

Assessing Phylogenetic Accuracy

A simulation study

Theodoor Heijerman

CENTRALE LANDBOUWCATALOGUS



0000 0670 0609

Promotor: Dr R. J. Post, hoogleraar in de diertaxonomie

Assessing Phylogenetic Accuracy

A simulation study

Theodoor Heijerman

Proefschrift

ter verkrijging van de graad van doctor
in de landbouw- en milieuwetenschappen
op gezag van de rector magnificus,
dr C. M. Karssen,
in het openbaar te verdedigen
op woensdag 27 september 1995
des namiddags te vier uur in de Aula
van de Landbouwniversiteit te Wageningen

BIBLIOTHEEK
LANDBOUWUNIVERSITEIT
WAGENINGEN

CIP-GEGEVENS KONINKLIJKE BIBLIOTHEEK, DEN HAAG

Heijerman, Theodoor

Assessing phylogenetic accuracy : a simulation study /
Theodoor Heijerman. - [S.l. : s.n.] (Wageningen : Ponsen &
Looijen)

Thesis Landbouw Universiteit Wageningen. - With ref. -
With summary in Dutch.

ISBN 90-5485-422-7

Subject headings: phylogenesis / animal taxonomy.

Cover: 'n kroezen boom

Stellingen

- 1 Het getuigt van te veel optimisme ten aanzien van het prestatievermogen van numeriek taxonomische methoden, wanneer deze worden aangeduid als fylogenie-reconstructie-methoden.

(dit proefschrift; F. J. Rohlf & M. C. Wooten, 1988. Evaluation of the restricted maximum-likelihood method for estimating phylogenetic trees using simulated allele-frequency data. — *Evolution* 42: 581-595).

- 2 Het beschikbaar komen van steeds gebruikersvriendelijker programma's voor cladistische analyses, zal de gemiddelde kwaliteit van de resultaten van deze programma's doen afnemen.

- 3 Het is sterk aan te bevelen om bij de presentatie van de resultaten van een cladistische analyse, niet alleen de meest parsimone, maar ook sub-optimale oplossingen te geven.

- 4 De toepassing van fenetische technieken om verwantschapsrelaties te bepalen, is ten onrechte in onbruik geraakt.

(dit proefschrift; D. L. J. Quicke, 1993. *Principles and techniques of contemporary taxonomy*. Chapman & Hall, Glasgow)

- 5 Classificaties dienen ten minste consistent te zijn met de beschikbare fylogenetische informatie. Omdat onze inzichten over verwantschapsrelaties aan verandering onderhevig zijn, zal de gewenste stabiliteit van classificaties aan de vereiste consistentie moeten worden opgeofferd.

- 6 Techniques should complement, not compete.

(dit proefschrift; A. W. Moss, 1983. Taxa, taxonomists, and taxonomy. In: *Numerical Taxonomy* (J. Felsenstein, ed.): 72-75. Springer-Verlag, Berlin; J. Kim, 1993. Improving the accuracy of phylogenetic estimation by combining different methods. — *Systematic Biology* 42: 331-340)

- 7 Natuurwaarde is een concept zonder waarde.

- 8 Gezien de aard van vele zogenoemde natuurbeheersmaatregelen, kan natuurbeheer in Nederland gelijk worden gesteld aan tuinieren.

- 9 Natuurbeschermingsbeleid, gebaseerd op Rode Lijsten van zeldzame en bedreigde insecten, zal uiteindelijk een averechts effect sorteren.

- 10 Bomen zijn bedrog.
 - 11 Hoge bomen maken veel wind.
 - 12 Colleagues should complement, not compete.
 - 13 Ook de Schepper moet een voorliefde voor Snuitkevers hebben gehad.
-

Stellingen behorende bij het proefschrift:

**Assessing phylogenetic accuracy
A simulation study**

Wageningen, 27 september 1995

Theodoor Heijerman

CONTENTS

1	Introduction	7
2	GENESIS: a simulation model of phylogeny The origin and early evolution of character state vectors	29
3	GENESIS: a simulation model of phylogeny A sensitivity analysis	57
4	Adequacy of numerical taxonomic methods A comparative study based on simulation experiments . .	77
5	Adequacy of numerical taxonomic methods Further experiments using simulated data	111
6	Adequacy of numerical taxonomic methods Why not be a pheneticist?	139
7	Discussion	161
8	Summary	167
9	Samenvatting	171
	Curriculum vitae	175

1

Introduction

"...we need to understand the diversity of living things for the same reasons that compel us to reach out towards understanding the origins and eventual fate of the universe, or the structure of the elementary particles that it is built from, or the sequence of molecules within the human genome that code for our self-assembly." May (1990:180)

How many species?

The world is a very diverse place: the number of described extant species is estimated to total approximately 1.7 million. Estimates for numbers of total extant species range from 8 to 100 million and an estimation of 30 million is generally accepted as a realistic one. Biodiversity is at crisis: Since 1600 there are more than 700 recorded extinctions of species (Reid, 1992). It was estimated that we may lose 100,000 species per year, mainly due to habitat destruction (data from *World Conservation Monitoring Center*, 1992). The description rate for new species is over 9,000 per year. This means that a species is much more likely to become extinct than to be described. Reid (1992) predicts that in the next 25 years 2 - 8 % of the current number of species on earth will face extinction. May (1990) states that "over half the species currently extant are likely to become extinct over the next 50 - 100 years". Other authorities agree "that 30 - 50% of all living species may go extinct in the next three to five decades" (Novacek & Wheeler, 1992, and references therein).

The number of extant species is only a fraction of the number of species that have ever lived on our planet and of which again only a fraction is or will be known to us as more or less imperfectly preserved fossils. More than 99 percent of all species that ever existed are believed to have become extinct (e.g. Raup, 1981).

Concepts of biodiversity

In general the concept of biodiversity refers to variety within the living world. Biodiversity can be defined in several ways and we can distinguish between genetic diversity, species diversity, ecosystem diversity and taxonomic diversity. Eldredge (1992) further defines phenotypic diversity as the amount of variation within or among populations, species or higher taxa (disparity). The most classical of these, species diversity, measures diversity as a function of the numbers of species present in a habitat or site (species richness) and their relative abundances. Species diversity measures are often used to quantify biodiversity and habitat quality.

Taxonomic diversity refers to diversity at higher levels of the hierarchical classification system, and is measured in terms of the number of phyla, classes, orders, families etcetera and of the number of species in these categories. Eldredge (1992) uses the term genealogical diversity to refer to the number of taxa within a monophyletic clade. Thus taxonomic diversity is a concept based on knowledge of classification. If classifications themselves should be based on evolutionary relationships, than taxonomic diversity cannot be assessed without reference to genealogical hierarchies.

Task of taxonomy

It is the primary task of taxonomy to document this diversity and to explain its nature and origin. Taxonomy can thus be defined as the theory and practice of describing the diversity of organisms and of the relationships among them. Many authors want to distinguish

between taxonomy and systematics. In Mayr's (1969) view, taxonomy is the "theory and practice of classifying organisms", whereas systematics is the "scientific study of the kinds and diversity of organisms and of all relationships among them". Wiley (1981) defines taxonomy as "the theory and practice of describing the diversity of organisms and ordering this diversity into a system of words that conveys information concerning the kind of relationship between organisms that the investigator thinks is relevant". Taxonomy may comprise nomenclature and classification as well as identification. Systematics is often understood as a broader area which includes taxonomy and adds to it the theoretical and practical aspects of evolution, genetics and speciation (e.g. Quicke, 1993). As there seems to be no general agreement on this issue, I would prefer to use these terms as near synonyms. The core assignment of taxonomy can be described as the assessment of relationships between taxa and, for practical purposes, the arrangement of these taxa into a classification system.

Relationship

In taxonomy, the term relationship in the broad sense may be used as including all possible biological relationships among organisms. Various different meanings of relationship may be distinguished. Phenetic or similarity relationship may be defined as "overall similarity as judged by the characters of the organism without any implication as to their relationship by ancestry" (Sokal & Sneath, 1963). As such phenetic relationship may also include breeding relationships and ecological relationships (Pankhurst, 1991). Phenetic similarity may be the result of evolutionary relationship: taxa may be similar because they are closely related by descent.

Evolutionary or phylogenetic relationship is relationship based on common ancestry or genealogical affinity. The use of the term genealogical relationship is sometimes reserved for blood-connections between individuals and generations within species.

Taxonomic relationship between taxa can only be established by examining the characters of the taxa under study. Similarities between taxa may be indicative of their degree of relationship. However, similarity can be thought of to contain different components. Similarity due to common ancestry is termed patristic similarity. Character evolution involves the change of one state, the primitive, preexisting state, into a new, derived state. Patristic similarity can therefore be further divided into two components, primitive similarity and advanced or derived similarity, which are termed symplesiomorphy and synapomorphy, respectively (Hennig, 1966).

Homoplastic similarity or homoplasy denotes similarity due to parallelism and convergence. Patristic similarity and homoplasy taken together constitute phenetic or overall similarity.

Taxonomic relationship can be visualized in the form of branching diagrams containing groupings of taxa. Such diagrams are termed dendrograms. Various kinds of dendrograms can be distinguished depending on the kind of similarity or relationship they are supposed to reflect. A cladogram is defined by Wiley (1981) as "a branching diagram of entities where the branching is based on the inferred historical connections between the entities as evidenced by synapomorphies". In other words, a cladogram is a common ancestry tree. A phenogram is "a branching diagram linking organisms by estimates of overall similarity as evidenced from a sample of characters" (Wiley, 1981).

Classification

Classification is the process of ordering organisms or taxa into clusters based on some criterion of taxonomic relationship and the subsequent naming of these clusters. Classification involves the translation of dendrograms (cladograms or phenograms) into a formal system of words. In general two types of classifications can be distinguished, viz. natural and non-natural (also arbitrary or artificial) classifications. A natural classification is one that contains groups

that are thought to really exist in nature. Therefore such a classification contains information about the evolutionary process. Artificial classifications on the other hand, do not specifically aim to reflect the historical process. In the definition of Wiley et al. (1991): "An artificial classification is a classification containing one or more artificial groups ...". However, there can be considerable disagreement on what is a natural taxon and what is a natural classification. According to one view a classification which has only a *limited use* is a special or artificial classification. "A natural taxonomy is a general arrangement intended for general use by all biologists" (Gilmour in Sneath & Sokal, 1973). A natural group is one that has a "high content of implied information" (Sneath & Sokal, 1973). This concept of naturalness, also termed Gilmour naturalness, implies that the more characters contribute to overall similarity, the more natural the resulting grouping of taxa will be, and the greater the information content of the classification. In another view, which will be discussed in more details later, natural taxa are monophyletic taxa and natural classifications are those established on the basis of phylogenetic relationship. A third view of naturalness identifies grades, which are taxa characterized by a general level of adaptation. Members of a grade are "characterized by a well integrated adaptive complex" (see Mayr, 1974) and to "the evolutionary taxonomist the existence of grades seems often more significant and more meaningful biologically than the mere splitting of phyletic lines" (Mayr 1974:107). Examples of well known grades are birds and reptiles. So in the opinion of some authors grades have a biological meaning, are the products of historical processes and thus could be considered natural groups.

Schools of taxonomy

There are basically three fundamental approaches to taxonomy. As a common goal each of these schools attempts to group taxa into a convenient classification. The approaches differ in the concept of

relationship that is applied and/or in the way they use the information on relationship to construct a classification. The three approaches are often referred to as schools of taxonomy and are known as phenetic taxonomy, phylogenetic taxonomy and evolutionary taxonomy.

Phenetic taxonomy

Phenetic taxonomy or phenetics is the approach in which taxa or operational taxonomic units (OTUs) are grouped into clusters on the basis of overall similarity. Similarity or dissimilarity can be estimated by similarity coefficients. There are many kinds of coefficients, like distance, association or correlation coefficients, but there are also probabilistic similarity coefficients and information theory measures of similarity. One is referred to Sneath & Sokal (1973) or Clifford & Stephenson (1975) for a full account.

The contribution of characters to the overall similarity may be effected by the different scales of measurement or differences in variability between characters. Therefore it is sometimes desirable to manipulate characters in order to equalize their contributions. This may be done by standardization or transformation (Sneath & Sokal, 1973; Clifford & Stephenson, 1975). A commonly used standardization procedure is normalization, in which the raw data are expressed as deviations from the mean in standard deviation units. This type of standardization results in all character means becoming 0, with a standard deviation of 1.

Clustering procedures use a resemblance matrix to produce phenograms. There are many kinds of clustering procedures; see e.g. Sneath & Sokal (1973), Clifford & Stephenson (1975) for more detailed information on clustering methods. Taxonomists generally use clustering procedures that are referred to as Sequential, Agglomerative, Hierarchic and Nonoverlapping (SAHN-clustering). Of all SAHN-procedures, the unweighted pair-group method using arithmetic averages (UPGMA or the group average method), appears to be the one most widely employed.

Phylogenetic taxonomy

An alternative approach, devoted to the study of evolutionary relationships, is phylogenetic systematics or cladistics. Natural taxa are groups of species that really exist in nature. They are the products of the evolutionary process; they exist whether we are able to perceive them or not. Species also are natural taxa. A natural taxon that is composed of two or more species constitutes a monophyletic group, also referred to as a clade. A monophyletic group can be defined as a group of species that includes an ancestral species and all of its descendants (Wiley et al., 1991). Evolutionary novelties (apomorphies) that arose in an ancestor species will be inherited by its descendants. Monophyletic groups can be postulated on basis of a shared possession of such novelties (characters in the derived state or synapomorphies).

For cladists, natural groups are monophyletic groups. Artificial taxa are taxa that do not exist in nature as the result of a unique evolutionary history. There are two kinds of artificial and thus non-monophyletic groups. Paraphyletic taxa are groups that do not include all the descendants of a common ancestor and are diagnosed by plesiomorphies. Polyphyletic taxa are defined as groups of which the ancestor belongs to another group, and are based on convergent characters. Examples of supposed paraphyletic groups are the Polychaeta, Oligochaeta, Turbellaria, Apterygota, Symphyta, Prosimiae, Anamnia, Reptilia, Pongidae, Invertebrata, Algae, Gymnospermae, Pisces. The Homeothermia and the Vermes are examples of polyphyletic taxa (Ax, 1987; Quicke, 1993).

Monophyletic groups are diagnosed by apomorphies. Therefore we must know the direction of character evolution (polarity), i.e. a method is needed to identify character states in a transformation series as being ancestral or derived. The most commonly used method to make polarity decisions is outgroup comparison. An outgroup may be defined as "any group used in an analysis that is not included in the taxon under study [the ingroup]" (Wiley et al., 1991). The outgroup rule states that of two or more states within a

group (ingroup) the state also occurring in the outgroup, may be inferred to be the plesiomorphic one. The most critical outgroup consists of the sistergroup, which is the closest related monophyletic group to the ingroup. Ax (1987) has introduced the term adelphotaxon for the concept of sistertaxa and defined adelphotaxa as "evolutionary species, or monophyletic species groups, of the first degree of phylogenetic relationship. They arise by the dichotomous splitting of a stem species common to them alone." See e.g. Wiley (1981), Watrous & Wheeler (1981) and Maddison et al. (1984) for further information and discussion of outgroup comparison and alternative methods of polarity determination.

Nowadays a large number of tree building procedures are available which can be classified into parsimony, compatibility and maximum likelihood techniques. These will be shortly presented below: See e.g. Wiley et al. (1991), Forey et al. (1992) and Quicke (1993) for more detailed discussions on the various approaches.

Parsimony approaches — Parsimony approaches aim at minimizing some measurement of tree length, that is, the number of evolutionary changes on the tree. The exact quantity to be minimized, the optimality criterion, depends on an underlying model of character evolution. A maximum parsimonious tree is a tree that is optimized for one of the criteria and in which the number of homoplasies are minimized and the number of synapomorphies are maximized.

Various parsimony procedures have been developed: Wagner parsimony, Fitch parsimony, Camin-Sokal parsimony, the polymorphism parsimony method. Swofford & Olson (1990) have developed the generalized parsimony method in which it is allowed to give weights (assign costs) to specific character state changes. In the case of Camin-Sokal parsimony, where reversals are not allowed, reversals are associated with infinite cost. The above mentioned parsimony procedures can all be considered special cases

of the generalized method. See Forey et al. (1992: table 4.1) for examples of cost-matrices.

Wagner parsimony and Fitch parsimony are the two methods that are most often employed. They are rather simple in their evolutionary assumptions. Dollo parsimony and the polymorphism procedure are only used for some type of data and also the Camin-Sokal parsimony procedure is rarely used. Also the general parsimony method may have advantages in some cases. On the other hand, differential weighting of character state changes is a problem analogous to character weighting and should be carefully considered before application (Forey et al., 1992; Wiley et al. 1991).

In the search for the most parsimonious tree, three searching strategies can be followed. During an exhaustive search all possible tree topologies are evaluated and the shortest tree is guaranteed to be found. As the number of possible tree topologies increases enormously with the number of taxa, this approach is only practical for data sets with not more than 10 taxa. Branch-and-bound methods are also guaranteed to find the minimum length tree but, unlike in the exhaustive search, not all possible topologies will have to be evaluated. Simply put, the tree length is calculated each time after the addition of a new taxon during the tree-building procedure. Does the tree length exceed the current minimum length, then there is no need to continue the current path. Branch-and-bound techniques make it possible to find minimum length trees in cases of up to 25 taxa. Heuristic methods start off with an initial tree, and through a process of rearranging branches the program tries to improve the tree. The initial topology is built by adding taxa in one of several ways (*As is, Random, Simple, Closest*). There is, however, the risk that a solution will be found which is a local optimum instead of a global one. In order to escape from a local optimum branch swapping may be applied. This involves rearranging the branches in the tree. Heuristic methods are not guaranteed to find the most parsimonious trees. For more details on branch swapping and addition sequences, see e.g. Swofford & Olson (1990) and Forey et al. (1992).

Compatibility approaches — Compatibility or clique analysis searches for the largest set (clique) of (true) characters that are mutually compatible. Characters that are in conflict with the largest set are false characters which are not informative as to evolutionary relationship. These characters are not used in further analysis and this seems to be an important drawback of the method. The major criticism of this procedure is (Forey et al., 1992) "that the tree constructed from the largest clique may be quite unparsimonious globally ...".

Zandee & Geesink (1987) have developed an approach for cladogram construction that incorporates elements from both parsimony and compatibility analysis. In contrast with the character compatibility method they refer to their method as group compatibility analysis. The following very abbreviated outline is extracted from Zandee (1984, 1987). As a first step building blocks or clada are constructed from the data matrix. There are two options to define clada: partial monothetic sets are defined by sets of unique character states; strict monothetic sets are defined by unique combinations of character states. Cladograms are built from these clada by a process of three-cladon statement permutations combined with local outgroup comparison. The analysis results in sets of non-overlapping clada; thus the procedure is based on the concept of cladon-compatibility. From the set of cladograms thus produced the 'best' ones can be selected using a number of selection criteria: 1) total of homoplasious states, 2) total of supporting states (fit), 3) homoplasy minus support, 4) total number of state changes, 5) redundancy and 6) consistency. This primary form of the group compatibility method may be extended with a so called secondary analysis to further analyze cladograms that are not fully resolved. Polychotomies may be dichotomized, but this will be at the expense of additional ad hoc statements. Like the method of character compatibility, one disadvantage of Zandee's method seems to be that its solutions can be considerably unparsimonious.

