J. Zool. Syst. Evol. Research 33 (1995) 42–47
© 1995 Blackwell Wissenschafts-Verlag, Berlin
ISSN 0947-5745

Received on 27 December 1994

Review

Department of Systematics and Animal Morphology, University of Bielefeld, Bielefeld, Germany

On the information content of characters in comparative morphology and molecular systematics

J. W. WÄGELE

Abstract

A review of the fundamental difference between single molecular-sequence positions, or numerical characters, and complex morphological characters is the subject of this study. It has been found that transformation series of single complex structures contain enough information to allow *a priori* determination of character order and that rooting of a dendrogram is possible without out-group comparison, while trees based on less-informative characters can usually only be rooted with out-group comparison. Furthermore, the quality of total information used is decisive in discriminating between hypotheses of relationships. Numerical methods for the inference of phylogenies have been found to be useful for high numbers of characters that have only a low information content, while the Hennigian procedure seems to be preferable or complex characters.

Key words: phylogenetic systematics – molecular systematics – character weighting – rooting of dendrograms – character-state polarity

1 Information contained in characters

DNA sequences contain information for the function and ontogenetic growth of cells and organs, but also traces that document phylogeny. Phylogenetically informative sequence fragments are not necessarily coding fragments, i.e. an organism contains more phylogenetically informative sequence fragments than coding fragments. In the following text, 'information content' refers to the information useful for phylogenetic studies. Since such information is also available from morphological characters, this study questions whether there is a fundamental difference between molecular and morphological homologies, but does not attempt to quantify the information content of characters. Although the number of putative synapomorphic positions could be counted, a comparable method of quantifying the value of morphological characters is not available. This paper was neither designed to compare all the advantages and pitfalls of molecular and morphological approaches (e.g. Hillis 1987; Wägele 1994), nor to present statistically testable ideas, but to draw attention to the fact that differences in the total information content used for an analysis are as important as differences between types of characters and types of cladistic analyses. To cite Charles Darwin (1871; 1974 reprint, p. 144), '...numerous points of resemblance are of much more importance than the amount of similarity or dissimilarity in a few points...'. Hennig (1950) spoke of the importance of the complexity of characters.

Typical characters used in phylogenetic analysis are molecular sequences, genes or the expressions of genes, and common functional units that represent only a very small proportion of the genome of a species. Complex characters can be divided into smaller, simpler characters, with less information. With the breakdown of complex characters, a subunit loses information-content quality, and its value for phylogenetic analysis and the estimation of probability that similarity indicates homology (Dohle 1989) is diminished if this subunit is considered as a single independent character. A substitution is the smallest possible apomorphy, while an unmodified sequence position is the smallest plesiomorphy of a descendant. By contrast, complex characters simultaneously contain, in comparison to an ancestral character, plesiomorphies and apomorphies of their components. One can safely assume that significant visible change in a genetically coded morphological structure is caused by a large number of substitutions at the molecular level (exceptions: e.g. cases where regulatory genes are involved).

It is safe to postulate that the genetic and phylogenetic information content of a complex morphological character is much higher than that of a sequence position. The latter unit – the smallest unit in systematics – is roughly analogous to physicists' quantum in that it possesses a quality of 'fuzziness', i.e. it is the most suitable unit for quantitative studies, but also a feature for which homology can not be established without additional information, in contrast to 'good' (complex) morphological characters. The highest information content is found in the complete organism. Between nucleotide and organism, all other genetic, biochemical, ultrastructural, histological or morphological characters are of intermediate quality (Fig.1).

Of course, a complex character is of no use if its information content and details are not explored. Superficial character analyses lead to erroneous hypotheses of homology, which is a major source of error in cladistic analysis (Wägele 1994). However, a quantitative description of differences between two character states is not possible in comparative morphology (Patterson 1988) without knowledge of the coding sequences and the factors influencing morphogenesis (Dohle 1989; Ingber 1993).

Nevertheless, when comparing homologous organs, the difference in complexity can often easily be estimated qualitatively: the pinhole eye of *Nautilus* is obviously simpler than the lens eye of *Sepia*; the metameric appendages of a polychaet

U.S Copyright Clearance Center Code Statement: 0947-5745/95/3301-0042$11.00/0

J. Zool. Syst. Evol. Research 33 (1995) 48
© 1995 Blackwell Wissenschafts-Verlag, Berlin
ISSN 0947-5745

Book Reviews

HEYER, W. R.: **Variation within the *Lepodactylus podicipinus-wagneri* Complex of Frogs (Amphibia: Lepodactylidae).** Smithsonian Contributions to Zoology 546 (1994). 124 pp., 46 figs, 55 tabs.

