0-7645-1696-5 £22.50 (paperback).

Associated web site:

www.dummies.com/WileyCDA/DummiesTitle/productCd-0764516965,page-1.html

Bioinformatics is an idea that originated in the late 1960s, when systematists first started seriously considering the practical application of computer technology to biology, in the form of specimen databases, interactive identification, and phylogenetic analysis. Subsequently, the widespread application to biology of the advances in information technology was probably inevitable, when computers became cheaper and more powerful during the 1970s. However, the idea of bioinformatics as a separate discipline had to wait for the development of molecular biology in the 1980s, when its full potential was realized in the mainstream disciplines of biology. Since then, bioinformatics has been the fastest growing field within the biological sciences, expanding rapidly during the 1990s to create a fusion of biologists, mathematicians and computer scientists the like of which had not been seen before.

This situation has, however, left the systematists in somewhat of an awkward position: although they can reasonably claim to be the founders of bioinformatics, and specimen databases, interactive identification and phylogenetic analysis have all developed along with the other parts of bioinformatics, many systematists are likely to feel that they have been left somewhat behind. After all, if the money that was spent on the Human Genome Project had been spent on biodiversity inventories instead, we would be a

long way towards having a global inventory by now, and the profession of systematics would be a lot healthier. This also means that keeping abreast of technological advances is becoming harder and harder for the average systematist; and it is hard enough just to understand what new possibilities are arising, let alone whether they might be of direct relevance to systematic research. Thus, it may be handy to have an introduction to the broader field of bioinformatics, so that when talking to a gene jockey or a computer whiz we can at least ask intelligent questions.

Sadly, there is no book called "Bioinformatics for Evolutionary Biologists / Systematists", even though other disciplines have such books; for example, geneticists have the book by Barnes and Gray (2003) and immunologists have the one by Lund et al. (2005). Therefore, there appears to be an unfilled niche here, which someone might like to fill, given that there seems to be a new book with the word "Bioinformatics" in the title published every month or so. In the meantime, systematists will have to delve into more generalist books in order to find what we are seeking. The two books reviewed here are intended to fill that more generalist niche.

While these books are not written specifically for systematists, they are both written for the type of person that a systematist might be: someone who knows quite a lot about biology and yet feels that they should know more about bioinformatics than they currently do but who doesn't (so far) feel the need to delve too deeply. This is a tricky sort of book to write, as there is a thin tightrope to walk between simplifying the information and being downright misleading. In fact, such authors probably can't win whatever they do, because any "expert" will always be able to see that their particular field has been "grossly distorted" during the abridgement.

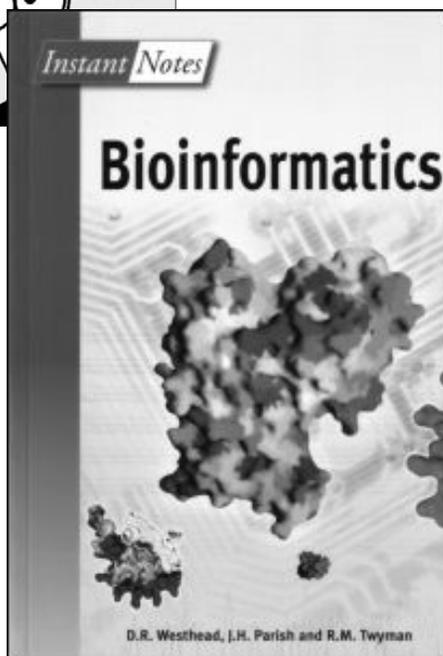
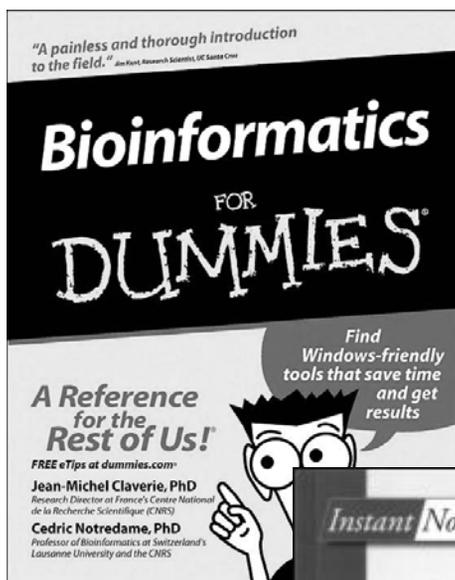
The format and style of each of the books is pre-determined by the series of which it is a part. The

Notes series is the more sober of the two, with the *Dummies* series having a tendency to sound like it doesn't take things too seriously — I am never sure who the “dummy” is supposed to be with this series of books: the reader, the writer, or even the publisher. The *Dummies* book is arranged as a series of practical exercises, which are best performed while sitting at an internet-connected computer, which gives it a strong flavour of a cookbook. Thus, a simplistic comparison might suggest that the *Notes* book is generally stronger on the theory while the *Dummies* book is generally stronger on the practice. However, this generalization is not always true, as the *Dummies* book gets the theory across although it does it in a very different way.