Maximum likelihood approaches — Maximum likelihood is a method developed for cases in which we have more information about the evolutionary process, that is, maximum likelihood requires a model describing the probabilities of evolutionary change. The optimum tree to be discovered is the tree that gives the highest probability of a data set being derived from it, given the probabilistic model. Maximum likelihood methods are only practical for small data sets of molecular characters. See Swofford & Olson (1990), Forey et al. (1992) or Quicke (1993) and references therein for more detailed discussions.

Evolutionary taxonomy

As Wiley (1981) pointed out, the approach of evolutionary taxonomy is difficult to define in a straightforward manner because "it is a heterogeneous discipline or an array of different points of view more than it is a method or system united by a single body of theory". The most important difference with the phylogenetic approach is the fact that non-natural, paraphyletic grouping are allowed in its classifications. The contrast between the two schools lies mainly in the way that classifications are constructed from a phylogenetic tree, rather than in how we should proceed to arrive at this tree. Evolutionary taxonomists too want their classifications to reflect evolutionary history and they also agree that classifications should be based on genealogy. However, not only the branching pattern but also the subsequent diverging of branches (anagenesis) should be reflected in a classification. This is a view already expressed by Darwin (1859): "Thus, on the view which I hold, the natural system is genealogical in its arrangement, like a pedigree; but the degrees of modification which the different groups have undergone, have to be expressed by ranking them under different so-called genera, sub-families, families, sections, orders, and classes." Evolutionary taxonomists are of the opinion that grades are also natural groups that really exist in nature. See e.g. Mayr (1981) for an evolutionist's view on the three schools of taxonomy.

Optimal trees and optimal classifications

Why bother about phylogeny?

The major task of taxonomy is to recover the historical course of evolution by unravelling genealogical relationships between species. A phylogenetic tree may be viewed as a graphic representation of this historical course. However, other areas of biology may also benefit from the products of taxonomic studies, i.c. phylogenetic trees displaying these genealogical relationships. Phylogenetic information is crucial for what is called the comparative method. To repeat Brooks & McLennan (1994): "It is therefore inappropriate to use the results of a non-phylogenetic systematic analysis as a phylogenetic tree in a comparative study". Historical biogeography (phylogenetic biogeography, vicariance biogeography, component analysis), the study of coevolution, palaeontology and other disciplines within biology are all heavily depending on phylogenetic information.

Desirable properties of classifications

What will be considered as desirable properties of classifications depends largely on the purpose of a classification. Classifications are of vital importance as storage-and-retrieval systems. Without such systems it would be impossible to specify what is being studied and to scan the literature for all kinds of information about the organisms being studied. It is clear that a classification, as a reference system, should be as stable as possible. Classifications are also expected to have some predictive value. This means that, if we have discovered a new species which we were able to classify to a known higher taxon based on some of its characters, we may predict the states of some other characters that were not studied. Also objectivity is sometimes listed as one of the desirable properties of classifications, which merely implies that there must be a standard method by which

a classification can be or should be constructed (e.g. Pankhurst, 1991).

For the majority of organisms no phylogenetic trees are available. In the absence of these, comparative biologists, biogeographers and in fact all biologists use (traditional) classifications as an information source for phylogenetic relationships. Or, to quote Wiley et al. (1991): "Most of the ideas of relationship that exist in the literature are embodied in classifications". Actually, many biologists would agree that classifications should be based on phylogeny. It is clear then, that the quality of the results of many of their studies, depends heavily on how accurately their classifications reflect phylogenetic relationships. From this two further and most important optimality criteria for classifications can be deduced. Classifications should fit to the true tree as close as possible, that is, classifications should be consistent with the phylogeny on which they are based. A classification is said to be consistent with phylogeny "if at least one of the possible phylogenies implied by it is the original phylogeny from which it was constructed" (Hull, 1964; see also Wiley, 1987; Wiley et al., 1991). As a consequence, to quote Wiley (1987) again, "all classifications containing even a single paraphyletic group are logically inconsistent with the phylogenies they are supposed to reflect and/or summarize." As a second prerequisite, classifications should be "informative regarding the common ancestry relations of the groups classified" (Wiley et al., 1991). A classification that is fully informative is one from which one can extract as much information as from the phylogenetic tree on which the classification was based. Such classifications are called isomorphic.

Evolutionary classifications, although indeed based on phylogeny, contain paraphyletic groupings. These kinds of classifications try to reflect not only the branching sequence, but also divergence within lineages. As a result they become inconsistent with phylogeny.

Natural groups as viewed by pheneticists are not necessarily identical with the monophyletic groups of cladists, and pheneticists

do not strive to construct classifications that mirror evolutionary history. Phenetic classifications may be consistent with the phenograms on which they are based, but they have a probability of containing non-monophyletic groupings in case of which they do not reflect evolutionary history. Even Darwin (1859) held a clear view on the significance of overall similarity for classification purposes:

"No one regards the external similarity of a mouse to a shrew, of a dugong to a whale, of a whale to a fish, as of any importance. These resemblances, though so intimately connected with the whole life of the being, are ranked as merely adaptive or analogical characters; ... the less any part of the organisation is concerned with special habits, the more important it becomes for classification."

It goes beyond the subject of this introduction to elaborate on how classifications can best be constructed. However, I agree with the three rules of phylogenetic classification as formulated by Wiley et al. (1991:102): 1) only monophyletic groups will be formally classified; 2) all classifications will be logically consistent with the phylogenetic hypothesis, and 3) classifications must be capable of expressing the sister group relationships among the taxa classified. So, any classification containing non-monophyletic groups should be rejected as a general reference system. In the absence of well corroborated phylogenies, also Miles & Dunham (1993) advise not to use taxonomic classifications as an alternative, because one cannot be sure that these classifications reflect phylogeny in an accurate manner. And once again Darwin (1859) can be cited, who wrote: "... the natural system is founded on descent with modification; ... the characters which naturalists consider as showing true affinity between any two or more species, are those which have been inherited from a common parent, and, in so far, ***all true classification is genealogical*** [my italics]; ...community of descent is the hidden bond which naturalists have been unconsciously seeking, and not some unknown plan of creation, or the enunciation of general propositions, and the mere putting together and separating objects more or less alike."

Is there a best way to uncover evolutionary history and is there a best way to classify, or, in other words, should one be an evolutionary taxonomist, a pheneticist or a cladist? Based on the foregoing considerations I would conclude that one should be a cladist: it is the task of taxonomy to discover natural, that is, monophyletic taxa. Monophyletic taxa can only be discovered by the use of apomorphic characters. Classifications should be natural classifications, that reflect evolutionary history and that are consistent with the phylogenetic tree from which they were derived.

Phylogenetic trees are hypotheses on the evolutionary history of species and taxa and constitute the foundation for phylogenetic classifications. As such they are a fundamental product of taxonomy. Present day diversity is the result of a unique history of descent with modification: there is only one true tree of life. It is this one true tree that taxonomists must try to discover and for which they have a large number of tree building techniques at their disposal. These phylogenetic trees form the basis for classification and provide a context for comparative biology, biogeography and other branches of biology, the results of which can only be as good as the phylogenetic trees on which they are founded. Or, to paraphrase Benton (1990): phylogenetic trees are the keys to determining why life is as it is. Therefore it is crucial to know how much confidence we can have in the accuracy of phylogenetic estimations.

Factors affecting the quality of phylogenetic estimations

The minimum requirement for a phylogenetic analysis to be able to produce a fully resolved tree would be a data matrix with a number of (binary coded) characters equal to the number of bifurcations minus one. Sokal (1983) has proposed to measure this aspect of the adequacy of the characters for resolving a cladogram by the ratio n/t , where n is the number of characters and t the number of OTUs (terminal taxa). If there would be no homoplasy in the data, if all characters would be mutually compatible, if all characters would

have been polarized correctly and if there would be no character correlation (if each bifurcation would be 'covered' by one apomorphy), then a phylogenetic analysis would definitely produce the correct tree. In practice, of course, such an ideal data matrix will never be available. Probably the two most important factors that may add phylogenetic *noise* to a data matrix are missing data and homoplasy. No data set will contain all possible characters. Missing characters with a high level of homoplasy, would pose no problem. However, we can only discover uninformative characters after we have performed a phylogenetic analysis. Therefore missing characters refer to any character or set of characters that is not used in the study, but that, if included in the data matrix, could affect the final tree topology (Eggleton & Vane-Wright, 1994). In analogy with missing characters, we can define missing taxa as taxa that were not included in the analysis, but that, if included, could effect tree topology. Taxa may be removed from phylogenetic analysis as a result of extinction. Extinct taxa can sometimes be used when they are still available as fossils. However, we can never be sure to have sampled all taxa, extinct or extant, within the monophyletic group under study. Fossils may provide extra information on character transformation; character interpretation may change when fossils are allowed in the analysis. There are examples that the inclusion of fossil taxa produced quite different results: e.g. Novacek (1992) and Wheeler (1992) point out that the topology of cladograms may be affected by the addition of new taxa (fossils) and this may be a serious problem since virtually all published trees are derived from extant taxa (Eggleton & Vane-Wright, 1994).

It is evident that convergence and parallel evolution (homoplasy) are important *noise*-producers that are capable of obscuring the phylogenetic *signal*. At the same time, homoplasy can only be discovered during a phylogenetic analysis, and this is even considered one of the major powers of parsimony analysis (Stewart, 1993). Too much noise can result in a most parsimonious tree being quite incorrect. To quote Stewart (1993:606): "Homoplasy in its various disguises, is the ultimate trickster of parsimony" and also

Dawkins (1986:269) noted that "The most interesting **bugbear**¹ of the taxonomist is evolutionary convergence." Application of the parsimony criterion implies that homoplasy is sufficiently rare and should be minimized in the final hypothesis. Nevertheless, evolution need not be parsimonious and the true level of homoplasy might be too high to permit successful reconstruction of the true tree.

How to measure the adequacy of phylogenetic estimations? — There are various ways of estimating the quality of a phylogenetic tree. One way is to measure the fit between data and tree. Two of the most used indices to measure how well the characters fit the tree are the consistency index (*CI*) and the retention index (*RI*). The consistency index measures the amount of homoplasy in the data; in the absence of homoplasy *CI* will have its maximum value 1. The retention index measures the actual amount of homoplasy as a fraction of the maximum amount possible. For phenograms a so-called cophenetic correlation coefficient (r_{cs} , Sneath & Sokal, 1973) can be calculated which measures the fit between the similarity values in the similarity matrix and those that can be deduced from the phenogram.

Felsenstein (1985) has developed bootstrap and jack-knife techniques to arrive at confidence limits for trees and also some other approaches have been advocated (see e.g. Forey et al. (1992) for a short discussion).

These measures and approaches may help us to decide which tree to choose from a set of competing trees, but we still cannot really judge the reliability of the resulting estimations. Since there is only one single true tree, we would need to know this true phylogeny in order to tell whether our estimates are correct or not.

There are a few cases in which a true phylogeny was known and which could be compared with the estimated ones (Baum, 1983, 1984; Fitch & Atchley, 1987; Hillis et al., 1992). In Hillis et al. (1992) an experimental approach to phylogenetics was proposed; a

¹ bugaboo

known phylogeny was generated of lineages derived from bacteriophage T7, by manipulation through the use of mutagens. This approach was disputed by Sober (1993). The ability of estimation methods to reconstruct the true topology of the artificial phage system need not be relevant to systems from nature. Besides, this experiment represented only a single simple evolutionary scenario. The question then is how the estimation methods would behave if the phage system had been made to evolve under another evolutionary model. Similar arguments can be put forward in the other cases of known phylogenies of real organisms.

In general, however, phylogenies of real organisms are unknown. In fact, tree building techniques, the quality of which we would want to inspect, were developed to estimate these unknown phylogenies. The only way to really assess phylogenetic accuracy is through the application of artificial data sets produced by simulations. By using computer simulations the relative efficiencies of phylogenetic estimation methods can be estimated under a wide variety of evolutionary conditions.

Aims and outline of this study

The principle objective of the current project is to develop and use a simulation model that is able to generate known phylogenies of imaginary species and to simulate the evolution of their character states. The character data of the resulting 'extant' species can be used as input for different estimation procedures and the estimated phylogenies can be compared with the single true tree. The agreement with the true tree can serve as an indication of the quality of the methods tested. The basic outline of such an evaluation experiment is given in figure 1.1. Because of the stochastic nature of the simulation model, many replicate simulations must be run for each evolutionary condition.

In chapter 2 the simulation model, called GENESIS, is presented. GENESIS offers options to simulate the evolutionary

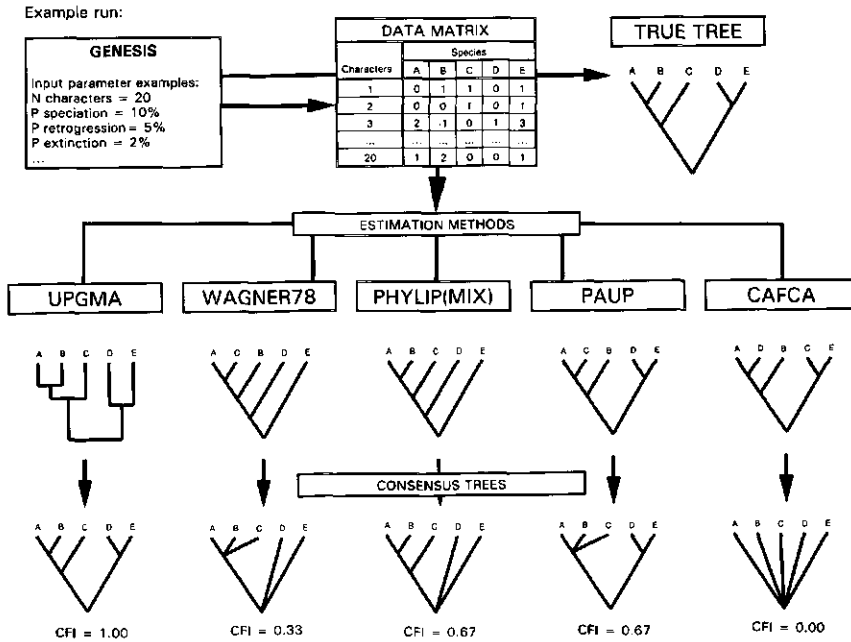


Figure 1.1 Simplified outline of an evaluation experiment.

process under a variety of evolutionary conditions. The results - true trees and their corresponding data matrices - of each simulation run are characterized by a number of descriptive statistics. In chapter 3 the results of an extensive sensitivity analysis of GENESIS are presented. Such an analysis reveals how the describing tree statistics are affected by changes in the input parameter values. Although it is impossible to examine the entire parameter space, such an analysis is a prerequisite for a proper understanding of GENESIS' performance under various evolutionary conditions. In chapters 4, 5 and 6 results of experimentations with GENESIS are presented; data sets produced by GENESIS were subjected to an analysis by a number of numerical taxonomic procedures. The adequacy of the estimation techniques was evaluated under a variety of evolutionary conditions. Chapter 7 presents a general discussion and in chapter 8 a recapitulation of the results is given.

References

- Ax, P., 1987.** *The phylogenetic system. The systematization of organisms on the basis of their phylogenies.* John Wiley & Sons, Chichester.
- Baum, B. R., 1983.** Relationships between transformation series and some numerical cladistic methods at the infraspecific level, when genealogies are known. In: *Numerical taxonomy* (J. Felsenstein, ed.): 340-345. Springer-Verlag, Berlin.
- Baum, B. R., 1984.** Application of compatibility and parsimony methods at the infraspecific, specific, and generic levels in Poaceae. In: *Cladistics: perspectives on the reconstruction of evolutionary history* (Th. Duncan & T. F. Stuessy, eds): 192-220. New York, Columbia University Press.
- Benton, M. J., 1990.** Phylogenetic trees and the unification of systematic biology. — *Trends in ecology and evolution* 5: 392-394.
- Brooks, D. R. & D. A. McLennan, 1994.** Historical ecology as a research programme: scope, limitations and the future. In: *Phylogenetics and ecology* (P. Eggleton & R. Vane-Wright, eds): 1-27. Academic Press, London.
- Clifford, H. T. & W. Stephenson, 1975.** *An introduction to numerical classification.* Academic Press, New York.
- Colless, D. H., 1970.** The phenogram as an estimate of phylogeny. — *Systematic Zoology* 19: 352-362.
- Darwin, Ch., 1859.** *The origin of species by means of natural selection or the preservation of favoured races in the struggle for life.* Reprinted: Penguin Books Ltd, Harmondsworth, England, 1968, edited with an introduction and bibliography by J. W. Burrow.
- Eggleton, P & R. I. Vane-Wright, 1994.** Phylogenetics and comparative biology. In: *Phylogenetics and ecology* (P. Eggleton & R. Vane-Wright, eds): 345-366. Academic Press, London.
- Eldredge, N., 1992.** Where the twain meet: Causal intersections between the genealogical and ecological realms. In: *Systematics ecology, and the biodiversity crisis* (N. Eldredge, ed.): 1-14. Columbia University Press, New York, Oxford.
- Felsenstein, J., 1978.** The number of evolutionary trees. — *Systematic Zoology* 27: 401-410.
- Felsenstein, J., 1985.** Confidence limits on phylogenies: an approach using the bootstrap. — *Evolution* 39: 783-791.
- Felsenstein, J., 1991.** *PHYLIP 3.4 manual.* University of California Herbarium, Berkeley, California.

- Fitch, W. M. & W. R. Atchley, 1987. Divergence in inbred strains of mice: a comparison of three different types of data. In: *Molecules and morphology in evolution: conflict or compromise?* (C. Patterson, ed.): 203-216. Cambridge University Press, Cambridge.
- Forey, P. L., C. J. Humphries, I. J. Kitching, R. W. Scotland, D. J. Siebert & D. M. Williams, 1992. *Cladistics. A practical course in systematics*. Systematics Association Publications, 10. Clarendon Press, Oxford.
- Hennig, W., 1966. *Phylogenetic systematics*. University of Illinois Press, Urbana, Illinois.
- Hillis, D. M., J. J. Bull, M. E. White, M. R. Badgett & I. J. Molineux, 1992. Experimental phylogenetics: generation of a known phylogeny. — *Science* 255: 589-592.
- Hull, D. L., 1964. Consistency and monophyly. — *Systematic Zoology* 13: 1-11.
- Maddison, W. P., M. J. Donoghue & D. R. Maddison, 1984. Outgroup analysis and parsimony. — *Systematic Zoology* 33: 83-103.
- May, R. M., 1990. How many species? — *Philosophical Transactions of the Royal Society of London B* 330: 293-304.
- Mayr, E., 1974. Cladistic analysis or cladistic classification? — *Zeitschrift für zoologische Systematik und Evolutionsforschung* 12: 94-128.
- Mayr, E., 1994. Biological classifications: towards a synthesis of opposing methodologies. In: *Conceptual issues in evolutionary biology* (E. Sober, ed.): 277-294. The MIT Press, Cambridge.
- Miles, D. B. & A. E. Dunham, 1993. Historical perspectives in ecology and evolutionary biology: the use of phylogenetic comparative analyses. — *Annual Review of Ecology and Systematics* 24: 587-619.
- Novacek, M. J., 1992. Fossils as critical data for phylogeny. In: *Extinction and phylogeny* (M. J. Novacek & Q. D. Wheeler, eds): 46-118. Columbia University Press, New York.
- Novacek, M. J. & Q. D. Wheeler, 1992. Introduction: Extinct taxa: accounting for 99.999...% of the earth's biota. In: *Extinction and phylogeny* (M. J. Novacek & Q. D. Wheeler, eds): 1-16. Columbia University Press, New York.
- Pankhurst, R.J., 1991. *Practical taxonomic computing*. Cambridge University Press, Cambridge.
- Quicke, D. L. J., 1993. *Principles and techniques of contemporary taxonomy*. Blackie Academic & Professional, London.
- Raup, D. M., 1981. Extinction: bad genes or bad luck? — *Acta Geologica Hispanica* 16: 25-33.