The aims of this paper are to describe the morphological variation within the *Lepodactylus podicipinus-wagneri* complex, reinterpret species limits, identify problem areas, and suggest aproaches that might provide resolutions. Over 6200 adults and juveniles were examined and variation analysis was performed on data from over 3000 specimens. The data set consists of a series of three morphological characters, of advertisement calls and habitat data. The morphological data are inadequate to perform a robust cladistic analysis as the data are very difficult to categorize in distinct, polarized states. The advertisement calls of the *podicipinus-wagneri* complex are not as distinct as those found in other species groups. Individuals of this complex demonstrate a broad array of vocalization types, the function of which is not well understood at present. Individual males are capable of producing distinct advertisement calls, but it is not clear which are which for all species. For most of the specimens it was possible to delimit species. The available material from most of Venezuela is inadequate to evaluate how many species occur there and which of them are conspecific with geographically adjacent species. Thirteen species are diagnosed as the result of the study, including a description of five new species. The distribution of most taxa within the *L. podicipinus-wagneri* complex are expected to be modified significantly by newly collected specimens. Only two distributions are considered robust, those of *L. natalensis* (LUTZ, 1930) and *L. podicipinus* (COPE, 1962). The provenance of some museum specimens is called into question. Unresolved problems are highlighted to encourage further studies to understand speciation processes and distribution patterns.

W. HERRE, Kiel

HARTL, B.; MARKOWSKI, J. (eds): **Ecological Genetics in Mammals**. Acta Theriologica, Vol. 38, Suppl. 2. Bialowieza/Poland: Mammal Research Institute, Polish Academy of Sciences 1993. 194 pp., num. figs. and tabs, paperback US $ 14,-. ISBN 83-900025-9-0

This book contains 14 articles from a meeting on 'Ecological Genetics in Mammals' held in Lódz in Poland in September 1992. The contributions deal with three major subjects: 1. Genetic variation, morphological variation, and developmental homeostasis; 2. Conservation genetics; and 3. Genetic variation, mating systems, and social organization. Many, though not all of the authors are Europeans (from Poland, Austria, Germany, Italy) who have collected data from European mammals (e.g. roe deer, *Capreolus capreolus*; brown hare, *Lepus europaeus*; wolf, *Canis lupus*; brown bear, *Ursus arctos*; alpine marmot, *Marmota m. marmota*; red deer, *Cervus elaphus*). Allozyme electrophoreses and some DNA techniques were used to detect and to determine genetic variation, but each article discusses a specific, genuine organismic problem of its own. A general, introductory paper at the beginning of each of the three main subjects completes the representation quite well. The booklet contains a unique collection of excellent papers on mammal ecology and population biology and is recommended as a good survey of the present development in this field. It might also be useful as a text for seminars on population biology and/or mammal ecology.

D. SPERLICH, Tübingen

Faith, D., 1990: Chance marsupial relationships. Nature **345**, 393–394.

Feduccia, A., 1980: The age of birds. Cambridge: Harvard Univ. Press. German translation: Es begann im Jura-Meer. Hildesheim: Gerstenberg Verlag.

Felsenstein, J., 1978: Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. **27**, 401–410.

– 1988: Phylogenies from molecular sequences, inference and reliability. Ann. Rev. Genet. **22**, 521–565.

Fitch, W.M., 1977: The phyletic interpretation of macromolecular sequence information, simple methods. In: Hecht, M.K.; Goody, P.C.; Hecht, B.M. (eds), Major patterns in vertebrate evolution. New York: Plenum Press. pp. 169–204.

Haeckel, E., 1896: Systematische Phylogenie der wirbellosen Thiere (Invertebrata). Zweiter Teil. Berlin: Verlag Georg Reimer.

Hendy, M.D.; Penny, D., 1989: A framework for the quantitative studies of evolutionary trees. Syst. Zool. **38**, 297–309.

Hennig, W., 1950: Grundzüge einer Theorie der Phylogenetischen Systematik. Berlin: Deutscher Zentralverlag. pp. 1–730.

– 1983: Stammesgeschichte der Chordaten. Fortschr. Zool. Syst. Evol.-forsch. **2**, 1–208.

– 1986: Wirbellose II, Gliedertiere. Thun: Verlag Harri Deutsch. pp. 1–335.

Hillis, D.M., 1987: Molecular versus morphological approaches to systematics. Ann. Rev. Ecol. Syst. **18**, 23–42.

– Huelsenbeck, J.P., 1992: Signal, noise, and reliability in molecular phylogenetic analyses. J. Heredity **83**, 189–195.

Ingber, D.E. 1993: The riddle of morphogenesis: a question of solution chemistry or molecular cell engineering? Cell **75**, 1249–1252.

Janke, A.; Feldmaier-Fuchs, G.; Thomas, W.K.; Haeseler, A.; Pääbo, S., 1994: The marsupial mitochondrial genome and the evolution of placental mammals. Genetics (in press).

Kitching, I.J., 1992: The determination of character polarity. In: Forey, P.L.; Humphries, G.J.; Kitching, I.J.; Scotland, R.W.; Siebert, D.J.; Williams, D.M. (eds), Cladistics — a practical course in systematics. Oxford: Clarendon Press. pp. 22–43.

