



PLANT SYSTEMATICS WORLD

Edited by Vicki A. Funk

■ A TRIBUTE TO PETER SNEATH

I never met Peter Sneath (17 Nov. 1923–9 Sept. 2011), but I have been hugely influenced by his work—perhaps to the point that without it, I wouldn't be doing what I am doing. This is also true of many others working in systematics, ecology, bioinformatics and biology in general.

Sokal & Sneath (1963) is an inspiring polemic, putting forward the idea that, in classification, we ought to be able to explain what we are doing. A reflection on three decades of its influence is provided by Sneath (1995). But it is Sneath & Sokal (1973) that I pored over as a student, and relied on for an undergraduate project in species delimitation from herbarium specimens (supervised by David Mabberley). I became slightly obsessed with the clarity and broad applicability of the methods. I horrified historians by suggesting a cluster analysis of World War II leaders (to decide whether Hitler or Stalin was 'worse'), anthropologists by suggesting a cluster analysis of human races based on morphology, and a botanical artist by suggesting we should represent each plant as a vector of numerical character states. The first idea now seems unworkable—I don't think evil lends itself to quantification. The last two ideas were not new, though in the 1990s were unfashionable, particularly any attempt to quantify 'race'. These days human race is a major area of research, though tending to be based on genetic data and aiming to understand origins and dispersal, rather than develop a classification (e.g., Chaubey & al., 2011).

If the Kew Index, funded by money left by Darwin, is the first bioinformatics database, then Sneath & Sokal (1973) is the first bioinformatics textbook. It is still one of the best. The clarity and notational unity make it timeless and comprehensible to anyone from a keen biology undergraduate to a professor. It is definitive. I refer postgraduates using clustering, ordination, and measures of similarity or distance to this book. Anyone wanting to know what sequential, agglomerative, hierarchical, nonoverlapping (SAHN) clustering methods are, for example UPGMA, ought to read it. As a Ph.D. student, I invoked Sneath & Sokal (1973) in persuading a senior academic to change the wording of an evolutionary practical hand-out, because the clustering method used—based on an erroneous, major textbook—was not UPGMA as advertised, but something more like UPGMC.

There are numerous occasions where people wish to group items according to similarity, and the influence of the book reaches far beyond biology. I was pleased to see it referred to in a recent textbook on chemoinformatics, for example (Leach & Gillet, 2007).



Peter Sneath in 2007.

In systematics in the 1980s, Sneath & Sokal's 'phenetic' approach became a target of hostile cladists, preferring to group organisms by presumed common descent rather than overall similarity. One argument against phenetics is that there are many clustering methods, and the choice between them is subjective; but there is only one phylogeny, so using it for classification is not arbitrary. At the time, it may have seemed there was only one way to reconstruct phylogeny—in practice Fitch parsimony (Fitch, 1971)—although distance and maximum likelihood methods had also been proposed (see Edwards, 2009). When reconstructing phylogeny today, there is certainly more than one method that can be argued to be 'best'. The main row now is between those preferring maximum likelihood with bootstrap support, on the one hand, or Bayesian Markov chain Monte Carlo, on the other. This is a specific instance of a century-old debate in statistics (e.g., Edwards, 1992), and is unlikely to be resolved in a simple way. Other decisions may include DNA vs. protein sequence, various details of the model of substitution, and, having decided what to aim for, which heuristic short-cuts to use. As with clustering, the choice of phylogeny reconstruction method has a subjective component. But these methods are all seeking to group by common ancestry, which is a real phenomenon, however difficult it may be to perceive. In phenetics, the very starting point of 'overall similarity' is subjective. The issue of taxonomic rank has



Peter Sneath and Robert Sokal in 1990.

been addressed inconclusively by both phenetic and phylogenetic approaches.

It is worth mentioning pattern cladism. This branch in the history of taxonomy invokes a phenetic justification for the use of parsimony, as the most logical way of grouping taxa according to their character states (e.g., Siddall & Kluge, 1997). This separation of organismal history from systematics is also a tribute to the phenetic taxonomic ideals of Sneath & Sokal (1973), although few pattern cladists would regard it as such.

It is interesting, too, that one of the original aims of SAHN clustering *was* phylogeny reconstruction (Michener & Sokal, 1957).