- Reid, W. V., 1992. How many species will there be? In: *Tropical deforestation and species extinction* (T. C. Whitmore & J. A. Sayer, eds): 55-73. Chapman & Hall, London.
- Sober, E., 1993. Experimental tests of phylogenetic inference methods. — *Systematic Biology* 42: 85-89.
- Stork, N. E., 1993. How many species are there? — *Biodiversity and Conservation* 2: 215-232.
- Sneath, P. H. & R. R. Sokal, 1973. *Numerical taxonomy. The principles and practice of numerical classification*. W. H. Freeman & Co., San Francisco.
- Sokal, R. R. & P. H. A. Sneath, 1963. *Numerical taxonomy*. W. H. Freeman & Co., San Francisco.
- Sokal, R. R., 1983. A phylogenetic analysis of the Caminalcules. I. The data base. — *Systematic Zoology* 32: 159-184.
- Swofford, D. L. & G. J. Olson, 1990. Phylogeny reconstruction. In: *Molecular Systematics* (D. M. Hillis & C. Moritz, eds): 411-501. Sinauer, Sunderland, Massachusetts.
- Watrous, L. E. & Q. D. Wheeler, 1981. The outgroup comparison method of character analysis. — *Systematic Zoology* 30: 1-11.
- Wheeler, W. C., 1992. Extinction, sampling, and molecular phylogenetics. In: *Extinction and phylogeny* (M. J. Novacek & Q. D. Wheeler, eds): 205-215. Columbia University Press, New York.
- Wiley, E. O., 1981. *Phylogenetics: the theory and practice of phylogenetic systematics*. John Wiley & Sons, New York.
- Wiley, E. O., 1987. The evolutionary basis for phylogenetic classification. In: *Systematics and evolution: a matter of diversity* (P. Hovenkamp et al., eds): 55-64. Faculty of Biology, Utrecht University, The Netherlands.
- Wiley, E. O., D. J. Siegel-Causey, D. R. Brooks, V. A. Funk, 1991. *The complete cladist: a primer of phylogenetic procedures*. Museum of Natural History, University of Kansas, Lawrence.
- World Conservation Monitoring Center, 1992. *Global biodiversity: status of the earth's living resources*. Chapman & Hall, London.
- Zandee, M. 1987. *C.A.F.C.A.: A Collection of APL Functions for Cladistic Analysis, version 1.8.6*; program and user manual.
- Zandee, M. 1987. A computing environment for cladistic analyses. Preliminaries of a user manual for CAFCA/PC. In: *Systematics and evolution: a matter of diversity* (P. Hovenkamp et al., eds): 111-140. Faculty of Biology, Utrecht University, The Netherlands.
- Zandee, M. & R. Geesink, 1987. Phylogenetics and legumes: a desire for the impossible? In: *Advances in legume systematics* (C. H. Stirton, ed.) 3: 131-167. Royal Botanic Gardens, Kew.

2

GENESIS: a simulation model of phylogeny

The origin and early evolution of character state vectors ¹

"This belief, that Darwinian evolution is 'random', is not merely false. It is the exact opposite of the truth." Dawkins (1986:49)

Abstract

A simulation model for phylogenesis is presented. The model, called GENESIS, creates a 'phantom' world of artificial species, which appear in the form of character state vectors. These species can be produced using different options of GENESIS, corresponding to different evolutionary scenarios. In other words, GENESIS can be used to produce data sets with different properties. The characteristics of the simulated evolutionary processes and their corresponding data matrices, are described by several tree statistics.

The data matrices can be subsequently used as test cases to evaluate the qualities of various methods of reconstructing phylogeny. These evaluations will be published in subsequent papers, as will also the results of a sensitivity analysis.

Introduction

If we knew the exact evolutionary history of a taxon, we could use this taxon as a test case to examine the merits of various phenetic

¹ Published as: Heijerman, Th., 1988. GENESIS: a simulation model of phylogeny. Part 1. The origin and early evolution of character state vectors. — *Zeitschrift für zoologische Systematik und Evolutionsforschung* 26: 609-622.

and cladistic approaches to cladogram estimation. However, it is not sufficient just to know the true phylogeny of this taxon, one also needs a character matrix which must satisfy a number of conditions. There must be sufficient characters that reflect the true evolutionary history of the taxon, and at the same time, there may not be too many 'bad' characters, that are incompatible with the true tree and that obscure the information contained by the 'good' characters. In other words, the data matrix should be sufficiently 'clean', and also contain enough information to produce a fully resolved tree. So, an ideal data matrix will appear to have a consistency index (C/I) equal to one, and every monophyletic group must be based on at least one synapomorphic character state. Thus the internal length, expressed as the number of character state changes on the non-terminal segments, must be at least equal to the number of internal segments, that is to $S - 2$, where S is the number of species.

Although there may be some rare cases in which the true phylogeny is known (e.g. Baum, 1984), one can never be sure of obtaining an ideal data matrix. Suppose, nevertheless, that we could find a good test case, then another difficulty would arise. We still cannot be sure whether a certain selected method will also give the best estimate of relationships between members of another taxon whose actual evolutionary course we do not know. This other taxon may have evolved under different and unknown evolutionary conditions. The quality of the available reconstruction methods will probably depend on the rate and exact pattern of its evolution. Moreover, the performances of the various reconstruction methods will be affected by the number of characters with incorrectly determined character state sequences, the number of missing recent species, and other 'artefacts' of the data matrix to be analyzed.

To evaluate and to estimate the quality of the various phylogenetic and phenetic methods, a simulation model of phylogenesis, called GENESIS, was developed.

GENESIS is designed to produce patterns of diversity and of character state distributions. Details of the process of development of these patterns can be controlled by the user of the model. Thus

GENESIS can produce sets of recent species with known phylogenies. These sets of species can be subjected to the various phylogeny reconstruction methods.

GENESIS can be made to produce ideal data matrices. One may expect every method to perform excellently on these perfect sets of data. One can subsequently produce data sets containing incompatible characters, by allowing back-mutations and homoplasous changes in character states. The *noise* produced by these 'bad' characters will make it a more difficult task for the reconstruction methods to find the true phylogeny indicated by the 'good' characters. The performance of the methods tested will depend, in different ways, on the amount of *noise* in the data set to be analyzed.

The simulation technique also allows for another kind of tests. By changing the values of input parameters and by selecting certain options of the model, one can control the behaviour of the system and thus produce sets of recent species based on different notions about the evolutionary process. Also these sets can be used as test cases.

Simulation models of evolution have been developed earlier. Anderson & Anderson (1975) used stochastic simulation models to examine the evolution of patterns of taxonomic diversity and showed that "the real-world diversity pattern could have been produced by a simple stochastic process ...". Raup (1977) briefly described a simulation model of the evolutionary branching process, making only a few biological assumptions. He used his model "primarily as an exploratory tool", concluding that if "a pattern commonly observed in the real world can be simulated readily by the program, then causes not included in the MBL model need not be called on for explanation and interpretation" (also Raup et al., 1973 and Raup & Gould, 1974). Tateno et al. (1982) simulated the evolutionary changes in nucleotide sequences for eight species and a given model tree, to examine "the accuracies and efficiencies of three different methods of making phylogenetic trees from gene frequency data ...". Also Sokal's Caminalcules (1983) and Wagner's Dendrogrammeaceae

(Duncan et al., 1980) can be regarded as the results of a single simulation of the evolutionary process. Leman & Freeman (1984) created evolutionary models to be used to predict "size and shape variation in families and genera under different evolutionary assumptions". Colwell & Winkler (1984) used and modified the models of Raup et al. (1973) and Raup & Gould (1974). Their model, named GOD, produces recent species with known phylogenies, and by using a subsequent program, WALLACE, they introduced patterns of geographic distributions into their simulations.

Fiala & Sokal (1985) developed a simulation model and used it to examine the accuracy of three taxonomic methods. Their model is rather like GENESIS, and was also developed for the same purpose, i.e. to compare estimated and true, though simulated, phylogenies. Although the model was already made public in their 1985 paper, the development of their model and of GENESIS have really been independent processes, and many of the similarities between the two models are indeed cases of homoplasy. Fiala & Sokal (l.c.) have provided their simulation program with a probability model to obtain random amounts of change in the character state (null) change probabilities. In their model reversals from the ancestral state are not allowed, whereas in GENESIS these reversals can freely occur. Fiala & Sokal have conducted their experiments using 20 OTUs, each possessing 25 characters. The simulations described in this paper have all been run with 50 OTUs, each with 100 characters. However the most conspicuous differences are in the technical design of the models. But by selecting the proper input-parameter values, comparable 'evolutionary context patterns' can be created.

Fiala & Sokal (1985:612) state that it "is neither practical nor desirable to explore the entire parameter space of the simulation model." Because of the many possible sets of input parameter values, one indeed cannot explore the entire parameter space of their model nor of GENESIS. But it seems possible and necessary to come to a better appreciation of the potentials of the model by studying its behaviour in different evolutionary contexts, as created by different sets of input parameter values. Because of the stochastic nature of

the model, many replicate simulations will have to be run for each separate set of input parameter values.

In this paper the general design of GENESIS is described, and results of some preliminary simulation runs are presented. The results of an extensive sensitivity analysis and of the evaluations of the various phylogeny reconstruction methods, will be presented in subsequent papers.

The simulation model: what exactly does it simulate?

Evolution may be defined as genealogical descent with modification, and involves the process of lineage branching (speciation) and lineage termination (extinction), as well as the process of character state change. Lineage splitting and character evolution are the basic processes that are simulated by GENESIS.

Furthermore, the evolutionary process may be considered a sequence of system conditions ordered in time. The condition or state of the system may be described, at any one moment, by a set of condition parameters, e.g. the number of recent species, the state of a certain character. This state of the system is the joint outcome of many separate events. Such events, for instance speciation and character state change, occur at a given time, with a certain probability. Thus, the outcome of the evolutionary process cannot be predicted exactly; the process must be viewed as being probabilistic instead of deterministic. The condition of the system at any one moment in the time sequence can be considered the joint result of the condition in the previous time step and the random processes. To simulate such a process, a stochastic simulation model is needed.

GENESIS can simulate evolution according to different conceptions of the evolutionary process, resulting in different patterns. There are two mathematical models that are often used to describe the pattern of diversification of the number of taxa. The simpler model involves an exponential increase in the number of taxa. In a second model the increase in number of taxa is described by a

logistic curve. An exponential curve may be expected best to describe the situation during phases of adaptive radiation, whereas the logistic model may be more appropriate during the later stages of the evolution of a taxon (e.g. Stanley 1977, 1985). Although for instance the histories of the family diversity of the Mammalia and Bivalvia, as presented by Stanley (1985, his fig. 4), seem to be good examples of the two models, they can only be used as approximations of what really happened in nature. For this pattern of diversity, GENESIS offers both a 'radiation' option and an 'equilibrium' option.

The number of species is allowed to increase exponentially with time in the radiation model, whereas this number is made to fluctuate in a random way around an equilibrium value in the 'equilibrium' version. Rates of evolutionary change imply rates of changes in the number of taxa and in evolutionary character states in the different lineages. These rates would vary between and within lineages. In GENESIS these evolutionary rates will differ only by chance.

Some data is available on rates of change of characters, e.g. the dimensions of the first lower molar of species of the genus *Pelycodus* (Gingerich, 1977). However rates of change in the total morphology of a species seem difficult to assess. Adaptive radiation is supposed to involve quantum evolution. Quantum evolution was first defined by Simpson (1944) as "the relatively rapid shift of a biotic population in disequilibrium to an equilibrium unlike an ancestral condition." Such shifts probably involve relatively rapid rates of change in characters. Eldredge & Gould (1972) recognized two patterns concerning the tempo and mode of evolution. Phyletic gradualism, the classical model, implies amongst other things, that new species can arise by transformation of an ancestral population. This transformation into the modified descendants is slow and even in all lineages. As an alternative, Eldredge & Gould (1972) presented the punctuated equilibrium hypothesis. Their model assumed that new species arise only by the splitting of a small sub-population from the ancestral one, thus by the splitting of lineages. And, moreover, the new species develop rapidly. It seems, however, that both phylogenetic patterns can be recognized in the fossil record. Nevertheless, there is still an

animated discussion in the literature about the pros and cons of the two models (e.g. Gould & Eldredge, 1983; Gingerich, 1984; Scudo, 1985).

The gradualistic and the punctualistic models are both available as options in GENESIS. In the 'gradualistic' version, character state changes occur at equal rates in both daughter lineages, whereas in the 'punctualistic' version (punctuated equilibrium model), these rates are unequal.

In their ultimate consequences, the two models lead to different kinds or concepts of species: chrono-species and cladistic-species, respectively. In GENESIS, however, species can only arise by lineage splitting (cladistic-species) and not by gradual transformation from the ancestral population (chrono-species).

Thus, GENESIS simulates a random process through time, during which lineages branch dichotomously or terminate, and characters change from one state into the other. The evolutionary process starts with one species at time zero. This species is the ancestor of all species subsequently generated, and all of its characters occur in the ancestral state, designated by "0". At every time step, the model makes several decisions about branching and evolution of single characters. All decisions are controlled by probability parameters, the values of which must be supplied by the user of the program.

Evolution of the branching pattern

Lineages terminate by extinction, speciation (pseudo-extinction) or termination of the simulation process. At every time step, Δt , the status of each species that has been generated so far is evaluated. If a certain species is extinct or already an ancestral species, the next species will be considered. For a recent species, there are three possibilities. The species becomes extinct, the species survives and splits into two daughter species (cladogenesis), or the species survives while no speciation occurs. Extinction and speciation are controlled by the extinction probability, P_{ex} , and the speciation

probability, P_{split} respectively. These are the only parameters that determine the branching pattern and thereby the diversity pattern of a simulation.

The radiation version of GENESIS — If speciation and extinction probabilities are constant over time, the change of the number of lineages through time may be described by the exponential curve:

$$N_t = N_0 \bullet e^{R \bullet t} \quad (1)$$

where N_t is number of recent species at time t ,
 N_0 is number of recent species at time 0 ($N_0 = 1$),
 t is the time,
 R is the relative rate of change,
 e is the base of the natural logarithm.

Equation 1 is only valid when time intervals are infinitely small. Otherwise equation 2 will apply:

$$N_t = N_0 \bullet (1 + R \bullet t/n)^n \quad (2)$$

where n is the number of time intervals.

The relative rate of change (R) is dependent on P_{ex} and P_{split} . So equation 1 can be written as:

$$N_t = N_0 \bullet e^{(N_{split} - N_{ex})t} \quad (3)$$

and equation 2 then becomes:

$$N_t = N_0 \bullet (1 + (N_{split} - N_{ex})t/n)^n \quad (4)$$

N_{split} and N_{ex} must be expressed as the number of speciations or extinctions respectively, per Δt . In time-homogenous models (Raup,

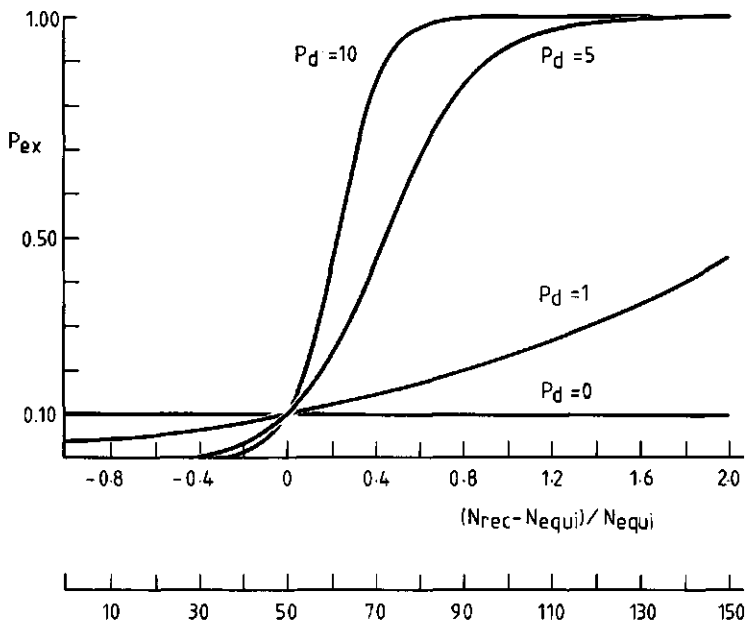


Figure 2.1 Probability of extinction (P_{ex}) as a function of the difference between the number of extant species (N_{rec}) and the equilibrium number of species (N_{equi}), for different values of the damping parameter (P_d). Probability of speciation (P_{split}) equals 0.10. On the lower x-axes the number of recent species is given when N_{equi} equals 50.

1985), N_{split} , or the probability of a speciation event (P_{split}), and N_{ex} , or the probability that a species will become extinct (P_{ex}), are constant over time. It can be shown (Raup, 1985, and literature cited there) that a clade will become extinct if $P_{split} \leq P_{ex}$. If $P_{split} > P_{ex}$, the probability of ultimate extinction will be $P_0 = (P_{ex}/P_{split})^{N_0}$. With an increasing value of $|P_{split} - P_{ex}|$ the probability of extinction decreases, as does the species diversity. In other words, the number of recent species will tend to become infinite largely within a finite time span.