Kumar, S.; Tamura, K.; Nei, M., 1993: Molecular Evolutionary Genetics Analysis (MEGA), Version 1.0. University Park, PA: Pennsylvania State University (computer program).

Lake, J.A., 1987: A rate-independent technique for analysis of nucleic acid sequences, evolutionary parsimony. Mol. Biol. Evol. **4**, 167–191.

– 1991: Lake replies. Trends Biochem. Sci. **16**, 289–290.

Larsen, N.; Olsen, G.J.; Maidak, B.L.; McCaughey, M.J.; Overbeek, R.; Macke, T.J.; Marsh, T.L.; Woese, C.R., 1993: The ribosomal database project. Nucl. Acid Res. **21**, 3021–3023.

Li, W.H.; Gouy, M., 1991: Statistical methods for testing molecular phylogenies. In: Miyamoto, M.M.; Cracraft, J. (eds), Phylogenetic Analysis of DNA Sequences. pp. 249–277.

Linkkila, T.P.; Gogarten, J.P. 1991: Tracing origins with molecular sequences: rooting the universal tree of life. Trends Biochem. Sci. **16**, 287–288.

Lorenzen, S., 1993: The role of parsimony, outgroup analysis, and theory of evolution in phylogenetic systematics. Z. zool. Syst. Evolut.-forsch. **31**, 1–20.

Nei, M., 1987: Molecular Evolutionary Genetics. New York: Columbia University Press.

Nelson, G.J., 1970: Outline of a theory of comparative biology. Syst. Zool. **19**, 373–384.

– 1978: Ontogeny, phylogeny, paleontology, and the biogenetic law. Syst. Zool. **27**, 324–345.

Patterson, C. 1982: Morphological characters and homology. In: Josey, K.A.; Friday, A.E. (eds), Problems in Phylogenetic Reconstruction. London: Academic Press. pp. 21–74.

– 1988: Homology in classical and molecular biology. Molecular Biology Evolution **5**, 603–625.

– Williams, D.M.; Humphries, C.J., 1993: Congruence between molecular and morphological phylogenies. Ann. Rev. Ecol. Syst. **24**, 153–188.

Remane, A., 1961: Gedanken zum Problem, Homologie und Analogie, Praeadaptation und Parallelität. Zool. Anz. **166**, 447–465.

Saitou, N.; Nei, M., 1986: The number of nucleotides required to determine the branching order of three species, with special reference to the human-chimpanzee-gorilla divergence. J. Mol. Evol. **24**, 189–204.

– Nei, M., 1987: The neighbour-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4**, 406–425.

Schlee, D., 1971: Die Rekonstruktion der Phylogenese mit Hennig's Prinzip. Aufs. Red. Senckenberg naturforsch. Ges. **20**, 1–62.

Sourdis, J.; Nei, M., 1988: Relative efficiencies of the maximum parsimony and distance-matrix methods in obtaining the correct phylogenetic tree. Mol. Biol. Evol. **5**, 298–311.

Sudhaus, W.; Rehfeld, K., 1992: Einführung in die Phylogenetik und Systematik. Stuttgart: G. Fischer Verlag.

Swofford, D.L., 1991: Phylogenetic Analysis using Parsimony (PAUP), Version 3.1. Champaign, IL: Illinois Natural History Survey (computer program).

– Olsen, G.J., 1990: Phylogeny reconstruction. In: Hillis, D.M.; Moritz, C. (eds), Molecular systematics. Sunderland: Sinauer Assoc. pp. 411–501.

Wägele, J.W., 1993: Rejection of the 'Uniramia' hypothesis and implications of the Mandibulata concept. Zool. Jb. Syst. **120**, 253–288.

– 1994: Review of methodological problems of 'Computer cladistics' exemplified with a case study on isopod phylogeny (Crustacea, Isopoda). Z. zool. Syst. Evol.-forsch. **32**, 81–107.

– ; Wetzel, R., 1994: Nucleic acid sequence data are not *per se* reliable for inference of phylogenies. J. Nat. Hist. **28**, 749–761.

Wagner, G.P., 1989: The biological homology concept. Ann. Rev. Ecol. Syst. **20**, 51–69.

Wheeler, W.C., 1990a: Combinatorial weights in phylogenetic analysis, a statistical parsimony procedure. Cladistics **6**, 269–275.

– 1990b: Nucleic acid sequence phylogeny and random outgroups. Cladistics **6**, 363–367.

Wiley, E.O., 1975: Karl R. Popper, systematics, and classification, a reply to Walter Bock and other evolutionary taxonomists. Syst. Zool. **24**, 233–243.

– 1981: Phylogenetics. The Theory and Practice of Phylogenetic Systematics. New York: Wiley and Sons.

Zharkikh, A.; Li, W.H., 1993: Inconsistency of the maximum–parsimony method, the case of five taxa with a molecular clock. Syst. Biol. **42**, 113–125.