The books also differ in the depth and breadth of coverage of the topics. The *Dummies* book concentrates very much on amino acid sequences, covering sequence annotation, database searches, sequence alignment and comparison, and protein structure prediction. From there it expands out into nucleotide sequences, including both protein-coding and RNA-coding. The *Notes* book also covers these subjects, but it includes many other topics as well, such as microarrays, proteomics, molecular pathways, and what it chooses to call cheminformatics and pharminformatics (although the discussion of most of these extra topics is mainly about molecular biology rather than about bioinformatics itself). So, the *Notes* book is literally a notebook, covering a lot of themes but each one in less detail. It is also more pedagogically formal in its presentation (i.e. it is much closer to a teaching book), while the *Dummies* book allows itself to wander from topic to topic (but in a controlled manner).

Interestingly, there are many relevant practical topics that are also not discussed by either book. For example, neither book tells you that doing a database search using a nucleotide sequence may produce no matches while a search of the same data translated to amino acids can produce a lot of (correct) matches. Neither book talks about all of the molecular data for which databases do not yet exist, such as AFLP, ISSR, RAPD and microsatellites (unless the person has sequenced

each microsatellite for each sample). Similarly, neither book warns you about the practical details of computerized data analysis. The biggest bane of my computing life is dealing with the fact that every computer program treats “nexus format”, “phylip format”, “fasta format”, etc in its own unique way, so that moving a file smoothly from one program to another is the exception rather than the rule. Furthermore, neither book talks much about computing itself, as both of them assume that the reader will be dealing only with user-friendly programs, and in the *Dummies* case only with



online programs at that. In science, it can be a serious mistake to treat user-friendliness or accessibility as the primary criterion for choosing a data analysis.

Overall, I think that each book is successful at what it sets out to achieve. The authors clearly have chosen to emphasize different information, and they have taken two very different approaches as to how to present that information. Which book you might choose will depend on your own inclination — I would choose the *Notes* book if I wanted breadth of coverage, but the *Dummies* book if I wanted practical details about a narrower range of topics.

However, there can be two sources of general criticism that can be levelled at in more detail. The first was alluded to above: expertise in a field inevitably leads to criticism of any attempt to simplify that expertise. In this case, the weakest part of each book for a systematist will obviously be the presentation of phylogenetic analysis. In fact, if you were to judge the intellectual quality of either book by first reading the phylogeny chapters then you probably wouldn't bother reading the rest of the book, but would instead merely use it to prop open the kitchen door.

A few comparative quotations from the glossaries [with comments] will illustrate the sorts of difficulties that both books get into:

- *Phylogenetics. Dummies*: Reconstruction of the evolution [sic] history of a gene family based on gene comparison. [Apparently gene trees are all there is, rather than species trees.] *Notes*: The branch of science that deals with resolving the evolutionary relationships between organisms. [Better, but it doesn't distinguish population genetics from systematics.]
- *Clade. Dummies*: Group of related species and their common ancestors. [Doesn't distinguish paraphyly from monophyly.] *Notes*: A group of organisms descended from a particular common ancestor (i.e. an ancestor and all of its descendants). [Saved by the bit in brackets.]
- *Alignment. Dummies*: Representation of two or more protein or nucleotide sequences where homologous amino acids or nucleotides are in the same columns while missing amino acids or nucleotides are replaced with gaps. [This emphasizes the philosophy only, which is homology assessment.] *Notes*: Arrangement of two or more nucleotides of protein sequences to maximize the number of matching monomers. [This emphasizes the practice only, which is entirely similarity.]

Given this background, it will not surprise you to learn that the *Notes* book presents an "example of an actual cladogram" that is unrooted but on which "clades" are recognized (which flatly contradicts the given definition of a clade), and which has a branch missing so that the diagram is not completely connected (and the *Dummies* book apparently had similar sorts of problems in the early printings, which have since been corrected). I won't list the other infelicities in the two books, or I might start foaming at the mouth. I will merely point out that the rest of each book is better than its phylogenetics section (and the same can be said of the book by Lund et al., 2005, which is actually much worse than these two). My main concern in raising this issue is

that these books probably reflect a view of phylogenetics that is quite widespread, where expertise in phylogenetic analysis is not credited to systematists as a group but is seen as being a general skill that anyone can acquire without much effort. The quality of phylogenetic analyses in the scientific literature is not always high, presumably because many practitioners do not realize the effort that is needed in order to do a good job. There is a real danger that too much emphasis on bioinformatics could turn biology from a science into a series of algorithms.

The second general criticism is that these two books take an entirely positive attitude to their subject, dwelling lovingly on the prospects and opportunities created by bioinformatics, so that both books come across as almost entirely uncritical. It is essential in science, in particular, that quality be assessable and assessed, but neither of these books addresses the issue of how this might be done using the bioinformatics tools that they extol and the data sources that they recommend. The best that either of them does is to have occasional reminders that database searches, for example, are not always reliable, along with some hints as to what to be wary of. My point should be seen as a general criticism of bioinformatics, in its current state of development, rather than a specific criticism of these particular books — the books are merely reflecting the general attitude of far too many bioinformatics practitioners, which is that quality control is someone else's problem. There is a strong tendency in bioinformatics to produce what mathematicians refer to as "analyses untouched by the human mind", and even to present this as a desirable thing. I don't see how any serious scientist could agree with this attitude.