The equilibrium version of GENESIS — In the 'equilibrium' model the number of species fluctuates around an equilibrium number of recent species, N_{equi} , the value of which must be specified by the user. This fluctuation is achieved by letting P_{ex} be a function of the difference between the number of recent species, N_{rec} , and N_{equi} :

Chapter 2

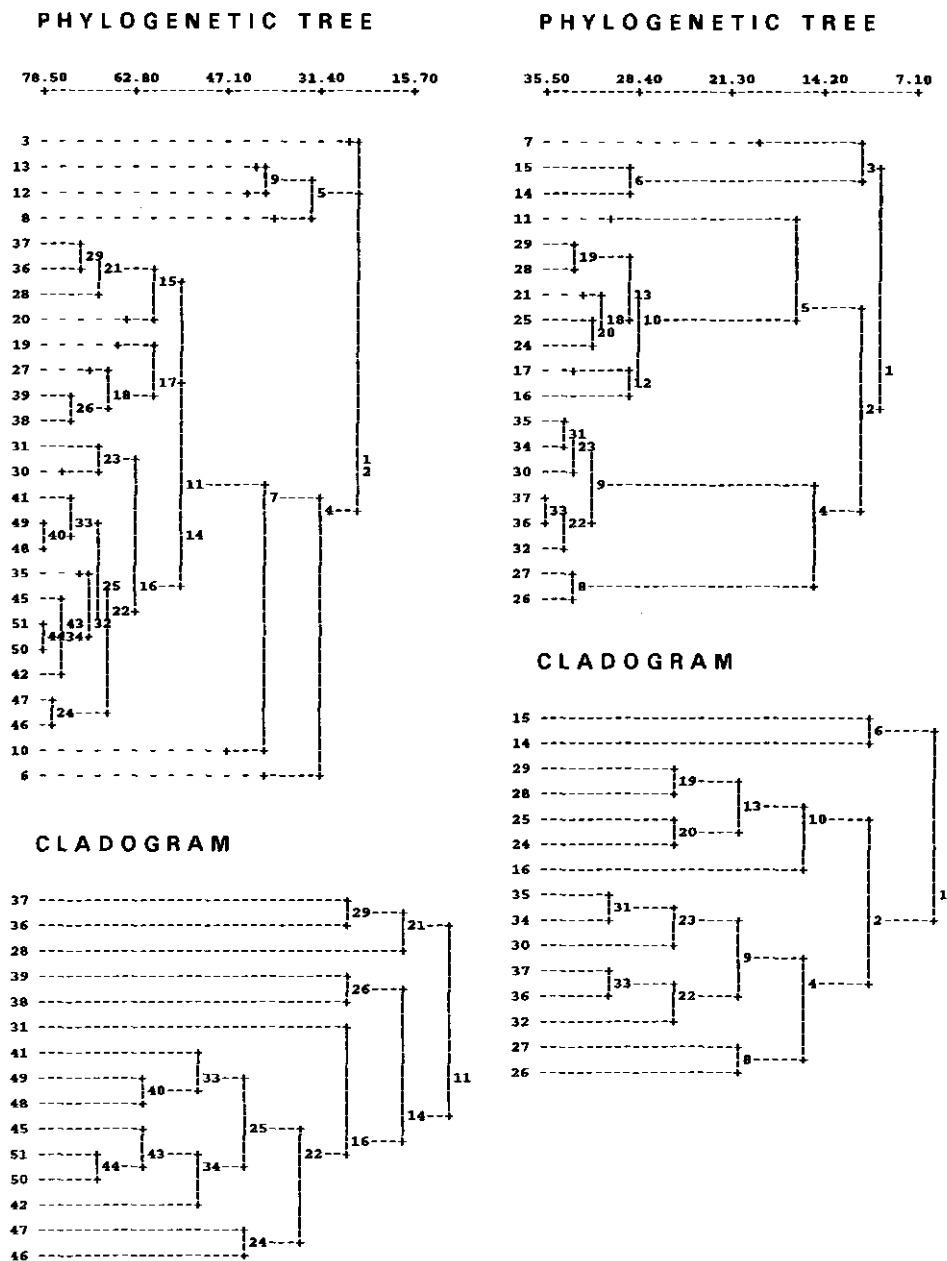


Figure 2.2 Two examples of phylogenetic trees and corresponding cladograms produced by GENESIS. The trees were generated using the 'radiation' version and the same values for the input parameters. $P_{split} = 0.10$, $P_{ex} = 0.04$. Differences between the two topologies are only caused by chance.

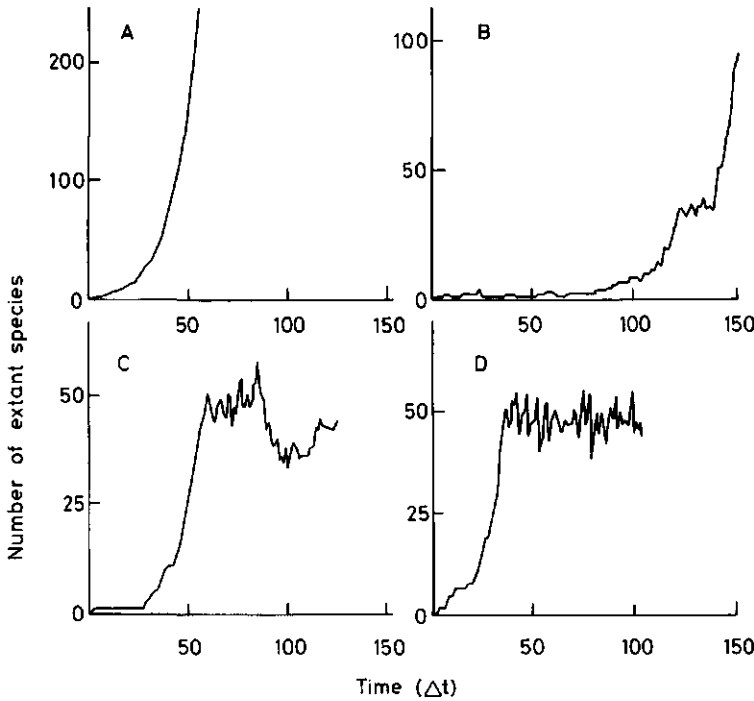


Figure 2.3 Diversity of species through time. Examples of results of different versions (scenario's) of GENESIS. A: Radiation, $P_{ex} = 0.0$; B: Radiation, $P_{ex} = 0.06$; C: Equilibrium, $N_{equi} = 50$, $P_d = 1.0$; D: Equilibrium, $N_{equi} = 50$, $P_d = 10$.

$$P_{ex} = f((N_{rec} - N_{equi})/N_{equi}) \tag{5}$$

At equilibrium, P_{ex} is made to equal P_{split} . One may wish to control the strength of this feedback mechanism, which is achieved by the use of a 'damping parameter', P_d (Fiala, 1983). Equation 5 then becomes:

$$P_{ex} = 1/(1 + ((1/P_{split}) - 1)(e^{- (P_d \cdot (N_{rec} - N_{equi})/N_{equi})})) \tag{6}$$

where parameters are as defined above. Such a model may be called a time inhomogeneous model (Raup, 1985).

By choosing different values for P_d one can change the strength of the feedback. If $P_d = 0$, there will be no feedback at all and P_{ex}

will equal P_{split} . With increasing values for P_d the feedback mechanism acts more powerfully (fig. 2.1).

The 'equilibrium' model includes two versions. In one version, the feedback mechanism is operative immediately from the start of the simulation at time zero. In the other version, the feedback mechanism will be activated only when N_{rec} has first become greater than N_{equi} . Before that moment, the model uses a constant probability of extinction, which must have been supplied by the user. The model thus initially operates according to the 'radiation' version.

Figures 2.2 and 2.3 show some results produced by different versions of the simulation model.

Character evolution

Character evolution may occur in two contexts: 1) in the anagenetic process (phyletic evolution) and 2) in the cladogenetic process. Anagenesis is the process of change of characters independent of the cladogenetic or branch-splitting process.

At every time step, each species still in existence will be considered and its character state vector updated; every single character may undergo random anagenetic change. This process is controlled by the character change probability, P_{ana} . Changes occur in discrete steps of uniform size. The direction of change is determined by the retrogression parameter, P_{ana} .

Another character change probability parameter, P_{clado} , controls character evolution occurring in the second context. Immediately after a branching event, i.e. in the same time interval and only during that time interval, every character of the two daughter species may change from one state into another. Again, there is a retrogression probability parameter, P_{clado} . In the 'gradualistic' version of GENESIS, the values of these parameters will be the same for the two daughter lineages. In the 'punctuated-equilibrium' version of the model, the values of these parameters only apply to the processes occurring in one of the daughter lineages. The corresponding parameters for the other lineage must be specified separately (fig. 2.4).

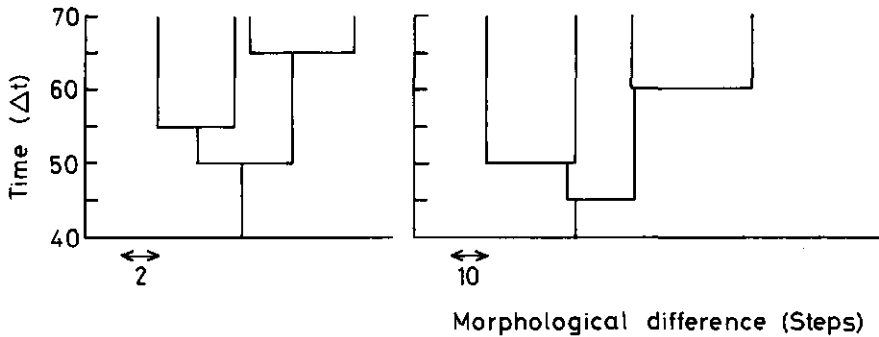


Figure 2.4 Parts of two phylogenetic trees as produced by GENESIS to illustrate the differences between the 'gradualistic' version (left) and the 'punctuated' version (right). In the 'gradualistic' version the average rate of character evolution is the same for both daughter lineages, whereas in the 'punctuated' version these rates are unequal. In the example, evolution proceeds 10 times as fast in one daughter lineage than in the other (punctuation parameter, P_p , equals 10).

The opportunity to distinguish between the two types of character evolution by setting P_{ana} and P_{clado} independently to their respective values, offers the possibility to choose for either the 'radiation' version, where P_{clado} will usually be set at zero, or the 'punctuated-equilibrium' version.

Every species may possess two basic types of characters. By default every character belongs to the kind that evolves without restriction and independently from the other characters, and independently from its own history as reflected by its present state. Character evolution is then only controlled by the character-change probability parameters. When comparing species, the occurrence of character state resemblances that are not due to inheritance (common descent), is only a matter of chance.

If desired, the user may specify a number of characters whose evolution is restricted. Changes in these selected characters will not be allowed to result in homoplasous similarities. These restrictions will result in a closer congruence of the character state trees with the phylogenetic tree. Homoplasous states that result from character reversals, however, cannot be prevented.

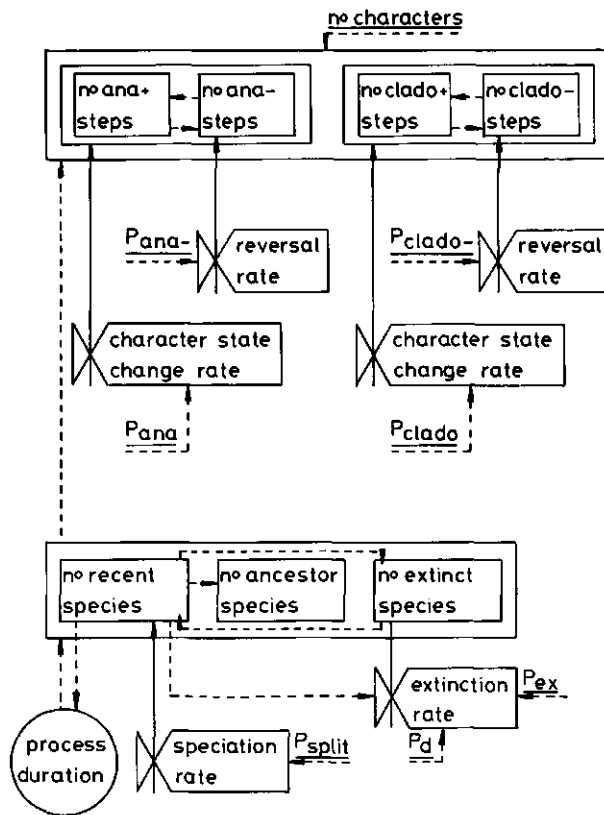


Figure 2.5 Simplified relational diagram to show the main elements of GENESIS. State variables are presented within rectangles, rates of change within valve symbols, auxiliary variables within circles, and parameters are underlined. Flow of information is designated by dotted arrows, flow of 'material', by solid arrows.

A summary of the general structure of GENESIS

Figure 2.5 summarizes the main structure and interrelations of the system that simulates the evolutionary process in a relational diagram. The model contains two main state variables; 1) the number of recent, ancestor, and extinct species and 2) the number of evolutionary steps. The total number of species is a function of the speciation rate, the extinction rate and the duration of the simulated evolutionary process. The number of evolutionary steps is determined

by the number of characters assigned to each species, the number of (recent) species, the character change rate, and the character reversal rate. The rates of change depend on one or more parameters; the rate of species extinction, however, may also be determined by the number of (recent) species. The duration of the process determines the number of (recent) species, whereas the number of (recent) species may in turn determine the duration of the process.

Input, output and options

GENESIS is an interactive program. Parameters are entered in response to messages that appear on the screen. The program is dimensioned for a maximum of 200 characters and 1200 species, including ancestral species and extinct species, but these values can, of course, be changed.

There are several ways in which the simulation process can be stopped. The duration of the process can be determined either by the number of extant species, the total number of species including ancestral and extinct ones, or the number of time intervals. Using the 'punctuated-equilibrium' version of GENESIS, one can choose for another duration-limiting parameter. The simulation can be halted when a specified number of (recent) species is reached, but only after a certain number of time intervals have elapsed.

GENESIS automatically characterizes each simulation run by a number of parameters and tree statistics. The basic parameters are the diversity parameters: the number of recent species, ancestral species and extinct species. A number of measures of evolutionary and phenetic change (Sokal, 1983), are calculated; path length ($L_{max(u)}$), Manhattan distance over all OTUs ($L_{max(l)}$), minimum length or total size ($L_{min(l)}$), and the actual length (L_{act}). These measures were used to define a number of tree statistics.

Sokal (1983:170) defined the reversal index (R_1) as:

$$R_1 = L_{max(u)} / L_{max(l)}$$

R_r measures "the amount of reversals and repeats in character state changes for the entire taxon considered as a bush, where reversals near the base will be weighted more heavily". High values of R_r indicate a high amount of reversals and repeated character state changes. If character reversals are absent, $R_r = 1$.

The dendritic index (DI) has been defined by Sokal (1983:171) as:

$$DI = (L_{max(u)} - L_{act}) / (L_{max(u)} - L_{min(l)})$$

DI measures the amount of shared evolution: if $DI = 0$, no shared evolution has occurred, and the taxon is a bush, whereas, if $DI = 1$, there is no homoplasy, in other words, there are no parallelisms nor reversals. Sokal (1983, his table 3) examined 19 data sets for real organisms, and presented minima and maxima of a number of tree statistics. The minimum and maximum value for DI in these data sets appeared to be 0.5354 and 0.9441 respectively.

Other indices of homoplasy have been suggested, e.g. the amount of excess, which equals $1 - DI$, and the homoplasy ratio (H). H equals $1/CI$, CI being the consistency index (Kluge & Farris, 1969), defined as:

$$CI = L_{min(l)} / L_{act}$$

(Sokal, 1983:172,173).

The minimum and maximum for CI as occurring in the 19 data sets were 0.1495 and 0.8621, respectively.

As a measure of symmetry of a tree, the index proposed by Colless (1982) has been used. This index (I_{col}), called *Colless2* in Sokal (1983), has been defined as the summation over all nodes (branching points) of the tree, of the difference in the number of terminal descendants on each side of the node divided by the score for complete asymmetry $((S(S - 3) + 1)/2$, S being the number of species) (Colless, 1982:103). For a perfectly symmetrical tree, $I_{col} = 0$; for a perfect asymmetrical one, $I_{col} = 1$. Minimum and maximum values in Sokal's table were 0.1061 and 0.4889, respectively.

The stemminess of a tree has been defined (Fiala & Sokal, 1985) as the summation over all nodes of the tree, excluding the basal one, of the stemminess of each node, divided by the number of nodes. This node stemminess can be calculated by dividing the length of its stem segment, by the total length of all internodes along the paths leading from the node to all the descendants, including the length of the stem segment itself. Stem length is measured in time units. Stemminess has been designed as a measure of the

topology of a tree. I have used a slightly modified stemminess index (I_{stem}) by using the number of evolutionary steps that have occurred along the segment as a measure of segment length, instead of time. If evolution proceeds at the same rate in both daughter species, the indices will yield the same results. If these rates are unequal, one will expect an effect on I_{stem} . Fiala & Sokal (1985) selected simulated trees based on their stemminess; 0.50 appeared to be a high value and 0.22 a low one. Their trees contained 20 OTUs.

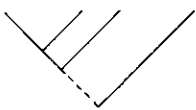
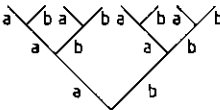
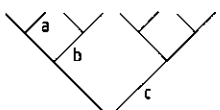




Tree topology	N OTUs	Stem length	Stemminess
	8 50	1 1	0.159 0.040
	8 8 8 8	a=1 b=1 a=1 b=2 a=1 b=10 a=1 b=20	0.269 0.263 0.230 0.220
	8 8 8	a=10 other 1 b=10 other 1 c=10 other 1	0.214 0.339 0.350
	9	1	0.240
	9	1	0.247
	9	1	0.255
	9	1	0.255

Figure 2.6 Examples of trees with different topologies, number of taxa and stem lengths, and values of corresponding stemminess indices (I_{stem}).

It remains rather obscure what this stemminess index expresses. The stemminess seems to describe several aspects of a tree topology at the same time, for instance the symmetry and inequality of stem lengths. Even the number of OTUs in the tree plays an important role in determining its value. The stemminess for a perfectly asymmetrical tree for 50 OTUs, with all segments of equal length, is 0.0403, the stemminess for a perfectly symmetrical one is 0.2082. Figure 2.6 shows the effects of several properties of a tree topology on stemminess.

As a measure of the adequacy of the characters to resolve the cladogram, Sokal (1983:175,176) proposed and used the ratio N_{char}/S , where N_{char} denotes the number of characters and S the number of species. Also the ratio $N_{char}/(2S - 3)$ has been proposed as such a measure (Sokal & Shao, 1985). Both these measures were calculated in a slightly different form, not using the number of characters, but the number of pseudocharacters (N_{pseu}), i.e. the number of characters after recoding to binary characters. Because all simulations are run with 100 characters and all are programmed to produce 50 recent species at the end, the adequacy as measured by both indices as originally defined will be 2.00 and 1.03, respectively. The minimum and maximum value for N_{char}/S , as presented by Sokal for the 19 data sets, were 1.16 and 6.62, respectively. This latter value, however, was derived from a data set on Pygopodidae (Kluge, 1976) with binary coded characters.

The relative rate of increase of recent species (RRI) can be defined as the summation of $N_{rec(t)} - N_{rec(t-1)}/N_{rec(t)}$ over all time intervals, t denoting an arbitrary time interval. The innate capacity for increase (IC) can be computed as $\ln(N_{rec})/T$, were T denotes the duration of the (simulated) evolutionary process. These statistics, of course, are only meaningful in describing patterns of exponential increase.