Author's address: Abt. Systematik und Morphologie der Tiere, University of Bielefeld, P.O.B. 10 01 31, D-33501 Bielefeld, Germany

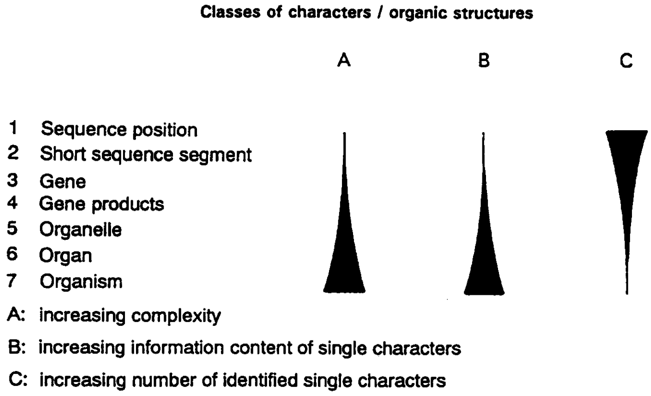

Classes of characters / organic structures
A
B
C
1 Sequence position
2 Short sequence segment
3 Gene
4 Gene products
5 Organelle
6 Organ
7 Organism
A: increasing complexity
B: increasing information content of single characters
C: increasing number of identified single characters

Fig. 1. Illustration of inverse relationship between information content of single characters and number of available characters (not drawn to scale)

are less complex than those of an ant. It is also possible to estimate the different information content of non-homologous characters, which is important for character weighting: an arthropod seta is less structurally complex than the stomach of a decapod crustacean, so the latter deserves a higher weight; pigment patterns of animals often vary within a species (= a character of low weight, or even of no use at all), while the basic construction of a feather is a unique feature that evolved only once.

2 No reconstruction of evolution without rooting of dendrograms

If one decides to study phylogeny with the help of sequence data and the computer programs available today (distance methods, parsimony methods etc.: Nei 1987; Lake 1987; Saitou and Nei 1986, 1987; Swofford and Olsen 1990; Swofford 1991; Bandelt and Dress 1992; Kumar et al. 1993) one difficulty soon arises: without knowledge of which species or groups of species can be used as out-groups, it is difficult to root the topologies of relationships. Comparative morphology enables the establishment of well-founded hypotheses on the phylogeny of major groups of organisms and on the evolution of organs (see the work of Charles Darwin (1871) or Ernst Haeckel (1896)). The main features of currently widely accepted trees have been established on the basis of characters with a high information content (Darwin 1871, 1974 reprint, p. 9: 'So that the correspondence in general structure, in the minute structure of the tissues, in chemical composition and in constitution, between man and the higher animals, especially the anthropomorphous apes, is extremely close').

2.1 Rooting using low-information-content characters

The information content of single-sequence positions is not high enough to find character-state order *a priori* (Swofford and Olsen 1990). This is also true for simple morphological characters (e.g. numbers of setae, chromatophore patterns, allometric data).

In the simplest case of sequences that evolved without back mutations and convergence or parallelism, each informative position causes a meaningful split in the group of species under study, i.e. each change in a character state produces two groups: one with the old state, one with the new state. These splits were produced historically in a time series, which must be reconstructed to find the phylogenetic tree, however, although, ideally, the true topology of the tree can be found with a good set of data using distance-matrix methods or maximum parsimony, the direction of evolution remains unknown. This well-known fact is illustrated by the example in Figure 2. Rooting with an outgroup helps to find the order of character states, but this is a risky procedure when using sequence data, due to convergence in the outgroup and back mutations in the ingroup etc., which, in practice, may be frequent. Instead of using an outgroup, a paralogous sequence can be used to root the tree in cases where gene duplication has been discovered (Linkkila and Gogarten 1991), however, the 'long-branch problems' remain the same (Lake 1991).

In the literature, some methods of finding the root without an outgroup have been proposed. Lorenzen (1993) remarked that, in a parsimonious tree, characters should show an encaptic order of apomorphic states. The idea is that monophyla of a young age are contained within larger, older monophyla. This of course causes a unique pattern of apomorphies of different ages which are ordered encaptically, however, when analysing sequence data, the problem is that apomorphic states can not be identified easily when a site is the character. The simple example in Figure 2 shows that the encaptic order approaches the true topology, but the root cannot be discovered without additional information, even if the topology is the single most parsimonious tree. Additional information such as the pattern of asymmetry in a number of putative apomorphic positions of a functional ingroup in comparison with the corresponding functional outgroup could be used. Such a pattern is not visible in simple examples (Fig. 2). Wheeler (1990b) recommended the use of an asymmetrical character-transformation matrix (Wheeler 1990a) for closely related species: if asymmetry in the character transformations is present, the outgroup taxon is the one that leads to the lowest cost. To find the outgroup taxon, all taxa are tested as outgroups and the cost of character transformations is calculated. Future research will show how useful this procedure is.