Though not now considered the preferred method, what is remarkable is that such half-century-old techniques are still used now and then, for example UPGMA by Lemberg & Freeman (2007). One attraction may be the incorporation of a molecular clock assumption, allowing interpretation of branch lengths in terms of time. This is also the greatest weakness of the method, because a molecular clock assumption is rarely valid. However, phylogenies with dated nodes and variations on a molecular clock assumption remain widely used. We are keenly interested in origins and timing and are willing to take some ‘hit’ in terms of accuracy to obtain them, though perhaps not such an extreme ‘hit’ as implied by UPGMA. Putting dates on nodes in a phylogeny remains methodologically difficult, and an active area of research. Though superseded for this purpose, it is clear that SAHN clustering was a pioneering approach in statistical phylogenetics.

Yet even in the most modern phylogeny reconstruction, phenetics retains a crucial role at early stages. For molecular phylogeny, where do the sequences come from? Often from a search of the public sequence databases using BLAST (Altschul & al., 1997). BLAST is a program that seeks maximum similarity between a query sequence and each sequence in a database; high similarity is taken as evidence of homology. This is a direct use of overall similarity as a proxy for homology. Even for obtaining novel sequences in the laboratory, we rely on the similarity of primer-pairs to amplify homologous sequences, which involves the same assumption.

Of course, there is more to systematics than phylogeny. Phenetic techniques, particularly ordination, are widely used in species delimitation, population genetics and ecology. Cluster analysis has experienced a total renaissance, being one of the major tools for analysis of high-throughput mRNA expression levels. In this



Richard Cowan, Peter Sneath and M.C. Vaughn at the Congress of Systematic and Evolutionary Biology, Boulder, Colorado, U.S.A., in 1973.

context, agglomerative clustering has been reintroduced to our undergraduate biology curriculum.

Peter Sneath is one of several heroes of his generation who have died recently who changed the world but whose recognition—though received—was not so enormous as the contribution. In a sense this is the highest tribute. Whoever invented the chair is unacknowledged, but our lives would be very different without it. For my own research in evolutionary bioinformatics, others in this category include Walter Fitch and Dennis Ritchie, without whom the field would be very different. In all three cases, I suspect it was their modesty and devotion to the subject that left them slightly out of the public eye. They certainly inspired those touched directly by their work.

Peter Sneath lived to see the end of the ‘cladistics wars’, the rise of computational methods and bioinformatics, and the explosion in use of his methods in a range of fields, if not always those originally envisaged. Most older arguments in systematics, based on authority and experience, have been replaced with arguments concerning data and methods of analysis which, though always flawed, are at least repeatable. It is as a pioneer of this ‘operational’ approach, more than for advocacy of any particular methods, that Peter Sneath has irreversibly and beneficially changed biological systematics.

Literature cited

- Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucl. Acids Res.* 25: 3389–3402.
- Chaubey, G. & al. 2011. Population genetic structure in Indian Austroasiatic speakers: The role of landscape barriers and sex-specific admixture. *Molec. Biol. Evol.* 28: 1013–1024.
- Edwards, A.W.F. 1992. *Likelihood*, exp. ed. Baltimore: Johns Hopkins University Press.
- Edwards, A.W.F. 2009. Statistical methods for evolutionary trees. *Genetics* 183: 5–12.
- Fitch, W.M. 1971. Toward defining the course of evolution: Minimum change for a specific tree topology. *Syst. Zool.* 20: 406–416.
- Leach, A.R. & Gillet, V.J. 2007. *An introduction to chemoinformatics*, rev. ed. Dordrecht: Springer.
- Lemberg, M.K. & Freeman, M. 2007. Functional and evolutionary implications of enhanced genomic analysis of rhomboid intramembrane proteases. *Genome Res.* 17: 1634–1646.
- Michener, C.D. & Sokal, R.R. 1957. A quantitative approach to a problem in classification. *Evolution* 11: 130–162.
- Siddall, M.E. & Kluge, A.G. 1997. Probabilism and phylogenetic inference. *Cladistics* 13: 313–336.
- Sneath, P.H.A. 1995. Thirty years of numerical taxonomy. *Syst. Biol.* 44: 281–298.
- Sneath, P.H.A. & Sokal, R.R. 1973. *Numerical taxonomy: The principles and practice of numerical classification*. San Francisco: W.H. Freeman.
- Sokal, R.R. & Sneath, P.H.A. 1963. *Principles of numerical taxonomy*. San Francisco: W.H. Freeman.

Daniel Barker

Centre for Evolution, Genes and Genomics, School of Biology,
University of St Andrews, St Andrews, Fife, KY16 9TH, U.K.
db60@st-andrews.ac.uk