Each simulation can also be graphically characterized. GENESIS may produce survival curves for recent species, duration curves for recent, ancestral and extinct species, and graphs of the number of recent species against time.

Table 2.1 Input parameter values for the four standard simulations. See text for definition of parameter symbols. 'Special characters' are characters in which changes are not allowed to result in homoplasous similarities.

Model version	Radiation		Equilibrium	
	Gradual	Punctuated	Gradual	Punctuated
	1	2	3	4
Input parameters:				
N_{char}	100	100	100	100
P_{split}	0.10	0.10	0.10	0.10
P_{ex}	0	0	— ¹	—
P_d	—	—	1	1
P_{ane+}	0.01	0	0.002	0
P_{ane-}	0	0	0	0
P_{clado+}	0	0.006	0	0.008
P_{clado-}	0	0	0	0
P_p	—	2	—	2
% 'special characters'	0	0	0	0
limit: N_{rec}	50	50	50	50
t_{min}	—	—	100	100

¹ "—" means not applicable.

All output from GENESIS is optional and is transferred to files with names supplied interactively. Output may include files containing only recent species and their character state vectors, ready to be used by different program packages such as CLUSTAN (Wishart, 1986), PHYLIP (Felsenstein, 1987), Wagner78 (Farris, 1978), PAUP (Swofford, 1985) and CAFCA (Zandee, 1985; Zandee & Geesink, 1987).

The main evolutionary scenarios

Several standard simulations were run, each with a well defined set of parameters. For the moment only the values will be presented that define these simulations, and the results of the standard simulations themselves. A sensitivity analysis has been conducted by varying one or a few parameters from the standard set, and examining the effects on the output as summarized by the tree statistics. The results of this sensitivity analysis will be presented in a separate paper.

Four standard simulations must be considered, each representing one of the main versions of GENESIS. For the 'equilibrium' version, only the option was considered, where the feedback mechanism becomes operative after N_{rec} has first become greater than N_{equi} . Table 2.1 lists the parameters of each of these versions.

Table 2.2 presents the mean values of the parameters and tree statistics for the four standard simulations or evolutionary scenarios. Standard deviations, coefficients of variation, minima and maxima were calculated too, but are not presented here. Parameters and tree statistics have been calculated for the entire phylogenetic tree, including its extinct species, and for the phylogenetic tree containing only the recent species, separately. The various statistics presented by Sokal (1983) have been calculated for the set of recent species only.

For Scenario 1, a character change probability of 0.01 was used, resulting in a mean number of about 560 evolutionary steps per simulation (table 2.2). This probability parameter as used in Scenarios 2, 3 and 4, was set at values such that the mean number of steps would be about the same as for Scenario 1. All simulations were run with 100 characters for each species, and a speciation probability of 0.1. In the 'punctuated' versions, character evolution in one daughter species was made to proceed at a rate twice that of the other daughter species ($P_p = 2$). The damping parameter was set at 1.0 in both 'equilibrium' versions, and the number of species in the equilibrium (N_{equi}) was set at 50. Homoplasous steps were allowed to occur freely in all characters. In the 'radiation' versions, the simulation process was made to continue until exactly 50 recent species were produced. In the 'equilibrium' versions at least 100 time steps must have passed, until the process is allowed to halt at exactly the moment that 50 recent species are in existence.

The results of the simulations (table 2.2) will be used as a reference for sensitivity analysis and further experiments with GENESIS. Let it here suffice that D/I falls within the ranges as given by Sokal (1983) for 19 data sets. The index C/I of the 'equilibrium'

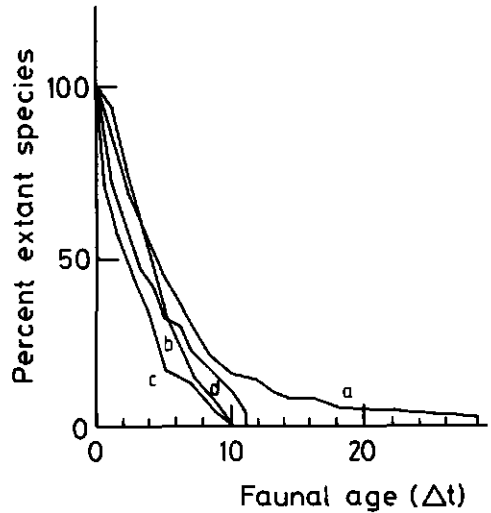
GENESIS, a simulation model of evolution

Table 2.2 Mean values of describing parameters and tree statistics for the four standard simulations. See text for definition of symbols.

Model version	Radiation		Equilibrium	
Standard simulation	Gradual	Punctuated	Gradual	Punctuated
	1	2	3	4
Diversity:				
N_{rec}	50.0	50.0	50.0	50.0
N_{ex}	0.0	0.0	287.9	283.4
N_{anc}	49.0	49.0	336.9	332.4
N_{tot}	99.0	99.0	674.8	665.9
Number of steps:				
<i>Clado</i> +	0.0	561.6	0.0	551.2
<i>Clado</i> -	0.0	0.0	0.0	0.0
<i>Ana</i> +	559.3	0.0	567.8	0.0
<i>Ana</i> -	0.0	0.0	0.0	0.0
Tree statistics:				
DI	0.81	0.78	0.90	0.89
DI^t	0.81	0.78	0.90	0.89
CI	0.33	0.31	0.57	0.57
CI^t	0.33	0.31	0.25	0.25
R_1	1.00	1.00	1.00	1.00
R_1^t	1.00	1.00	1.00	1.00
I_{col}	0.10	0.11	0.12	0.10
I_{col}^t	0.10	0.11	0.05	0.05
I_{stem}	0.24	0.19	0.33	0.28
N_{pseu}/S	3.61	3.49	2.34	2.32
$N_{pseu}/(2S-3)$	1.86	1.80	1.22	1.20
RRI	0.10	0.11	0.04	0.04
ICI	0.09	0.10	0.03	0.03
Duration of:				
recent species	5.27	5.43	4.91	4.54
extinct species	0.00	0.00	5.01	5.22
ancestor species	5.75	5.50	5.47	5.65
process	45.45	42.40	116.16	122.08
Number of runs	100	25	25	25

t = index calculated for the entire tree, that is the tree of recent species and extinct species included. Other indices are calculated from the tree of recent species only.

Figure 2.7 Examples of typical survival curves as produced by standard simulations 1 - 4 of GENESIS (a, b, c and d, respectively). The survival curves do not differ significantly between standard simulations, and they resemble Lyellian curves from palaeontological literature.



versions had higher values than the maximum given by Sokal, indicating a relative low degree of homoplasy. For all four standard simulations, the symmetry of the generated trees, as expressed by I_{col} seemed high compared with Sokal's data. Because evolutionary rate is equal in both daughter lineages and proportional to time, in the 'radiation' versions, the I_{stem} may be compared with the stemminess index as used by Fiala & Sokal (1985). The values fall in the lower region of the range found for their simulated trees. The mean value of the I_{stem} for Scenario 1 approaches the value for a perfectly symmetrical tree with all segments equal in length.

Figure 2.7 gives typical survival curves from the four standard simulations.

Discussion

How far do the processes simulated by GENESIS indeed mimic evolutionary processes we believe to have occurred in nature? How far do the patterns produced and represented in the data sets of the recent species generated by GENESIS, resemble real data sets? Can the evolutionary process indeed be thought of as a stochastic process in which chance plays the major role? Or must we agree with Dawkins (1986) in his conclusion that "... living things are too

Dawkins (1986) in his conclusion that "... living things are too improbable and too beautifully 'designed' to have come to existence by chance"?

The question whether evolution is by mere chance or not is simplistic. The role played by chance in the evolutionary process is elegantly treated by Dawkins (1986) in his *Blind Watchmaker*. Using the example of the randomly typing monkey, Dawkins clearly demonstrated that, although mutation is random, the "most important ingredient (in the Darwinian recipe) is cumulative selection which is quintessentially non-random." Selection is not explicitly incorporated in GENESIS. Extinction is simulated as a random process, and some character states do not occur just by chance. However this does not imply that the results are irrelevant. As Raup et al. (1973) say: "We do not suggest that evolution be viewed as a haphazard process, independent of basic relations of cause and effect. Rather we suggest that an evolutionary event may depend upon the joint occurrence of many underlying causes, each having a specific probability of occurrence at a given time, so that the event itself can be predicted in a statistical sense, even though it does in fact have a conventional cause".

Sokal (1983) examined the degree to which the Caminalcules data set resembles data sets on real organisms. He concluded from a few criteria, that the Caminalcules are good imitations of real organisms. The same holds for the data sets produced in the four scenarios of GENESIS.

There is, however, no way of verifying that the processes and patterns, as simulated and produced by GENESIS or any other model of phylogenesis, are realistic in the sense that they have ever actually occurred in nature. In this respect simulation models of evolution are speculative models and always will be; they cannot be verified.

Viewed in the light of our objective of evaluating the merits of the various methods of reconstructing phylogeny, GENESIS need not be a verifiable simulation model at all. However, the assumptions underlying GENESIS need to be in agreement with the assumptions

of the reconstruction methods tested. By selecting the proper input parameters, this can be easily achieved.

GENESIS actually summarizes the processes of lineage branching and character evolution believed to occur in nature and produces patterns of diversity and of character state distributions, which are indeed similar to the patterns encountered in the real world. GENESIS can be used to discover features of the evolutionary process that are important because they determine the accuracy of methods of phylogeny reconstruction. Data sets produced by GENESIS are suitable as test cases to study the qualities of these methods.

A good understanding of the qualities of the various methods used for reconstructing phylogenies is a prerequisite for a critical and judicious use and GENESIS may be a useful tool in evaluating these qualities.

Zusammenfassung

GENESIS: Ein Simulationsmodell der Phylogenie

Die Ausgangssituation und die frühe Evolution der Zustandsvektoren der Merkmale

Ein Simulationsmodell für phylogenetische Prozesse wird vorgestellt. Das Modell heisst GENESIS und erzeugt eine Phantomwelt an artifiziellen Arten, welche in der Form von Zustandsvektoren von Merkmalen auftreten. Diese Arten können von GENESIS unter Berücksichtigung verschiedener Optionen erzeugt werden, die den verschiedenen Evolutionsabläufen entsprechen; oder umgekehrt, GENESIS kann dazu benützt werden, Gruppen von Daten zu erzeugen, die unterschiedliche Eigenschaften erkennen lassen. Die Eigenschaften der simulierten Evolutionsprozesse und deren entsprechende Daten-Matrizen werden durch eine Zahl von verschiedenen verzweigungs-Statistiken beschrieben.

Die Daten-Matrizen können im folgenden verwendet werden, die Güte der verschiedenen Methoden der Stammbaumkonstruktion zu testen. Über solche Tests und über Ergebnisse, die die Empfindlichkeit der Analysen prüfen, wird in späteren Publikationen berichtet werden.

Acknowledgements

I acknowledge the invaluable help of the late Prof. Dr Ir R. Cobben. I thank Drs T. De Winter, P. De Vrijer, R. Zandee and C. Booij, for comments on earlier versions of the manuscript and I want to thank D. Jansen and staff of the Computer Centre of the Agricultural University, Wageningen, for help during the creation of GENESIS. Mr J. C. Rigg kindly corrected the English text.

References

- Anderson, S. & C. S. Anderson, 1975. Three monte carlo models of faunal evolution. — *American Museum Novitatus*, No. 2563.
- Baum, B. R., 1984. Application of compatibility and parsimony methods at the infraspecific, specific, and generic levels in Poaceae. In: *Cladistics: Perspectives on the reconstruction of evolutionary history* (Th. Duncan & T. F. Stuessy, eds): 192-220. New York, Columbia University Press.
- Colless, D. H., 1982. (Review of) Phylogenetics: The theory and practice of phylogenetic systematics. — *Systematic Zoology* 31: 100-104.
- Colwell, R. K. & D. W. Winkler, 1984. A null model for null models in biogeography. In: *Ecological communities: Conceptual issues and the evidence* (D. R. Strong, D. Simberloff, L. G. Abele & A. B. Thistle, eds). Princeton, Princeton University Press.
- Dawkins, R., 1986. *The Blind Watchmaker*. Harlow, Longman.
- Duncan, T., R. B. Phillips, W. H. Wagner Jr, 1980. A comparison of branching diagrams derived by various phenetic and cladistic methods. — *Systematic Botany* 5: 264-293.
- Eldredge, N. & S. J. Gould, 1972. Punctuated equilibria: an alternative to phyletic gradualism. In: *Models in paleobiology* (T. J. M. Schopf, ed.): 82-115. San Fransisco, Freeman, Cooper.
- Farris, J. S., 1978. *WAGNER78*; manual, documentation and a FORTRAN IV Wagner program.
- Felsenstein, J., 1987. *PHYLIP; Phylogeny Inference Package, version 3.0*; manual, documentation and several PASCAL programs. University of California Herbarium, Berkely, California.
- Fiala, K. L., 1983. A simulation model for comparing numerical taxonomic methods. In: *Numerical Taxonomy* (J. Felsenstein, ed.): 87-91. Springer-Verlag, Berlin.

- Fiala, K. L. & R. R. Sokal, 1985. Factors determining the accuracy of cladogram estimation: evaluation using computer simulation. — *Evolution* 39: 609-622.
- Gingerich, P. D., 1984. Punctuated Equilibria - Where is the evidence. — *Systematic Zoology* 33: 335-338.
- Gingerich, P. D., 1977. Patterns of evolution in the mammalian fossil record. In: *Patterns of evolution, as illustrated by the fossil record. Developments in palaeontology and stratigraphy, 5* (A. Hallam, ed.): 469-500. Amsterdam, Oxford, New York, Elsevier.
- Gould, S. J. & N. Eldredge, 1983. Darwin's gradualism. — *Systematic Zoology* 32: 444-445.
- Kluge, A. G., J.S. Farris, 1969. Quantitative phyletics and the evolution of anurans. — *Systematic Zoology* 18: 1-32.
- Leman, C. A., P. W. Freeman, 1984. The genus: a macroevolutionary problem. — *Evolution* 38: 1219-1237.
- Raup, D. M., 1977. Stochastic models in evolutionary palaeontology. In: *Patterns of evolution, as illustrated in the fossil record. Developments in palaeontology and stratigraphy, 5* (A. Hallam, ed.): 59-78. Amsterdam, Oxford, New York, Elsevier.
- Raup, D. M., 1985. Mathematical models of cladogenesis. — *Paleobiology* 11: 42-52.
- Raup, D. M. & S. J. Gould, 1974. Stochastic simulation and evolution of morphology - towards a nomothetic paleontology. — *Systematic Zoology* 23: 305-322.
- Raup, D. M., S. J. Gould, T. J. M. Schopf & D. S. Simberloff, 1973. Stochastic models of phylogeny and the evolution of diversity. — *Journal of Geology* 81: 525-542.
- Scudo, F. M., 1985. Darwin, Darwinian theories and Punctuated Equilibria. — *Systematic Zoology* 34: 239-242.
- Simpson, G. G., 1944. *Tempo and mode of evolution*. New York, Columbia University Press.
- Sokal, R. R., 1983. A phylogenetic analysis of the Caminalcules. I. The data base. — *Systematic Zoology* 32: 159-184.
- Sokal, R. R., K. Shao, 1985. Character stability in 39 data sets. — *Systematic Zoology* 34: 83-89.
- Stanley, S. M., 1977. Trends, rates, and patterns of evolution in the bivalvia. In: *Patterns of evolution, as illustrated by the fossil record. Developments in palaeontology and stratigraphy, 5* (A. Hallam, ed.): 209-250. Amsterdam, Oxford, New York, Elsevier.
- Stanley, S. M., 1985. Rates of evolution. — *Paleobiology* 11: 12-13.

- Swofford, D. L., 1985.** *PAUP, Phylogenetic Analysis Using Parsimony, version 2.4*; manual, documentation and program. Illinois Natural History Survey, Champaign, Illinois.
- Tateno, Y., M. Nei & F. Tajima, 1982.** Accuracy of estimated phylogenetic trees from molecular data. I. Distantly related species. — *Journal of Molecular Evolution* 18: 387-404.
- Wishart, D., 1986.** *CLUSTAN: A cluster analysis package, Version 3.2*. Edinburgh University, Program Library Unit.
- Zandee, M., 1985.** *C.A.F.C.A.: A collection of APL functions for cladistic analysis*. On-line CAFCA documentation.
- Zandee, M. & R. Geesink, 1987.** Phylogenetics and legumes: a desire for the impossible? In: *Advances in Legume Systematics, Part 3* (C. H. Stirton, ed.): 131-167. Royal Botanic Gardens, Kew.

3

GENESIS: a simulation model of phylogeny

A sensitivity analysis ¹

Abstract

In a former paper (Heijerman, 1988) a simulation model of phylogeny, GENESIS, was presented. This paper describes the results of a sensitivity analysis of GENESIS. The analysis is performed by changing the input parameter values and estimating the relative effects on the model's output, as summarized by several tree statistics. The results show that none of the statistics tested can be classified as an unambiguous estimator of accuracy of methods for estimating phylogenetic trees. The sensitivity analysis increases the insight into the behaviour and applicability of the model. This is a prerequisite for a correct interpretation of the results of the evaluation experiments that will be carried out using GENESIS.

Key words: Numerical taxonomy — Phylogeny — Simulation model — Sensitivity analysis — Tree statistics

Introduction

GENESIS, a simulation model of phylogeny, was constructed to provide a means for evaluating the accuracy of various phylogeny reconstruction methods (Heijerman, 1988). It was argued, that it is neither possible nor necessary to validate simulation models of phylogeny.

¹ Published as: Heijerman, Th., 1990. GENESIS: a simulation model of phylogeny. Part 2. A sensitivity analysis. — *Zeitschrift für zoologische Systematik und Evolutionsforschung* 28: 81-93.

In for example agronomy, simulation models are developed that reflect as accurately as possible the real agrosystem. These models are used to describe and predict the growth and yield of a crop, and the implications of the model can be tested against field data.

Although GENESIS does not actually model a unique system or process, it should be classified as a speculative model, because there is no way to test its reality and generality. Speculative models, however, can and must be put to the test of usefulness and applicability.

GENESIS only makes some very basic assumptions about the evolutionary process itself; ancestor species give rise to daughter species by a splitting process and characters change in the course of evolution. Species contain information about and may thus reflect this course of evolution in their character states. The processes of branch splitting and character evolution may be simulated in different ways, that accord with different evolutionary scenarios.

These basic assumptions are very simple and non-controversial. Moreover, they agree with the assumptions implied by methods for estimating phylogenetic trees.

The output produced by GENESIS, in the form of a set of recent species and their character state vectors, should also correspond to something that is real. This 'reality' cannot be tested by comparing the generated data sets with data sets on real species and characters. It is important that the output data set contains sufficient information about the process that produced a particular data set. If evolution is simulated under different conditions, that is, in different evolutionary scenarios, then the resulting data sets should also be different. The question is whether or not GENESIS can put this relevant information in its output data sets.

In order to test GENESIS' usefulness and applicability, and to come to a better understanding of its behaviour and performance in different evolutionary contexts, I carried out a sensitivity analysis, the results of which are presented in this paper.