A different method was used by Eigen et al. (1989). Those authors studied the degree of randomization of tRNA in mitochondria, bacteria and eukaryotes, in comparison with reconstructed consensus-node sequences. For moderately diverged positions, the node sequence with the smallest relative mutational distance must be closest to the root (Eigen et al. 1988, 1989).

2.2 Rooting with highly informative characters, and the advantages of *a priori* determination of character-state polarity

If it is so difficult to root trees in order to discover the direction of character-transformation series, one must question why zoologists have been so sure that poriferans and cnidarians are historically old groups, evolving much earlier than amphibians or dragonflies, even without analysing data from the fossil record. Comparative morphology enables one to make these conclusions with a certainty not available with the type of sequence analysis that has become popular in the past few years. The cause of this difference is the unequal information content of the data. Previously published molecular phylogenies used DNA-sequences of between 250 bp and (rarely) 20 000 bp. These tiny samples of the genome cannot describe the very different complexity of the structure of different organisms. Therefore, comparative morphology is an approximate method for the comparison of large quantities of genetic information.

```
                  Potential autapomorphies of
                  Mammalia:                                  Amphibia:

                                     1111111122223333333        133
                  1222445555555677771155889911285556778        6456
                  9116683344445900033459283800522478594      676190
                  0675247905670913536781559617966495992      924955

Homo sapiens      GCCGUAGGGCGGGCGCGGGCGGACGGGGCCGCUCCU       UCAAUA
Mus musculus      .....................................      ......
Rattus norw. 1    ............C........................      ......
Rattus norw. 2    .....................................      ......
Oryctolagus 1     ...-......G..........................      ......
Oryctolagus 2     ......??.............................      ......
Alligator m.      CGUACC------CCUUUAU-UACGUAAC?AUAUCU?C      ...?..
Gallus gallus     C.?A--C?---CG.CGUUA??.ACGUAAC.U..?U?G      ...?..
Heterodon p.      CGUAC?------CCUUUA?UUACGU.ACAAUAU?UUC      ......
Xenopus l.        CGUACC---CGCCCUUUAUUU.CGUAACAAUAUCU.C      CGGGCG
Bufo valliceps    CGUAC?C?CCUC?CUUUAUUUACG.A.CAAUAUCU-C      CGGGCG
Hyla cinerea      C??ACC?C?CUC?CUUUAUUUACGUAACAAUAUCU-C      CGGGCG
Ambystoma m.      CGUACC------CCUUU?UUUACGUAACAAUAUCUUC      CGGGCG
Echinorhinus      CGUACU-----CCCUUUAUUUACGUAACAAUCUCU-C      ......
Squalus a.        CGUACU-----.CCUUUAUUUACGUAACAAUCUCU-C      ......
Fundulus h.       CGUACCCC.CACCC.U.AUUUACGUAAC.AUAUCU-C      A...C.
Sebastolobus      UGUACG--CUGCCCUUUAUUUACGUAAC.AUCUCU-C      A.....
```

Fig. 4. Selection of positions of sequences used for Figure 3. Positions supporting splits of Mammalia/remaining taxa and Amphibia/remaining taxa are shown

the sister-group of the mandibulates, based on a large amount of morphological (Hennig 1986) and physiological data (Wägele 1993), is founded more on plain information than the 'discovery' that onychophorans are arthropods, with myriapods as the earliest branch (Ballard et al. 1992), based on 245 informative sites of 12SrRNA (Wägele and Wetzel 1994). On the sequence level, complex characters such as large inversions, rearrangements, or the opening of primarily circular DNA-strands (Bridge et al. 1992; Janke et al. 1994) may be events of a higher quality than a simple substitution, but these are not frequently used.

Neither morphologists nor molecular systematists can guarantee the uncontroversial assent of their colleagues in relation to their results (review: Patterson et al. 1993). However, at the moment, informative morphological characters are easier to obtain: the feather establishes the monophyly of birds, the radula that of molluscs. Sequence data are useful when other types of characters are not available (very closely or distantly related species, bacteria etc.), but the information content of sequences must be examined to avoid the formulation of absurd hypotheses.

Finally, we should not forget that apparent clues to the correct identification of species in the field must, for the foreseeable future, still rely on descriptions of morphology. Furthermore, study of evolution is more than just the analysis of relationships. The 18SrDNA tells us nothing about the wonderful adaptations of microchiropterans to nocturnal insects hunting.

Acknowledgements

The author is grateful to Prof. H.J. Bandelt (University of Hamburg) for fruitful discussions on the theory of molecular systematics, and to Prof. A. Dress (University of Bielefeld) for some useful comments. Mr G. Stanjek kindly wrote a very useful computer program for the selection of putative autapomorphies.

