

Bioconsensus.—

Janowitz, M. F., Lapointe, F.-J., McMorris, F. R., Mirkin, B., and Roberts, F. S., editors. (DIMACS series in discrete mathematics and theoretical computer science, v.61) 2003. American Mathematical Society. 242 pp. ISBN 0-8218-3197-6. \$75.00

“Consensus is what many people say in chorus but do not believe as individuals.”
Abba Eban (1915-2002), Israeli diplomat.

Group decision making is as old and as ubiquitous as human societies. Should the gladiator live? How much harvest should be stored for the winter? Which is the best movie of the year? Who should be the Democratic presidential nominee? Which is the true evolutionary tree? The formal theory of voting and social choice dates back to the eighteenth century members of the French Academy of Sciences, Marquis de Condorcet (Condorcet, 1785) and de Borda (de Borda, 1784). The Rev. C. L. Dodgson, better known as Lewis Carrol, “wrote extensively on committees, elections, and proportionate representation” (Black, 1958). The modern developments in the field originate in Kenneth J. Arrow’s 1951 doctoral thesis (Arrow, 1963), who developed a mathematical system to consider possible voting schemes and showed that any group choice system is, sadly, either inconsistent, arbitrary, or unstable. The theory of combining conflicting choices and rankings into a “representative” object spans many areas of mathematics and, recently, computer science.

In the past fifteen years the mathematical and computational techniques developed in the context of group choice and consensus decisions have started to be applied to biological problems, mainly in systematics, taxonomy, and phylogenetics. This is not to say that until then systematists faced with the task of combining multiple phylogenetic hypotheses just threw their hands up in despair. However, much of it was done ad hoc, using an implicit heuristic and expert knowledge and intuition. Since computers and automated methods have started to be used in phylogeny reconstruction, analysis of hundreds of taxa (Bush et al., 1999; Soltis et al., 1999; Savolainen et al., 2000; Berbee, 2001) has become commonplace and phylogenies with thousands of taxa (Källersjö et al., 1998) are regularly attempted. Many computational approaches use some optimization criterion, such as maximum parsimony and maximum likelihood, to evaluate possible phylogenies. As a result,

they can return hundreds or more (e.g., Maddison et al., 1992) equivalently good trees. Such answer is quite unsatisfying, so in the final stage these trees are combined into a “representative” tree. Computers cannot use intuition to do this, so there was a need for a formal tree consensus method. While some formal consensus methods already existed (Adams, Nelson, strict, semi-strict, and majority rule), these methods are fairly limited, and there were no mathematical or practical guarantees on their topological accuracy. The mathematical and computational field of BIOCONSENSUS explores different options for combining biological data (mainly phylogenetic trees at this point), establishes a formal framework to develop new consensus methods that stress various aspects of the data, compare their relative merits, and evaluate their practical performance.

In September 2000, the Center for Discrete Mathematics and Theoretical Computer Science¹ (DIMACS) began its four-year focus on Computational Molecular Biology². This program is bringing together mathematicians, statisticians, computer scientists, and biologists to collaborate on various aspects of molecular biology. The book “Bioconsensus” is the result of two working group meetings on Bioconsensus held at DIMACS in 2000 and 2001. It is a collection of mathematical and computer science papers on consensus methods in various biological applications. The papers are organized into three parts: the very theoretical mathematical foundations of phylogenetic consensus theory, computational consensus methods and algorithms in various areas of biology, and practical considerations for the consensus techniques in phylogeny reconstruction.

Part I consists of five mathematical papers. This section requires more than just a passing knowledge of mathematics and, in the first article’s authors’ own words, “a tolerance for mathematical abstraction”. There are whole pages covered with mathematical symbols. The first paper (by Day and McMorris) puts the subject of consensus of phylogenetic hypotheses in the broader and older context of the social choice theory. It gives a historic and mathematical overview of the main results and techniques in this area and shows the mathematical consequences

¹A consortium of Rutgers University, Princeton University, AT&T Labs, Bell Labs, NEC Laboratories America, and Telcordia Technologies, founded as a National Science Foundation Science and Technology Center

²The special focus on Computational Molecular Biology is a continuation of the DIMACS 1994–2000 special focus on Mathematical Support for Molecular Biology

of those for the phylogenetic consensus. A very good historic overview of social choice theory and the social and legal consequences of Arrow's (Condorcet's) voting paradox is given in a more accessible language in (Block, 1998). The second (by McMorris and Powers) and third (by Powers) papers of "Bioconsensus" extend two specific Arrow-type impossibility results from the social choice theory to the context of combining phylogenetic trees. That is, they prove that if some reasonable conditions on a tree consensus method are imposed, then there exists an input tree with more "influence" over the outcome than the rest of the input trees. The fourth paper (by Bryant et al.) is very different. It gives an upper bound on the expected size of the a maximum agreement subtree of two input trees, randomly generated either under uniform or Yule-Harding models. The final fifth paper (by Crown and Janowitz) in this section goes back to the axiomatic representation of consensus methods. The authors state a mathematical condition that the properties of a consensus rule must satisfy to ensure that a consensus object exists. Overall, Part I of "Bioconsensus" is aimed at mathematicians that work or have a desire to work on bioconsensus problems. It provides the mathematicians with the abstractions of the biological problems, known results, and open problems. A mathematician can read this section and start working in this area without any prior biological knowledge or consulting any biological papers or experts.

Part II of the book is more computational. It deals mainly with algorithms for specific biological problems involving various types of consensus methods and the papers spend more time on biological motivation. The five papers in this section use very diverse mathematical and computational techniques to solve very different biological problems. A graph theoretic model is proposed to detect and eliminate errors in physical mapping of DNA; a hierarchical clustering algorithm is used to build evolutionary trees (based on full genomes); a genetic algorithm attempts to categorize an organism's life-cycle; geometric pattern matching techniques are employed in the comparison of long DNA sequences; and classical combinatorial optimization techniques are the basis for new phylogenetic supertree method. The section is definitely more about breadth than depth. It presents a sample of biological problems that use consensus methods in some way and can be solved computationally.