Methods

In the preceding paper (Heijerman, 1988), four evolutionary scenarios were defined and the results of the corresponding standard simulations were presented. In this paper, the responses of GENESIS to input parameter changes are investigated. It is impossible to study the entire parameter space of GENESIS, because there are too many input parameters that can be changed, and thus some were kept constant over all simulations.

Unless otherwise stated, all simulations were carried out using 100 characters per species and the speciation probability (P_{split}) was set at 0.10. In the radiation version of GENESIS the simulations were stopped when the number of recent species reached 50. In the equilibrium version simulations were also stopped at 50 species, but only after at least 100 time steps had elapsed. Because evolution is simulated as a stochastic process, simulations were performed in sets of at least 25 runs.

Four scenarios

A more detailed description of the four evolutionary scenarios is given in Heijerman (1988).

In evolutionary Scenario 1, branch splitting occurs according to the 'radiation' version of GENESIS and character changes occur according to the 'gradual' version. The probability of character state change is 0.01, as in the standard version of this scenario. In the sensitivity analysis, the values of three input parameters were subjected to changes, namely the probability of species extinction (P_{ex}), the probability of reversed character state changes (P_{ana}) and the number fraction of compatible characters ($F_{compchar}$).

In evolutionary Scenario 2, branch splitting also occurs according to the 'radiation' version. Character changes occur according to the 'punctuated equilibrium' version, which introduces unequal evolutionary rates. This is accomplished by the 'punctuation' parameter (P_p), which determines how much faster evolution proceeds, expressed as the number of character state changes per

time-step, in one daughter lineage compared to the other. The speciation probability is set at a value such that, whatever the value of P_p , the total number of character changes will remain approximately the same. In this scenario four parameters were subjected to changes, namely P_{ex} , the probability of reversed character state changes (P_{clado}), $F_{compchar}$ and P_p .

In Scenario 3, species evolution is simulated according to the 'equilibrium' version and character evolution according to the 'gradual' version. Three input parameters are subjected to changes, P_{clado} , the damping parameter (P_d) and $F_{compchar}$.

Finally, in Scenario 4 the branch-splitting process is according to the 'equilibrium' version, and character change according to the 'punctuated' version. Apart from the three parameters mentioned for Scenario 3, also P_p is subjected to change.

Descriptive statistics

The output data sets as produced by GENESIS, were characterized by a number of tree statistics. These statistics were calculated for the tree of recent species, and for the entire tree, containing recent as well as extinct and ancestor species.

Different types of descriptive statistics were calculated:

Tree Length Measures: The consistency index (CI), the dendritic index (DI) and the reversal index (R_1) (Sokal, 1983).

Measures of Tree Topology: The Colless index of symmetry (I_{col}) (Colless, 1982) and the stemminess index (I_{stem}). I_{stem} was calculated in a slightly different form compared to the stemminess index as originally defined by Fiala & Sokal (1985); instead of time, I used the number of evolutionary changes as a measure of branch-length. However, I_{stem} suffers from a minor disadvantage, namely, if all segments leading to a node in the tree are of length zero, then calculation of the stemminess of this particular node is impossible. Therefore, in the analysis presented here, I have only calculated I_{stem} for trees in which no such nodes occur. Consequently, unlike the other indices, the average values for I_{stem} are not always based on the same number of runs. In part of the analysis I have also used

the stemminess index as defined by Fiala & Sokal (1985), hereafter referred to as I_{stemFS} .

Measures of Diversity: The user must specify the number of recent species (N_{rec}) that must be present at the end of the simulation, but the number of ancestor species (N_{anc}) and/or extinct species (N_{ex}) that will be produced depends on the extinction probability (P_{ex}) (in Scenarios 1 and 2), or the value of the damping parameter (P_d) (in Scenarios 2 and 4).

Measures of Adequacy: In cases where $P_{ex} = 0$ (Scenarios 1 and 2) or $P_d = 1$ (Scenarios 3 and 4), the value of P_{split} is chosen so that the total number of evolutionary steps will be about 560 (table 2 in Heijerman, 1988). N_{char} is set at 100, but the number of pseudocharacters (N_{pseu}) is determined by the total number of evolutionary changes. Two measures of adequacy were used, N_{pseu} / S and $N_{pseu} / (2S - 3)$ ($S =$ number of recent species). Both are defined in terms of the number of pseudocharacters rather than in terms of the number of characters as in Fiala & Sokal (1985).

To characterize the output data sets, R_1 , CI , DI , I_{col} , I_{stem} and I_{stemFS} as calculated for the tree of recent species, will be used.

General design of the analysis

The responses of indices to input parameter changes is described in all four scenarios. The simulations were run to produce 50 recent species each with 100 characters. The results are presented in the form of 3D-graphs. The responses are quantified using a simple method that is outlined below, where its application can more easily be seen.

To investigate the accuracy and applicability of this method, the analysis was repeated in much greater detail for Scenario 1. The results of this detailed analysis are also presented graphically, and a regression analysis was carried out to quantify the responses of input parameter changes.

Finally, to investigate whether the results depend on the number of species and characters used, I also performed a sensitivity analysis using 20 species each with 50 characters.

Results

Responses of indices to parameter changes in all scenarios

The parameter space investigated is determined by the combinations of a limited number of values for P_{ex} (0, 0.03, 0.06), P_{ana-} (0, 0.30, 0.50) and $F_{compchar}$ (0, 0.50, 1.00) (Scenarios 1 and 3) or for P_{clado-} (0, 0.30, 0.50), $F_{compchar}$ (0, 0.50, 1.00), P_d (1, 10) and P_p (1, 20) (Scenarios 2 and 4).

Figure 3.1 illustrates the responses of the measures of tree length to changes in input parameters for all four scenarios, and in figure 3.2 the responses of the tree topology measures to input changes are presented.

It appears from figure 3.1 that the variation in the reversal index, R_1 , may be explained by changes in the value of the reversal probability. In Scenarios 3 and 4 the effect of respectively P_{ana-} and P_{clado-} is less pronounced than in Scenarios 1 and 2. In Scenarios 1 and 2, R_1 is also slightly affected by P_{ex} .

The dendritic index (DI) measures the amount of parallellisms and reversals. If there are no parallellisms nor reversals, DI will equal 1. If the taxon is a 'bush', DI equals 0, which means that character states are fully incompatible on the cladogram. Since DI is a measure of the compatibility of the characters on the tree, it is expected that $F_{compchar}$ as well as the reversal probability, will determine its value. Indeed, DI is affected by $F_{compchar}$ in all scenarios, although the effect of P_{ana-} or P_{clado-} appears to be rather weak.

The consistency index (CI) is another measure of the amount of convergence. Where there is no convergence on the tree, CI will equal 1 and where there is a maximum amount of convergence, CI will equal 0. It appears that CI is rather strongly affected by changes in $F_{compchar}$ in all four scenarios, and also a marked effect of P_{ana-} or P_{clado-} can be observed.

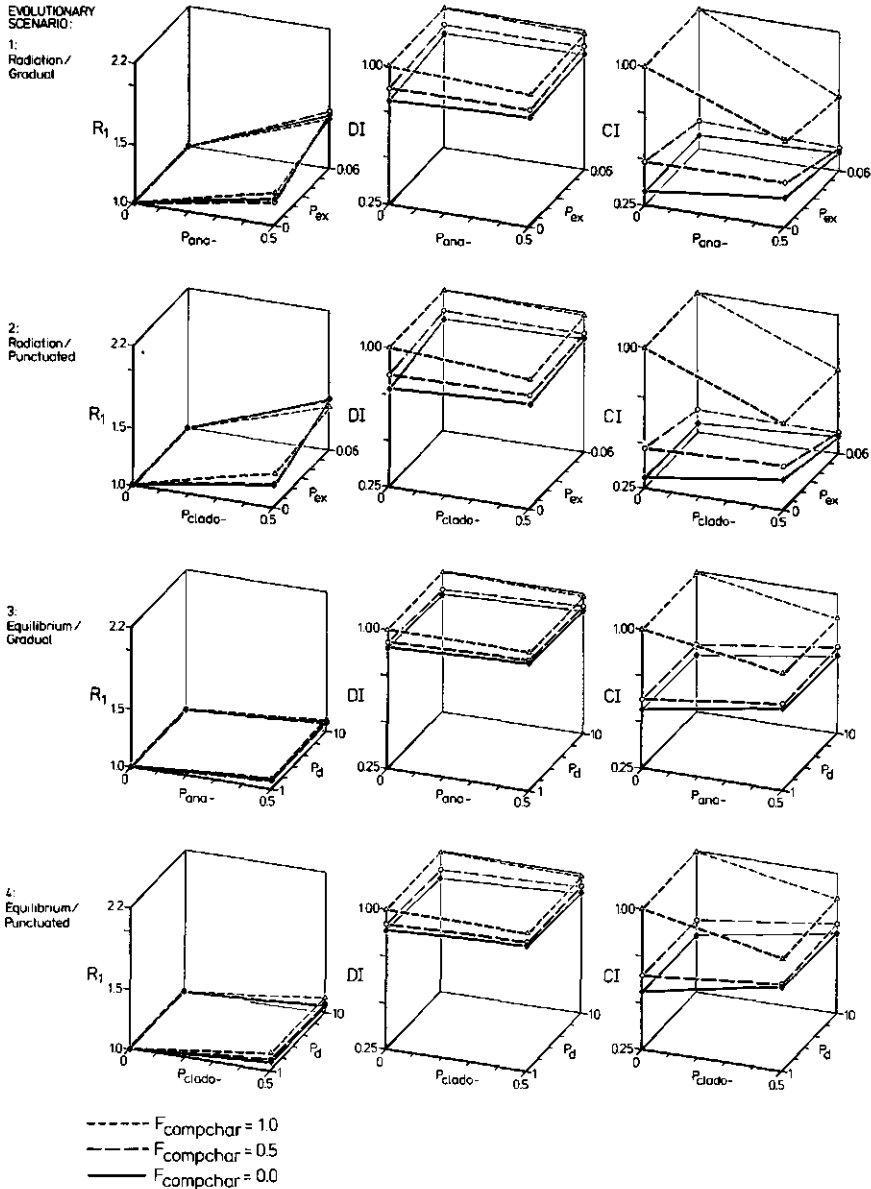


Figure 3.1 Responses of changes in input parameters on three tree length measures, the reversal index (R_1), the dendritic index (DI) and the consistency index (CI), in the four evolutionary scenarios. Input parameters: retrogression parameter (P_{ana-} , P_{clado-}), extinction probability (P_{ex}), damping parameter (P_d) and/or number fraction of compatible characters ($F_{compchar}$). Value of the punctuation parameter (P_p) in Scenarios 2 and 4: 20.

Chapter 3

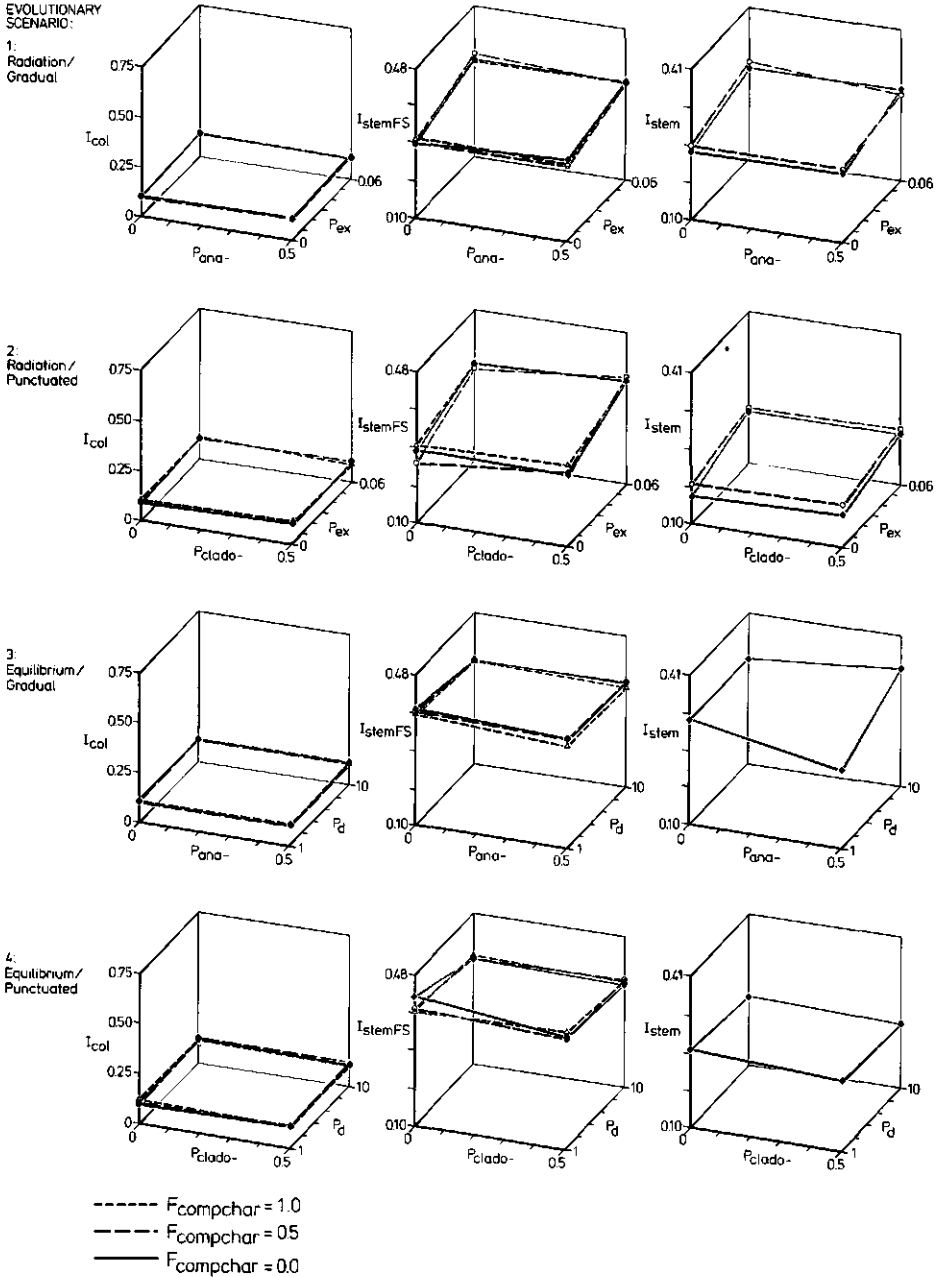


Figure 3.2 Responses of changes in input parameters on three tree topology measures: the Colless index of symmetry (I_{col}) and two stemminess indices (I_{stemFS} and I_{stem}), in the four evolutionary scenarios. Input parameters as in figure 3.1.

Figure 3.2 shows that the Colless index of symmetry, I_{col} , which theoretically ranges from 0 (perfect symmetry) to 1 ('Hennigian-comb'), is not affected in any of the scenarios. The stemminess indices I_{stem} and I_{stemFS} behave similarly being only slightly affected by the input parameters. It seems that the effect of P_{ex} is most noticeable, although still rather weak.

Quantification of the responses in all scenarios

Although the responses of changes in input parameters on tree statistics can be evaluated graphically, a simplified method may be used to quantify these responses. The method developed also allows for a comparison of results within and between evolutionary scenarios.

In this method, the change in the value of a tree statistic, say R_1 , is calculated when one input parameter, say P_{ana-} , changes maximally from 0 to 0.50, while at the same time the values of the other input parameters (P_{ex} and $F_{compchar}$) are kept constant. This change in R_1 is calculated for the combinations of the extreme values of the other parameters; in this example for the next four situations: ($P_{ex} = 0, F_{compchar} = 0$), ($P_{ex} = 0, F_{compchar} = 1.00$), ($P_{ex} = 0.06, F_{compchar} = 0$) and ($P_{ex} = 0.06, F_{compchar} = 1.00$). The impact of P_{ana-} on R_1 can then be calculated by averaging the four values thus obtained ('impact factor'). The importance of the other two input parameters were calculated in the same way. For Scenario 1 the following values of the 'impact factors' of P_{ex} , P_{ana-} and $F_{compchar}$ on R_1 were found: 0.099, 0.350 and 0.017 respectively, indicating that R_1 is mainly affected by P_{ana-} and, to a lesser extent also by P_{ex} . The same procedure was repeated for the other tree statistics.

To make comparisons between scenarios easier, the value of the 'impact factor' is expressed as a proportion of the maximum possible impact. When there are no reversals, R_1 has its minimum value of 1. Theoretically R_1 has no maximum value, but I used 2.58 as the maximum value since it was the highest that was recorded among more than 15.000 simulations. Thus the range of R_1 equals 1.58 (2.58 - 1). Theoretically CI ranges from 0 (maximum amount of convergence) to 1 (no convergence) and the same holds for DI .

Table 3.1 Results of the sensitivity analysis using the RIF-method in the 50-species simulations. Numbers are RIF-values (see text for explanation). Results are based on simulation sets of 50 runs each.

Evolutionary scenario	Independent variables	Dependent variables					
		R_1	DI	CI	I_{col}	I_{stem}	I_{stemFS}
Radiation/ Gradual	P_{ana-}	22	2	19	0	2	2
	$F_{compchar}$	1	14	49	0	3	2
	P_{ex}	6	0	3	1	11	11
Radiation/ Punctuated	P_{clado-}	21	3	18	1	1	2
	$F_{compchar}$	3	16	51	1	4	2
	P_{ex}	6	4	3	1	8	14
	P_p	0	0	1	1	7	2
Equilibrium/ Gradual	P_{ana-}	6	2	12	0	9	3
	$F_{compchar}$	0	9	32	0	-	2
	P_d	0	1	1	0	10	3
Equilibrium/ Punctuated	P_{clado-}	6	3	14	1	4	2
	$F_{compchar}$	2	10	31	1	5	2
	P_d	1	1	1	0	2	4
	P_p	0	0	1	0	6	3
Range		1.58	1	1	1	0.41	0.50

Also I_{col} varies from 0 (perfect symmetry) to 1 ('Hennigian comb'). The minimum value for I_{stem} will approach 0 for a perfectly asymmetrical tree with very unequal stemlengths. I_{stem} calculated for a perfectly symmetrical tree of 50 species and with all segments equal in length is approximately 0.22. However, 0.413 was the highest value observed among the simulations and therefore 0.41 is used as the range of I_{stem} . Also the minimum value of I_{stemFS} approaches zero. The highest value of this index found among all simulations is 0.501, and the highest value as found by Fiala & Sokal (1985) 0.50. Thus the range for I_{stemFS} is 0.50.

Table 3.1 gives the corrected, relative 'impact factors' (RIFs) for all scenarios, and we can see that the results are in good agreement with figures 3.1 and 3.2.