Zusammenfassung

Ein fundamentaler Unterschied zwischen der einzelnen Sequenzposition oder auch numerischen Merkmalen und komplexen morphologischen Merkmalen ist ihr Informationsgehalt. Merkmalsreihen komplexer Strukturen enthalten meist genügend Information, um *a priori* die Bestimmung der Lesrichtung zu ermöglichen. Die Feststellung des Ursprunges eines Dendrogramms ist somit ohne kladistischen Außengruppenvergleich möglich, während Bäume (Topologien), die auf wenig informativen Merkmalen beruhen, allgemein nur mit dem kladistischen Außengruppenvergleich 'gewurzelt' werden können. Die Qualität der insgesamt verwendeten Information ist entscheidend für die Wahl zwischen alternativen Verwandtschaftshypothesen. Numerische Methoden der Rekonstruktion der Phylogenese sind nützlich bei Verwendung einer *großen* Zahl informationsarmer Merkmale; das Hennigsche Verfahren ist für komplexe Merkmale vorzuziehen.

References

Ax, P., 1988: Systematik in der Biologie. UTB. Stuttgart: G. Fischer Verlag.

Ballard, J.W.; Olsen, G.J.; Faith, D.P.; Odgers, W.A.; Rowell, D.M.; Atkinson, P.W., 1992: Evidence from 12S ribosomal RNA sequences that onychophorans are modified arthropods Science **258**, 1345–1348.

Bandelt, H.J.; Dress, A.W., 1992: Split decomposition: a new and useful approach to phylogenetic analysis of distance data. Molecular Phylogenetics Evolution **1**, 242–252.

Bridge, D.; Cunningham, C.W.; Schierwater, B.; Desalle, R.; Buss, L.W., 1992: Class-level relationships in the phylum Cnidaria: evidence from mitochondrial genome structure. Proc. Natl. Acad. Sci. **89**, 8750–8753.

Cedergren, R.; Gray, M.W.; Abel, Y.; Sankoff, D., 1988: The evolutionary relationships among known life forms. J. Mol. Evol. **28**, 98–112.

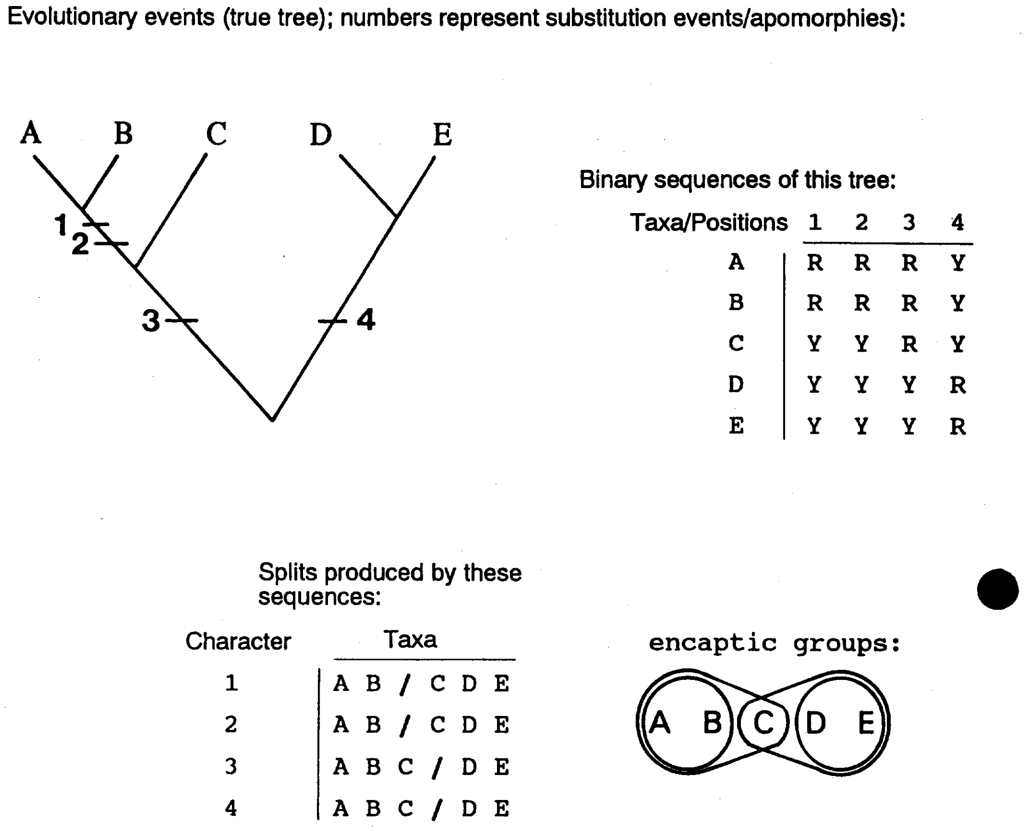
Darwin, C., 1871: The descent of man and selection in relation to sex, reprint 1974. Detroit: Gale Research Co. pp. 1–672.

Dohle, W. 1989: Zur Frage der Homologie ontogenetischer Muster. Zool. Beitr. **32**, 355–389.

Donoghue, M.J.; Maddison, W.P., 1986: Polarity assessment in phylogenetic systematics, a response to Meacham. Taxon **35**, 534–545.