The lack of critical thinking can also be seen in a number of university bioinformatics courses, especially those based on the principle of using on-line servers for data analysis. Many of these courses (especially short courses) simply equate computer programs with their analyses, and therefore deal only with those *programs* they consider to be "popular" rather than with which *analyses* are "best". They thus concentrate on what resources are available rather than on any assessment of the quality or usefulness of those resources. Their emphasis on "quick" methods only gives beginners a wrong impression, emphasizing the need to get a result rather than to be judicious about that result. In this environment, we cannot expect much in the way of a critical approach to data and analysis in bioinformatics. This attitude is not unusual in biology, of course, being also shown in the popularity of Microsoft Excel as a tool for scientific data analysis in spite of a long series of

published critical evaluations demonstrating its inaccuracies and inadequacies, thus reflecting the primacy of convenience over quality for many people.

This leads me to wonder exactly what it is that scientist users of bioinformatics are going to do about quality control, since the expert practitioners and developers have so far shown little inclination to do very much. We, also, have done relatively little so far, and yet we must do something if we are to benefit from bioinformatics without being lead astray. For example, we cannot simply download data from a database, combine it with our own data, and then produce a phylogenetic analysis. This is because most databases are uncurated, which means that there is no quality control whatsoever, and so you can't tell the good information from the bad. Every bioinformatics project on which I have worked has required me to spend as much time checking inconsistencies and contradictions in the downloaded data as it has collecting the new data, which seems almost to defeat the purpose of having the database in the first place. Even curated databases have their problems (e.g. I once found a gross error in a structure in the Rfam database), but then at least there is someone you can contact about it to have the problem fixed. In introductory books such as the ones discussed here, a chapter on quality control would add a suitable sobering note to the general euphoria, without necessarily detracting from the overall positive ambience.

For anyone wishing to pursue bioinformatics in more detail, the next step up from these two books consists of undergraduate textbooks. The two currently most popular books are those of Higgs and Attwood (2005) and Lesk (2005), although the book by Gibson and Muse (2004) contains much that will be of interest to evolutionary biologists. The next step up from there will take you into the specialist literature, where there is a rapidly expanding number of books available on various topics (especially those related to computing itself). Sadly, none of these books recognize the role that systematists played in the origin of bioinformatics, as this word is now seen as being synonymous with "molecular bioinformatics", the rest presumably being seen as "taxonomic computing" (a term that should not be despised, given the ongoing importance attached to the electronic dissemination of systematic data). Nor do any of the books present things in a way that would be directly relevant to practical phylogenetics studies. This seems to leave an unfilled niche in the publishing world, which one of you might feel inspired to fill.

References

- Barnes, M. R., and I. C. Gray (eds). (2003) *Bioinformatics for geneticists*. John Wiley and Sons, New York.
- Gibson, G., and S. Muse. 2004. *A primer of genome science, second edition*. Sinauer Associates, Sunderland, MA.
- Higgs, P. G., and T. K. Attwood. 2005. *Bioinformatics and molecular evolution*. Blackwell Publishing, Oxford.
- Lesk, A. M. 2005. *Introduction to bioinformatics, second edition*. Oxford University Press, Oxford.
- Lund, O., M. Nielsen, C. Lundegaard, C. Kesmir, and S. Brunak. (2005) *Immunological bioinformatics*. MIT Press, Cambridge, MA.

Nitrogen fixation in Acacias

Dan Murphy

Royal Botanic Gardens Melbourne

***Nitrogen Fixation in Acacias: an untapped resource for sustainable plantations, farm forestry and land reclamation.* By John Brockwell, Suzette D. Searle, Alison C. Jeavons and Meigan Waayers. ACIAR Monograph No. 115, 132 pp. Published by Australian Centre for International Agricultural Research (ACIAR), 2005. ISBN 186320489.**

This book is a review of the literature dealing with the associations between acacias and root-nodulating rhizobial bacteria, but it also includes original observations and interpretation of this literature. A particular emphasis has been placed on an examination of the symbiosis between legumes and rhizobial bacteria, the ability of this symbiosis to fix atmospheric nitrogen, and how this may be enhanced and utilised.

After a brief introduction (Chapter 1), I initially thought that the chapters of most interest to readers of this newsletter would be chapters two and three, which focus on the plants and bacteria, including their taxonomy and systematics. However, these chapters only make up a minor proportion of the book, and are not the most readable parts. The chapter on the acacias (Chapter 2 – The Plants) is relatively brief, and the discussion of taxonomy a little confusing because it does not explain why taxonomic changes are necessary in *Acacia*. I thought it may have been useful to include the implications of the non-monophyly of *Acacia sensu lato* and how this has impacted on the understanding of the patterns of nodulation and symbiosis. For example, now that we know that members of the mimosoid legume tribe Ingeae (including genera