Responses of indices to input parameter changes in Scenario 1

In figure 3.3 the input parameter space used for the detailed sensitivity analysis in Scenario 1 is given. In this figure, the parameter space is delimited by the extreme values of the three parameters. $F_{compchar}$ ranges from 0 to 1.0 and these are the extreme values. Also P_{ana-} has a natural upper and lower limit of 0.50 and 0 respectively. The lower extreme value for P_{ex} is 0, and the upper extreme value would be 0.10 in the present situation, where P_{split} also equals 0.10. But in this case (where $P_{ex} = P_{split} = 0.10$) the simulation could never result in 50 recent species. When P_{ex} is set at 0.09, 0.08 or even 0.07, there is very little chance the simulation will be able to proceed until it reaches 50 recent species. For example, when P_{ex} equals 0.09, only 8 out of 300 runs were able to produce 50 recent species, in the other runs all lineages came to an end too soon. In the case of P_{ex} being equal to 0.08, 50 out of 357 runs were completed successfully, for $P_{ex} = 0.07$ the result is 50 runs out of 194, and for $P_{ex} = 0.06$ the result is 50 out of 147. Therefore 0.06 was chosen as a practical and acceptable maximum value for P_{ex} .

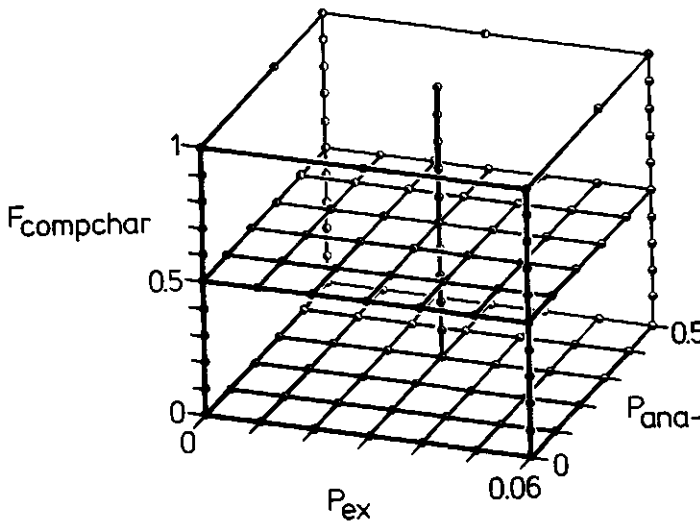


Figure 3.3 Partial parameter space (Scenario 1) showing points used for regression analysis. Each point represents 25 simulation runs.

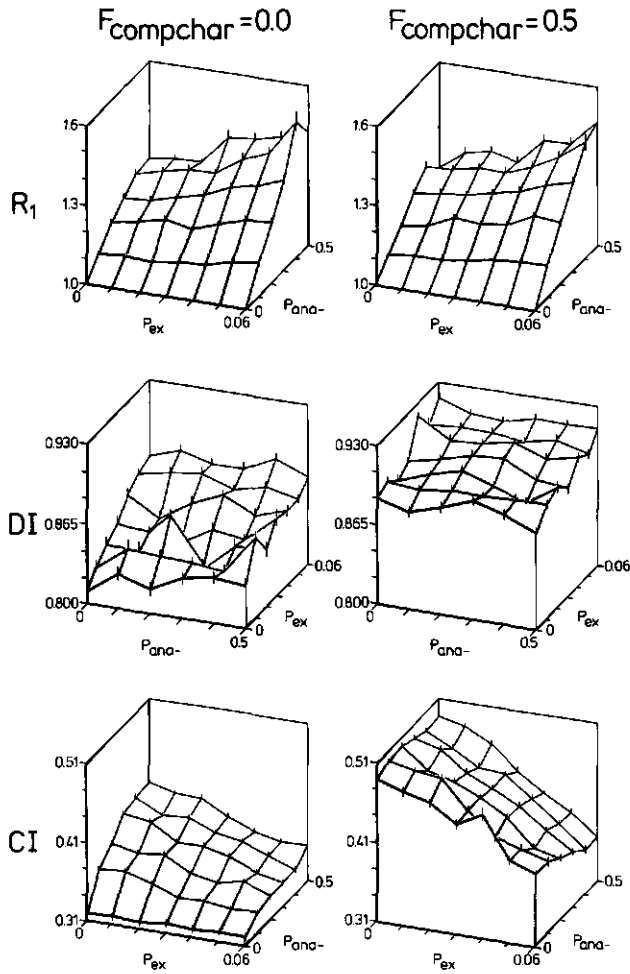


Figure 3.4 Responses of changes in input parameters on the reversal index (R_1), the dendritic index (DI) and the consistency index (CI), in Scenario 1. Input parameters: retrogression parameter (P_{ana-}), extinction probability (P_{ex}) and the number fraction of compatible characters ($F_{compchar}$). Vertical dashes denote the standard deviations of the mean.

The 133 points in this space indicate the combinations of input parameter values that were used in the simulation runs. As evolution is simulated as a stochastic process, the response at each point was estimated by 25 runs. Thus the analysis for Scenario 1 is based on $133 \times 25 = 3325$ runs.

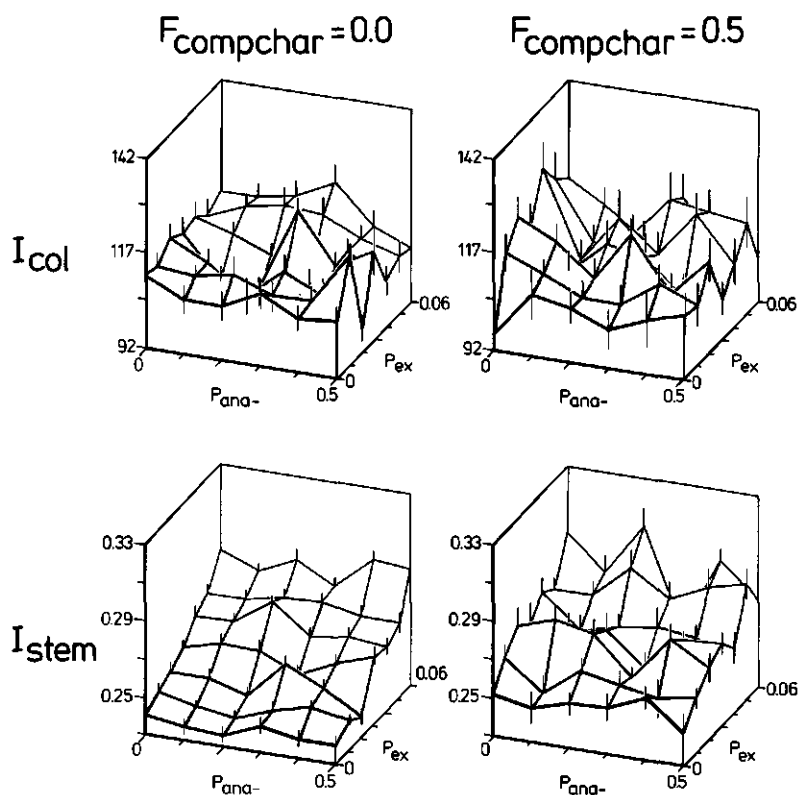


Figure 3.5 Responses of changes in input parameters on the Colless index of symmetry (I_{col}) and the stemminess index (I_{stem}), in Scenario 1. Input parameters as in figure 3.4.

In figure 3.4 the responses of the three tree length measures to changes in three parameters is illustrated; P_{ex} ranges from 0 to 0.06 and P_{ana-} ranges from 0 to 0.50, and $F_{compchar}$ equals either 0 (first column) or 0.50 (second column).

It is clear that the reversal index R_1 is affected by P_{ana-} , and also slightly by P_{ex} , but not by $F_{compchar}$. The trends shown in figure 3.4 for R_1 are very distinct and also the standard deviations of the means are very small. DI is affected by changes in $F_{compchar}$ and P_{ex} , but hardly at all by P_{ana-} . Figure 3.4 shows that CI is affected by $F_{compchar}$ and P_{ana-} and also strongly by P_{ex} .

In figure 3.5 the effects of the input parameters on two tree topology measures are illustrated. The Colless index of symmetry

(I_{col}) is not affected by P_{ex} , P_{ana} or $F_{compchar}$. The stemminess index (I_{stem}) is affected by changes in P_{ex} but not by P_{ana} and $F_{compchar}$.

Quantification of the responses in Scenario 1

A multiple regression analysis was carried out to estimate the relationships between each tree statistic and the input parameters P_{ex} , P_{ana} and $F_{compchar}$. The analysis was performed using the information of all 133 points in the parameter space of figure 3.3. The dependent variables are the mean values of the tree statistics ($n = 25$). The independent variables used in the regression model are P_{ex} , P_{ana} , $F_{compchar}$, $(P_{ex} \bullet P_{ana})$, $(P_{ex} \bullet F_{compchar})$, $(P_{ana} \bullet F_{compchar})$, P_{ex}^2 , P_{ana}^2 and $F_{compchar}^2$. The prediction equations for each tree statistic are given in table 3.2 together with their coefficients of multiple determination (R^2).

The R^2 values in table 3.2 indicate that for R_1 , Cl and DI , the regression models used, describe the data adequately. Only 74 percent of the variation of I_{stem} is explained and the variation in I_{col} is not explained at all.

Table 3.2 Prediction equations for regression of tree statistics (indices) on the independent variables, and the value of the coefficient of multiple determination (R^2), for Scenario 1. Partial regression coefficients are given only if significant (t-test at $p < 0.05$).

index	R^2	constant	P_{ex}	P_{ana}	$F_{compchar}$	$\frac{P_{ex} \bullet P_{ana}}{P_{ana} \bullet F_{compchar}}$	$\frac{P_{ex} \bullet F_{compchar}}{F_{compchar}}$	$\frac{P_{ana} \bullet F_{compchar}}{F_{compchar}}$	P_{ex}^2	P_{ana}^2	$F_{compchar}^2$
R_1	.98	.9984		.0087		.0008					-.0001
Cl	.94	.3415	-.0084	.0018	.0008		-.0001	-.0001			.0001
DI	.97	.8171	.0077	.0003	.0011		-.0001	-.0000			.0000
I_{col}	.03	.1080	.0005								
I_{stem}	.74	.2367	.0082				-.0001	-.0000			.0000

Since the independent variables have different ranges, the relative importance of the variables cannot be inferred directly from their partial regression coefficients in the regression equations.

Comparing the results of the simple 'RIF-method' for Scenario 1 and the regression analysis, it appears that both methods lead to very much the same conclusions. Therefore, I feel justified in using

the 'RIF-method' as it gives reliable results, both in a shorter time and with a saving of CPU_time.

Analysis based on 20 species each with 50 characters

The analysis, as presented above, is based on simulations that have generated 50 recent species each with 100 characters. The 'RIF-method' was also applied in simulation experiments using 20 species each with 50 characters. The probability of character state change was set at 0.03 in Scenario 1, which resulted in approximately 350 evolutionary steps. This probability was adjusted to equalize the total number of evolutionary changes (L_{act}) in the other standard simulations (for $P_{ex} = 0$, $F_{compchar} = 0$ and/or $P_d = 1$). The ranges for the variables were determined in the same way as in the 50-species simulations. In table 3.3 the results for the 20-species simulations are presented. In general, the results of this RIF-analysis seem to be in good agreement with those presented in table 3.1. However, there are some differences in the results for I_{stem} , especially in the Scenarios 2 and 4. In the 20-species simulations this index is clearly affected by P_p , which is not surprising since unequal stem-lengths reduce the values of I_{stem} . As yet, I cannot explain why this effect should be absent in the 50-species simulations. Furthermore there are differences in the relative magnitudes of the values for R_1 , I_{col} , and the stemminess indices, which tend to have higher values in the 20-species simulations.

General conclusions and relevance of sensitivity analysis

From the sensitivity analysis it is clear that the tree length measures can, to a certain degree, be used to characterize the output data sets. The indices R_1 , DI and CI clearly have differing responses to input parameter changes. However, the analysis demonstrates that one index can be affected by several input parameters at the same time. This is most clearly illustrated by CI , but in fact all indices are influenced by more than one input variable.

Chapter 3

Table 3.3 Results of the sensitivity analysis using the RIF-method in the 20-species simulations. Numbers are RIF-values (see text for explanation). Results are based on simulation sets of 25 runs each.

Evolutionary scenario	Independent variables	Dependent variables					
		R_I	DI	CI	I_{col}	I_{stem}	I_{stemFS}
Radiation/ Gradual	P_{ana}	25	5	25	2	2	2
	$F_{compchar}$	2	15	38	2	3	2
	P_{ex}	6	4	3	2	7	9
Radiation/ Punctuated	P_{clado}	22	5	19	1	2	2
	$F_{compchar}$	2	19	45	2	2	3
	P_{ex}	6	7	4	2	9	7
	P_p	1	2	4	2	18	2
Equilibrium/ Gradual	P_{ana}	9	2	11	1	3	2
	$F_{compchar}$	1	9	28	2	12	1
	P_d	1	1	3	2	12	4
Equilibrium/ Punctuated	P_{clado}	8	2	12	2	1	2
	$F_{compchar}$	1	10	27	2	2	2
	P_d	2	2	2	1	6	1
	P_p	1	1	2	2	17	2
Range		2.74	1	1	1	0.52	0.64

The tree topology measures seem to be hardly (I_{stem}) or not at all (I_{col} , I_{stemFS}) influenced by changes in input parameters in any of the scenarios. This means that they do not describe properties of trees that can be effectively controlled by the user of GENESIS by choosing certain parameter values or evolutionary scenarios. Nevertheless these properties may well be correlated with the relative efficiency c.q. accuracy of reconstruction methods to discover the true tree.

For a correct interpretation of the effects of a change in one input parameter on the output data, one needs to take care that the conditions defined by the other parameters remain as constant as possible. Consequently, the probability of character state change was adjusted to equalize the number of evolutionary steps among the standard simulations. However, this adjustment will be rather difficult in cases where $F_{compchar}$ has a high value (0.75 - 1.00) and also when $P_{ex} > 0$, especially in situations where the value of P_{ex} fluctuates throughout a simulation as in Scenario's 3 and 4.

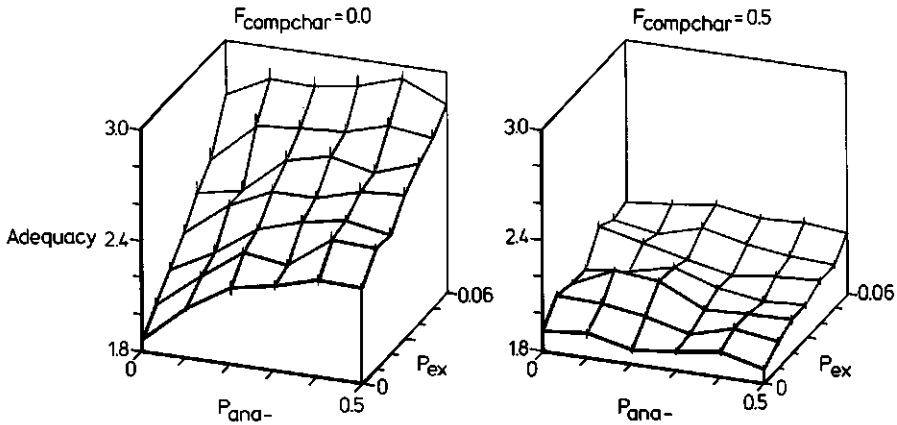


Figure 3.6 Responses of changes in input parameters on the adequacy as measured by $N_{pseu} / (2S - 3)$, (N_{pseu} = number of pseudo-characters, S = number of species), in Scenario 1. Input parameters as in figure 3.4.

Also Fiala & Sokal (1985) were not able to "devise a reliable adjustment criterion" for L_{act} . Preliminary evaluation experiments with GENESIS revealed that the accuracy of reconstruction methods increases with the number of characters c.q. pseudocharacters. Also Rohlf & Wooten (1988) found that the number of characters has a strong effect on the accuracy. The adequacy of the data will therefore be an important predictor of accuracy. From figure 3.6 it appears that the adequacy, expressed as $N_{pseu} / (2S - 3)$, does indeed depend on P_{ex} and $F_{compchar}$ in Scenario 1. Table 3.4 presents the RIF-values of this index for all scenarios. Note that only the relative magnitudes of the values should be taken into account. In spite of some efforts to equalize the number of evolutionary steps between scenarios, it is evident from this table, especially if we look at Scenarios 1 and 2, that the adequacy depends on the values of the retrogression parameter, $F_{compchar}$ as well as those of P_{ex} .

Summarizing, none of the tree statistics investigated seems to describe a single property of the evolutionary process in an unambiguous way. However, as Rohlf & Wooten (1988) rightly pointed out, even if an index exists that "serves as a good predictor of how well a phylogenetic tree can be estimated, one encounters the problem that one must know the true tree in order to compute the index."

Table 3.4 RIF-values for adequacy ($N_{pseu} / (2S - 3)$). The range has been determined in an arbitrary way, by choosing 0.51 as the lower limit and 5.08 as the upper limit. These limits correspond with average character lengths equal to 1 and 10 respectively. The lowest and highest values recorded were 1.03 (average character length: 4.06) and 3.45 (average length: 13.94). Results based on simulation sets of 50 runs each.

Independent variable	Evolutionary scenario			
	1	2	3	4
P_{ana} / P_{clado}	8	10	3	3
$F_{computer}$	20	22	5	5
P_{ex}	11	13	-	-
P_p	-	1	-	0
P_d	-	-	0	1

Both Fiala & Sokal (1985) and Rohlf & Wooten (1988) concluded from their evaluation experiments that the tree topology (stemminess) is the most important predictor affecting accuracy. Fiala & Sokal (1985) found that all the methods tested by them (UPGMA-clustering, Wagner parsimony and character compatibility) gave similar results and that none of them were very accurate. Rohlf & Wooten (1988) found that for large numbers of characters all the methods they investigated (UPGMA-clustering, Wagner parsimony and Restricted maximum-likelihood method) were very "similar in their ability to estimate the true cladograms." Kim & Burgman (1988) state that, in general, their results are in agreement with previous work, but that some of their findings are in conflict. Among the methods they employed (UPGMA-clustering, Wagner parsimony and maximum likelihood) they found maximum likelihood to perform the best. Moreover they even found that under certain conditions the accuracy of the parsimony method increased with a decrease in the numbers of characters (loci) (see their figure 4).

These three studies are all based on simulation models with different assumptions about the evolutionary process, and also different numbers of characters per species and recent species generated. Therefore it seems difficult to assess to what extent these studies and their results are comparable. Kim & Burgman (1988) likewise 'complain' that "it is difficult to identify the reason for conflicting results".

The present sensitivity analysis has shown that none of the tree statistics tested can be expected to be an unambiguous estimator of accuracy. A sensitivity analysis may reveal the extent to which the various statistics are interdependent and how far they depend on the input parameters. I feel that a good understanding of the behaviour of the simulation model used for evaluating the accuracy of reconstruction methods is an essential prerequisite for a correct interpretation of the results. This understanding will be of valuable help in designing the evaluation experiments. Moreover, it will make comparison with the results of other similar studies and identification of the cause(s) for conflicting results more easy.