Eigen, M.; Winkler-Oswatitsch R.; Dress, A., 1988: Statistical geometry in sequence space: a method of quantitative comparative sequence analysis. Proc. Natl. Acad. Sci. USA **85**, 5913–5917.

– Lindemann, B.F.; Tietze, M.; Winkler-Oswatitsch, R.; Dress, A.; von Haeseler, A. 1989: How old is the genetic code? Statistical geometry of tRNA provides an answer. Science **244**, 673–679.



Evolutionary events (true tree); numbers represent substitution events/apomorphies):

Binary sequences of this tree:

Taxa/Positions	1	2	3	4
A	R	R	R	Y
B	R	R	R	Y
C	Y	Y	R	Y
D	Y	Y	Y	R
E	Y	Y	Y	R

Splits produced by these sequences:

Character	Taxa
1	A B / C D E
2	A B / C D E
3	A B C / D E
4	A B C / D E

encaptic groups:

Fig. 2. Example demonstrating that the true tree cannot be reconstructed by searching the encaptic order of characters when character states are not polarized. The ancestor sequence of the example is YYYY

Of course, morphologists do not compare sequences but, rather, the results of complex interrelated gene expression.

In comparative morphology, *a priori* polarization of a transformation series is possible (Hennig 1950; Donoghue and Maddison 1986) when the complexity of characters is large. It is reasonable to suppose that a sophisticated construction (e.g. lens eye) cannot appear by chance without a precursor. This implies that structurally and functionally simpler organs with high probability are historically older than more complex and efficient ones, as long as simplicity is not the result of reduction. 'Simple' does not always imply 'primitive', but these secondary phenomena (reductions) can be identified when they occur in a specimen of high anatomical organization. In these cases, the species cannot be as phylogenetically old as the simplified organ would suggest (e.g. sense organs in parasites, in deep-sea species, in dwarfish subterranean species etc.). Thus, an outgroup comparison is not always necessary for polarization of transformation series or for rooting of trees (descending analysis: Donoghue and Maddison 1986; Sudhaus and Rehfeld 1992).

The cladistic outgroup comparison has the disadvantage of unnecessary uncertainties (Lorenzen 1993): convergence in an outgroup can be the source of erroneous polarization of states of the ingroup, finally producing the wrong tree (Wägele 1994). It seems that it is often thought that analysing characters carefully involves too much work. Wherever characters do allow the *a priori* determination of polarity, using the automatic rooting via outgroups would cause a loss of valuable information.

Transformation series can also be polarized with the help of the orderliness of ontogeny (Nelson 1970, 1978; review: Kitching 1992), i.e. relying on Haeckel's biogenetic principle, in combination with the principle of parsimony. This procedure also requires complex characters to ensure that homology can be corroborated.

The *a priori* determination of character-state polarity is at least as difficult when the sequence position or a numerical feature is the character used. As explained above, this is a consequence of the characters' low information content. The experienced taxonomist knows that some characters vary more than others. The reason for this variability in 'weak' characters is probably that either epigenetic factors or few mutations already have an effect on morphology, which implies in turn a higher probability of convergences. Unfortunately, the use of 'weak' characters is not always avoidable; closely related species often possess no complex synapomorphies. The same criteria as those for molecular data must be applied: if the information content of simple characters is low, a large amount of data is necessary to avoid stochastic errors.

W. Hennig, the founder of the strict logic used by cladists, was well aware of the difference between simple and complex characters (Schlee 1971). His method requires *a priori* determination of polarity (and, therefore, high-information-content characters), a procedure accepted by many modern systematists when morphological characters are used (Wiley 1975, 1981; Ax 1988; Sudhaus and Rehfeld 1992). The outgroup comparison used is based on test samples (Lorenzen 1993), in contrast to the cladistic ('automatic') type of outgroup comparison, where control of determination of character polarity is not aimed at.

3 Identification of homologies and homoplasies

It is well known that low information content in characters reduces the probability of homology of similarities. An alignment similarity ('G' in position 132) is a potential homology and, at the same time, a potential convergence (Patterson 1988), a one-to-one comparison of characters of two species (Wagner 1989) is not possible. Only two methods can produce a hypothesis of homology (with no guarantee for the single position): 1. Reference to a dendrogram, when character distribution is compatible with the topology; and 2. Reference to the identity of aligned sequence segments.

Using morphological characters, this problem only exists if complexity is low. The lens eye of cephalopods can easily be identified as convergence with the vertebrate eye. Morphological characters offer an opportunity of clarifying the question of 'homology vs. convergence' by analysis of details (e.g. of ultrastructure), i.e. increasing the amount of data, and, as a result of that, the probability that identity is caused by common descent, while single-sequence positions and simple morphological characters do not contain additional information (see e.g. 'long-branch problems' caused by convergence: Fitch 1977; Felsenstein 1978; Lake 1987; Cedergren et al. 1988). Homology of characters with a high information content can be ascertained without the congruence test proposed by Patterson (1982; see also criteria for homology: Remane 1961; discussion in: Dohle 1989). A hypothesis of monophyly can be substantiated with *a priori* data (i.e. by character analysis carried out independently of the phylogenetic analysis) whenever the probability of homology in an apomorphic state can be confirmed. Character analysis is therefore the most important step in a phylogenetic study based on morphology.