The third and final part of the book is titled "Practical Considerations", although it would be better

named "Phylogenetic Consensus". It contains six papers that discuss consensus and supertree methods for phylogenetic trees. All the papers use less mathematical notation than the ones in the previous sections and are readable by non-mathematicians. The first paper (by Bryant) is a mathematical survey and classification of existing consensus methods. It is, like the previous articles, written for mathematicians: in the terminology section the terms "taxa" and "monophyletic" are defined, while "tree leaf" and "most recent common ancestor" are not. However, this is the only existing comprehensive survey of the phylogenetic consensus methods. It is clearly written and fairly accessible, with many examples. The second paper (by Thorley and Wilkinson) is a review of supertree methods. This is the first paper in this volume that explicitly addresses (although very briefly) *biological* considerations when using a particular method. The third paper (by Wilkinson and Thorley) discusses the reduced consensus method, the only consensus method that, to resolve conflicts in the input trees, may return a tree that does not contain all the input taxa. While it is stated as a consensus method, it is more appropriate viewed as a supertree construction method. The fourth paper (by Lapointe and Cucumel) is a review of statistical methods for assessing the quality and reliability of a consensus tree. It addresses such issues as consensus of trees with branch lengths, testing for the randomness of a consensus tree, support for existing branches, etc. The paper contains examples to illustrate the concepts discussed and requires only a practitioner's knowledge of statistics (rather than a statistician's). The last two papers of the book use simulations to assess the accuracy of a particular consensus method. The penultimate paper (by Levasseur and Lapointe) is concerned with the average consensus, the only consensus method that takes branch lengths into account. The final paper (by Bininda-Emonds) deals with the MRP supertree construction technique viewed as a consensus method. A major experimental design flaw of both papers is that the trees used are very small (10 taxa in the first paper and 8 or 32 in the second) and very few input trees are considered (2 or 10). Thus, while the conclusions of both papers may be valid (branch lengths need to be taken into consideration when combining trees and MRP consensus is about as topologically accurate as Majority Rule), better designed large-scale experiments must be performed to state those with certainty. Overall, this section is a good *tour de force*

of the phylogenetic consensus and supertree methods and the practical computational and statistical tools available in this area. The three review papers contain a comprehensive list of references an interested reader can use to obtain more information. What is glaringly missing is the discussion on the biological merits of the various consensus and supertree methods and the validity of the use of these methods in general. A more in-depth treatment of phylogenetic supertree methods is about to appear in the book “Phylogenetic supertrees: the book” by Kluwer Academic Publishers.

“Bioconsensus” presents mathematicians and computer scientists with a new application area for consensus research, provides them with sufficient background, a sample of results and important open problems. Only the last section of the book is of interest to anybody who is concerned with practical phylogenetic analysis. However, I think this is an important book that can stimulate collaborative research between mathematicians, statisticians, computer scientists, and biologists. Hopefully, as the field of bioconsensus develops, the ideas presented in this volume will become chapters in various interdisciplinary textbooks.

References

- ARROW, K. J. 1963. Social choice and individual values. John Wiley, New York, second edition.
- BERBEE, M. L. 2001. The phylogeny of plant and animal pathogens in the Ascomycota. Physiological and molecular plant pathology.
- BLACK, D. 1958. The theory of committees and elections. Cambridge University Press, Cambridge.
- BLOCK, C. D. 1998. Truth and probability – Ironies in the evolution of social choice theory. Washington university law quarterly, 76:975–1037.
- BUSH, R. M., FITCH, W. M., BENDER, C. A., AND CO, N. J. 1999. Positive selection on the H3 hemagglutinin gene of human influenza virus A. Molecular biology and evolution, 16:1457–1465.
- DE BORDA, J.-C. 1784. Mémoire sur les élections au scrutin. Histoire de l’académie royale des sciences.
- MARQUIS DE CONDORCET (M. J. A. N. DE CARI-TAT) 1785. Essay on the application of analysis to the probability of majority decisions.
- KÄLLERSJÖ, M., FARRIS, J. S., CHASE, M. W., BREMER, B., FAY, M. F., HUMPHRIES, C. J., PEDERSEN, G., SEBERG, O., AND BREMER, K. 1998. Simultaneous parsimony jackknife analysis of 2538 rbcL DNA sequences reveals support for major clades of green plants, land plants, seed plants and flowering plants. Plant systematics and evolution, 213:259–287.
- MADDISON, D. R., RUVOLOVO, M., AND SWOFFORD, D. L. 1992. Geographic origins of human mitochondrial DNA: Phylogenetic evidence from control region sequences. Systematic biology, 41 (1):111–124.
- SAVOLAINEN, V., CHASE, M. W., HOOT, S. B., MORTON, C. M., SOLTIS, D. E., BAYER, C., FAY, M. F., BRUIJN, A. Y. D., SULLIVAN, S., AND QIU, Y. L. 2000. Phylogenetics of flowering plants based on combined analysis of plastid atpB and rbcL gene sequences. Systematic biology, 49: 306–362.
- SOLTIS, P. S., SOLTIS, D. E., AND CHASE, M. W. 1999. Angiosperm phylogeny inferred from multiple genes as a tool for comparative biology. Nature, 402:402–404.
- Tanya Y. Berger-Wolf, Laboratory for High Performance Algorithm Engineering and Computational Molecular Biology, Department of Computer Science, University of New Mexico, Albuquerque, NM 87131, USA.*