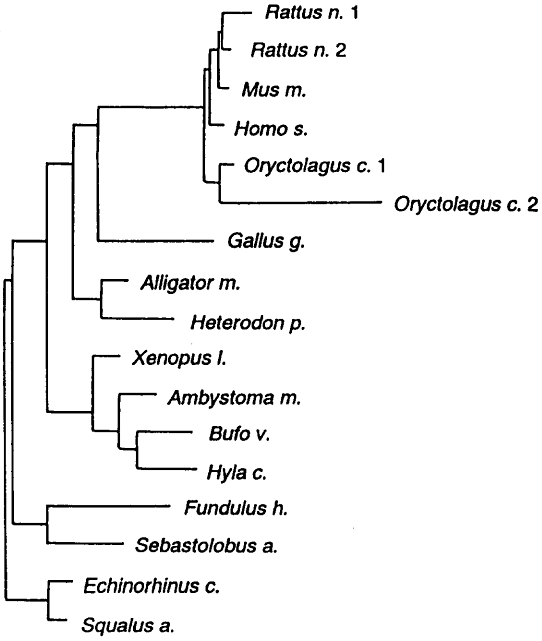


Fig. 3. Dendrogram of vertebrate species based on complete 18SrRNA sequences. Source of aligned sequences: RDP database (Larsen et al. 1993); total number of sites including alignment gaps: 3939. Distance analysis using MEGA computer program (Kumar et al. 1993; neighbour joining and Tajima-Nei distance)

4 Complexity in sequence data

The idea that taxa-specific segments of sequences can be used as complex characters (signatures) should be familiar to a morphologist. The probability of obtaining a short sequence of five nucleotides twice by chance is very low: 4^{-5} (about 10^{-3}). Doubling the number of nucleotides increases the probability of homology in two identical sequence segments by three orders of magnitude. It is safe to postulate homology for identical sequence segments (in contrast to single-sequence positions). Using such segments as single characters would allow weighting in dependence of segment length.

However, in molecular systematics, a difference between the quantity of homologous substitutions accumulated in shorter segments in comparison to a similar number of substitutions distributed over the length of the sequence is usually not emphasized. Complexity at a sequence level is represented by the total number of putative apomorphies that support monophyly in a group, as in parsimony analysis. Signatures could, nevertheless, be useful. They can be identified *a posteriori* as a characteristic of a monophylum and then be used for diagnostic purposes (e.g. for the design of taxa-specific primer).

The following example illustrates a problem connected with *a priori* analysis of the 'information content' of sequences (this paper does not aim to discuss the voluminous literature on the reliability of inference methods). Figure 3 shows a tree of vertebrate species calculated from 18SrRNA sequences (1800–1900 bp) selected from an alignment of the RDP-database (Larsen et al. 1993). These sequences contain only 170 parsimony-informative positions. Of these, 27 support monophyly of Mammalia, and some further positions show convergence in the non-mammalian taxa (Fig. 4). These supporting positions can be considered as the 'minimum number of putative autapomorphies'. Depending on tree topology, maximum parsimony could identify further putative autapomorphies due to the reconstruction of an ancestral-node sequence. Monophyly of Amphibia is supported by only five positions, a sixth one (3595 in Fig. 4) shows a convergence in the fish *Fundulus heteroclitus*. Monophyly of Anura and of the Archosauria (represented by the two species *Gallo gallo* and *Alligator mississippiensis*), is not supported (Fig. 3). This means that substitutions that are apomorphic for the Archosauria have probably been overlaid by subsequent events and/or, possibly, that only a few apomorphies were present in their common-ancestor sequence. Morphology strongly supports monophyly of the Archosauria (Feduccia 1980; Hennig 1983). Thus, one must conclude that this alignment of 18SrRNA is not very informative for a comparison at this taxonomic level.

It has often been suggested that, by increasing the length of sequence data, the analysis improves (Sourdis and Nei 1988; Zharkikh and Li 1993). With a given rate of substitutions or model of sequence evolution, it is possible to calculate, in simulations, how long sequences must be to find the correct topology of a tree (Saitou and Nei 1986; Felsenstein 1988; Li and Gouy 1991). But, in practice, analysis of the quality of data is still not satisfactory: the number of putative autapomorphies supporting monophyly in a group is often not counted (see also branch-length estimation proposed by Hendy and Penny (1989)), instead, bootstrap percentage values are widely used as evidence for the quality of the data, however, random data also produce most parsimonious trees (Faith 1990; Hillis and Huelsenbeck 1992). It is not surprising that morphologists rely more on phylogenetic systems that are based on complex homologies and distrust specific phylogenies obtained from short sequences. The statements that onychophorans are protarthropods (= pararthropods) and that the chelicerates are