Zusammenfassung

GENESIS: Ein Simulationsmodell der Phylogenie Sensitivitätsanalyse

Das prinzip des Simulationsmodells GENESIS wurde bereits in einer früheren Publikation (Heijerman, 1988) dargestellt. Die vorliegende Arbeit stellt die Ergebnisse einer Sensitivitätsanalyse dar. Bei dieser Analyse wurden die verschiedenen Eingabewerte variiert und deren Effekte auf die Ausgabewerte ermittelt, so wie sie die verschiedenen Stammbaum-Statistiken zusammenfassen. Die Ergebnisse zeigen, daß keine der getesteten Statistiken als eindeutig bestes Entscheidungskriterium für die Güte der verschiedenen Konstruktionsmethoden angesehen werden kann. Die Analyse hilft aber, mehr Einblick in das Verhalten und die Brauchbarkeit der Modelle zu erlangen. Dies ist jedoch für eine korrekte Interpretation der Ergebnisse der Simulationsexperimente, die mit GENESIS durchgeführt werden können, Voraussetzung.

Acknowledgements

I am grateful to drs P. De Vrijer, A. De Winter, R. Daamen and R. Zandee for discussions and comments on the draft. Mrs C. Hengeveld corrected the English text, and Mr P. Kostense made the drawings.

References

- Colless, D. H., 1982. (Review of) Phylogenetics: The theory and practice of phylogenetic systematics. — *Systematic Zoology* 31: 100-220.
- Fiala, K. L. & R. R. Sokal, 1985. Factors determining the accuracy of cladogram estimation: evaluation using computer simulation. — *Evolution* 39: 609-622.
- Heijerman, Th., 1988. GENESIS: a simulation model of phylogeny. 1. The origin and early evolution of character state vectors. — *Zeitschrift für zoologische Systematik und Evolutionsforschung* 26: 409-424.
- Kim, J. & M. A. Burgman, 1988. Accuracy of phylogenetic estimation methods under unequal evolutionary rates. — *Evolution* 42: 596-602.
- Rohlf, F. J. & M. C. Wooten, 1988. Evaluation of the restricted maximum-likelihood method for estimating phylogenetic trees using simulated allele-frequency data. — *Evolution* 42: 581- 595.
- Sokal, R. R., 1983. A phylogenetic analysis of the Caminalcules. I. The data base. — *Systematic Zoology* 32: 159-184.

4

Adequacy of numerical taxonomic methods

A comparative study based on simulation experiments ¹

"Techniques should complement rather than compete". Moss (1983:75)

Abstract

A simulation model, GENESIS, was developed to examine the relative merits of several phylogenetic and phenetic methods. For a description of the model and the results of a sensitivity analysis, see Heijerman (1988, 1990). GENESIS was designed to generate artificial data sets of 'species' with known phylogenies. These data sets were subjected to character analysis by various numerical taxonomic methods (UPGMA clustering, Wagner parsimony analysis and component-compatibility analysis). The results of analysis were compared with the true phylogeny. The agreement between the true tree and the reconstructed tree was used as a measure of quality (adequacy). By varying the input parameters of GENESIS, output produced under different evolutionary scenarios was obtained and the relative adequacies of the methods in relation to these evolutionary conditions were evaluated. The overall differences in adequacy between Wagner parsimony as performed by PAUP, PHYLIP (MIX) and Hennig86, and UPGMA clustering with product moment correlations of unstandardized characters, were rather small. These methods were more adequate than Wagner parsimony with Wagner78 and group compatibility with CAFCA. The adequacy of the trees as

¹ Published as: Heijerman, Th., 1992. Adequacy of numerical taxonomic methods. A comparative study based on simulation experiments. — *Zeitschrift für zoologische Systematik und Evolutionsforschung* 30: 1-20.

estimated by Wagner78, PAUP, PHYLIP and CAFCA depended on several tree properties, the consistency index being the most important one.

Key words: Numerical taxonomy — Phylogeny — Simulation model — Cladogram estimation

Introduction

It is a principle goal of systematists and evolutionary biologists to construct natural classifications of living organisms. Such classifications are considered useful because they not only categorize for identification purposes, but they also reflect phylogenetic relationships. Because of this, they can serve as the best general reference system, having predictive value for phylogenetic relationships.

There are two major and seemingly conflicting approaches to classification: phenetics and phylogenetic systematics. In the phenetic approach, species are grouped together on the basis of overall similarity, whereas phylogenetic systematists attempt to base their classifications on the actual phylogeny. Adherents of the two approaches both claim to have succeeded in demonstrating the superiority of their own classifications or the inferiority of the classifications of the proponents of the other school (e.g. Farris, 1977; Funk 1983; Schuh & Polhemus 1980; Sokal & Rohlf, 1981). Many proponents of the phenetic or the phylogenetic 'school' seem to favour their own approach and, simultaneously, to disapprove of the alternative one, in a rather dogmatic way. To illustrate this, Funk (1983:21) can be cited: "The conclusion is that only by abandoning the ideas of overall similarity and character weighting can we find the natural classification that we seek." And Sneath (1983:28) states: "The objection to basing general biological classification on phylogeny is simply stated, quite apart from considerations of the purpose of classification: it is not desirable for science to be based on the unverifiable." In this, I fully agree with Moss (1983:74, 75):

"Diversity of approach and methodology is extremely useful in our attempts to interpret patterns of biological diversity", and "... it is foolish to claim victory for one side or for one method, or to insist that a modern taxonomist limit his techniques ...".

In many studies, various criteria have been used to investigate the differences in performance between these two main approaches. In general, those methods that result in stable and natural classifications that have predictive value are considered superior. Some authors have used another approach to examine the relative merits of the various phenetic and phylogenetic methods, using artificial data sets. Sokal (1983a, 1983b) has employed this approach by analyzing the character data set of the Caminalcules, a group of artificially generated organisms with a known phylogeny. He examined several phenetic and phylogenetic (cladistic) methods and found both the Wagner method and the Camin-Sokal method, as furnished by the PHYLIP program package, to produce trees that were closest to the true cladogram.

Similar studies, but based on computer simulations, were published by Fiala & Sokal (1985), Rohlf & Wooten (1988) and Kim & Burgman (1988). Fiala & Sokal (1985) examined three methods (Wagner parsimony, character compatibility and UPGMA clustering) and they concluded that "... none of the methods is very accurate, that the differences among them are rather small, and that historical effects (the branching pattern of a phylogeny) may outweigh biological effects in determining the accuracy with which a phylogeny can be reconstructed." Rohlf & Wooten (1988) found that the "overall level of accuracy of tree reconstruction depends on the topology of the true phylogenetic tree". Also the accuracy was found to increase if more characters are used, UPGMA to perform better if data sets with a small number of characters are tested, the Wagner tree method and the restricted maximum likelihood method to perform better for large numbers of characters, and the restricted maximum likelihood method to be "clearly superior to the Wagner tree method". However inspection of their table 1 and figure 5 (Rohlf & Wooten, 1988:588, 590) suggests that the differences between the methods are small. Kim

& Burgman (1988) also found the maximum likelihood method to perform better than parsimony or phenetic clustering.

In this paper, the results will be presented of experiments designed to examine the relative adequacy of various numerical taxonomic methods to estimate true trees, as produced by the simulation model GENESIS.

GENESIS was constructed to produce recent species, with known character state distributions and known phylogenies. Examples of input parameters of the model are the speciation probability, the extinction probability, the probability of character state change and the probability of character state reversals. By selecting the proper values of input parameters and proper options of the model, one can produce sets of recent species based on different views about the evolutionary process. GENESIS can produce ideal data sets by selecting the proper values of parameters and by changing these values, one can introduce a certain amount of *noise* in the data. The properties of the data sets, as produced by the various runs of GENESIS, can be described by several tree statistics. For more details about GENESIS, one is referred to Heijerman (1988) and to Heijerman (1990), where the results of a sensitivity analysis are presented.

GENESIS was run to produce data matrices, together with their corresponding true trees, under different evolutionary conditions. The efficiencies of the various methods to find the true trees were evaluated in relation to several properties of the evolutionary process simulated, as expressed in properties of the true trees generated. The adequacy of the estimation methods was measured by the consensus of the true tree with the estimated tree.

The simulation model developed and used by Fiala & Sokal (1985) shows some similarities with GENESIS. Among other things, they studied the effects of several tree length measures on the accuracy of the phylogenetic estimates by computing regressions. However in the present study, some tree statistics were included in the analysis that were not evaluated by Fiala & Sokal. Also Fiala & Sokal evaluated one UPGMA method, the Wagner method as performed by Wagner78 (Farris, 1978), and a

character compatibility analysis as performed by Fiala's program CLINCH. In the present study, four UPGMA techniques were evaluated, three programs that calculate Wagner trees, and a recently published group compatibility program (CAFCA; Zandee, 1988). Unfortunately none of the authors of the simulation models mentioned performed a sensitivity analysis. So the reasons for different results will be difficult to identify.

Methods

Numerical taxonomic methods evaluated

As a phenetic method the UPGMA (unweighted pair-group methods using arithmetic averages) technique was applied, using the CLUSTAN package (Wishart, 1982). Phenograms were constructed, based on matrices of euclidian distances and product moment correlations, calculated both from non-standardized as well as from standardized data sets. For details on these methods see Sneath & Sokal (1973).

Minimum length trees were constructed using the distance Wagner procedure of Wagner78 (Farris, 1978), the Wagner parsimony method as provided by the MIX program in the PHYLIP package, version 3.0 (Felsenstein, 1987), PAUP, version 2.4 (Swofford, 1985) and Hennig86, version 1.5 (Farris, 1988).

Wagner78 was run using the HOM option only, to request computation of the total homoplasy and the deviation ratio. As a first taxon, the ancestor (all characters are in state "0") was added to the character matrix. The results are considered not to depend on the order of species in the input matrix and so one single run should be sufficient. Unexpectedly, however, the results were found to be dependent on the species order and so Wagner78 was run ten times with different orderings of species in the input matrix.

The MIX program was used to carry out the Wagner parsimony method, using the "A" (ancestral states) option, and the ancestor was added to the data set of recent species. Since

the results depend on the order of the species in the input matrix, Felsenstein (1987) advised at least ten runs. Because MIX runs rather slowly (VAX 8600), it was run only four times with different orderings.

The PAUP documentation recommends one to try a variety of combinations of options, to find the most parsimonious trees. Four combinations were tried, all of which used SWAP = GLOBAL, ROOT = ANCESTOR, MULPARS, STATS and FVALUE. The combinations differed with respect to the values of the ADDSEQ and HOLD options. Combination 1 used ADDSEQ = CLOSEST, HOLD = 1; combination 2 used ADDSEQ = ASIS, HOLD = 5; combinations 3 used ADDSEQ = SIMPLE, HOLD = 10; and combination 4 used ADDSEQ = ROOTLESS and HOLD = 15. All characters were treated as ordered characters (default option).

SWAP "selects the level of branch-swapping to be performed", and "in GLOBAL swapping, each possible subtree is removed from a tree and reinserted at all other positions on the tree". MULPARS "initiates a search for multiple equally parsimonious trees via branch-swapping". ROOT "selects the procedure to root the tree(s) ..." and if ROOT = ANCESTOR "an OTU designated as the hypothetical ancestor is placed at the base of the tree". The ancestor (all characters in state "0") was added as the first species in the character matrix. ADDSEQ "sets the method used to determine the order in which OTUs will add to the tree during stepwise OTU addition". For details on the four options (SIMPLE, ROOTLESS, ASIS and CLOSEST), one is referred to the PAUP manual. HOLD "specifies the number of trees to be held at each step of the tree construction". The STATS and FVALUE option are used to request the computation of the tree length, the consistency index, and the *f*-value as defined by Farris (1972). For further details on the specified options and their values, one is referred to the PAUP manual.

To calculate minimum-length trees with Hennig86, the mhennig* and bb options were used, an option recommended by Farris (1988) and found by Platnick (1989) to perform best. The mhennig* command "applies branch-swapping to each of the

initial trees, retaining no more than one tree for each initial one", and the bb command "applies extended branch-swapping to the trees in the current tree file, producing a new tree file. The shortest trees found are retained", and "... bb will generate all trees it can find". As in PAUP, all characters were treated as ordered (default option).

As a compatibility method, CAFCA, Version 1.9 (Zandee, 1988) was applied. A primary analysis was run, using the default cladon option (partial monothetic sets). Again as the outgroup the ancestor was added as a species with state "0" for all characters. All characters were treated as ordered. CAFCA offers six selection criteria for cladograms. Criterion 1 is the number of homoplasious events, i.e. "all character states requiring more than 1 step to explain their distribution over terminal taxa in the cladogram (reversal, parallelism, convergence)". Criterion 2 (support) concerns the total of single origins, which equals the number of characters that "require 1 step for their origin explaining their distribution". Criterion 3 is referred to as "The balance between events of homoplasy and support, irrespective of the number of steps (default criterion)". Criterion 4 is the total number of state changes and "considers all steps needed to explain the distribution of all character states, without differentiation as to the quality of these steps". Criterion 5 is the redundancy index (*Ri*; Geesink & Zandee, in preparation). Finally Criterion 6 is the consistency index, defined as the ratio of the theoretical minimum of steps given the number of character states, and the actual number of steps needed in the cladogram to explain all distributions of character states over taxa (Kluge & Farris, 1969). In this study, Criteria 3, 4 and 6 were used.

There are two differences in the way characters were treated, between CAFCA on the one side, and Wagner78, PAUP and Hennig86 on the other. CAFCA was used treating all characters as ordered, that is, a priori polarized (e.g. 0 → 1 → 2 → 3, etc.). However, if a more parsimonious solution would result, CAFCA will change the ordered sequence of character states (e.g. 1 → 0). Secondly, suppose that the transformation series of a certain character is as follows: 0 → 1 → 2 → 3, and that character

states 0, 1 and 3 are present in recent species, whereas state 2 only occurred in a species now extinct. In such a case CAFCA will treat the difference between states 1 and 3 as a single evolutionary step. In the other methods, this difference will be two evolutionary steps. Because of this one must be careful in comparing the results of CAFCA and the other phylogenetic estimation methods, with regard to the total length of the estimated trees as well as their consistency index.

MIX and PAUP can only produce fully resolved bifurcating trees, and this may lead to internal tree segments of zero length. This implies that the set of output trees may contain some identical trees. Wagner78, Hennig86, CAFCA and UPGMA can produce polyfurcating trees, but the trees as produced by UPGMA were always fully resolved.

The quality measure used

As a measure of the adequacy the consensus fork index (*CFI*, also CI_c , e.g. Rohlf, 1982; Shao, 1983) was calculated. *CFI* is defined as the total number of branching points in the strict consensus tree (exclusive of the basal one), divided by the number of OTUs minus 2. If the true tree and the estimated tree have exactly the same topology, their strict consensus tree will be fully resolved, resulting in *CFI* equal to 1. If there is no agreement at all between the two trees, the consensus tree will be a bush, and *CFI* will equal 0. Also various other consensus indices were calculated. The results of an evaluation of the properties of the various consensus indices will be presented in another paper.

Tree statistics

To study whether and how the adequacy of the various taxonomic methods depends on or relates to properties of the true tree, some tree statistics were computed and evaluated. The reversal index (R_1) (Sokal, 1983) measures "the amount of reversals and repeats in character state changes ...". The dendritic index (*DI*) (Sokal, 1983) measures the amount of shared evolution. As another

measure of homoplasy, the consistency index (C/I) (Kluge & Farris, 1969; Sokal 1983) was calculated. As a measure of the adequacy of the characters to resolve the cladogram, the adequacy (I_{adeq}) was calculated based on binary coded characters ($I_{adeq} = N_{pseu} / (2S - 3)$, where N_{pseu} is the number of pseudocharacters or binary coded characters, see Sokal, 1983). The following indices were used as measures of the topology of a tree; the Colless index of symmetry (I_{col}) (Colless2 in Sokal, 1983); the stemminess (I_{stemFS}) as defined by Fiala & Sokal (1985); the modified stemminess index (I_{stem}) (Heijerman, 1988); the Sackin index of symmetry (I_{sack}) (SI_a in Shao, 1983; for original description, see Sackin, 1972).

Two tree statistics were also computed for the estimated trees: C/I (only for PAUP, Hennig86 and CAFCA) and I_{sack} (for all estimation methods).

Experimental design

GENESIS can simulate the evolutionary process according to four major evolutionary scenarios. In Scenario 1 and 2, speciation and extinction probabilities are constant over time and the number of extant species increases exponentially with time ('radiation' version of GENESIS). In Scenarios 3 and 4, the number of species fluctuates around an equilibrium number of recent species (the 'equilibrium' version). As for the evolution of characters, in Scenarios 1 and 3, the rate of character evolution is the same in both daughter lineages ('gradualistic' version of GENESIS). In Scenarios 2 and 4, the average rate of character change is unequal in the two daughter lineages ('punctuated' version) (see Heijerman, 1988, for further details).

In this paper, the results will be presented of experiments with data sets produced by simulations according to Scenario 1 only; the other scenarios will be dealt with in a subsequent paper. All simulations were run to produce 20 species, each with 50 characters. The probability of speciation (P_{split}) was set at 0.10, and the probability of character state change (P_{ana+}) at 0.03. In three 'experiments', three input parameters were subjected to

change: the number fraction of characters that are fully consistent with the tree ($F_{compchar}$) (experiment 1), the probability of a reversed character state change (P_{ana-}) (experiment 2), and the probability of extinction (P_{ex}) (experiment 3). All these parameters may influence the degree by which the resulting character state distributions reflect the evolutionary history of the extant taxa. History is only truly (100% correct) reflected when $P_{ex} = 0$, $P_{ana-} = 0$ and $F_{compchar} = 1$. $F_{compchar}$ was set at 0, 0.2, 0.4, 0.6, 0.8 and 1.0 with $P_{ex} = 0$ and $P_{ana-} = 0$ (experiment 1), P_{ana-} was set at 0, 0.1, 0.2, 0.3, 0.4 and 0.5, with $P_{ex} = 0$ and $F_{compchar} = 1.0$ (experiment 2) and P_{ex} was set at 0, 0.01, 0.02, 0.03, 0.04, 0.05 and 0.06 with $P_{ana-} = 0$ and $F_{compchar} = 0.6$ (experiment 3). The choice for the parameter values was based on experience gained during the sensitivity analysis of GENESIS (Heijerman, 1990). Every simulation was run at least ten times with the same input values. There are 17 different combinations of input parameter values. The total number of data matrices and true trees generated amounts to 188. All these data matrices were used as input for the numerical taxonomic methods to be examined.

The four UPGMA strategies only produced a single phenogram for each analysis. The consensus (*CFI*) of each phenogram with the true tree was calculated.

Wagner78 also produced only one tree per run. However as the results proved to be dependent on the order of species in the input matrix, the order was randomly reshuffled ten times and Wagner78 was rerun on each of the ten resulting data sets. From these ten solutions, the shortest tree was selected for further analysis.

Each of the four combinations of options of PAUP may produce several equally parsimonious trees. From the four sets of equally parsimonious trees, the set containing the shortest trees was selected. If two or more options provided trees with the same length, the largest set was selected.

The results produced by PHYLIP depend on the order of species in the input matrix and therefore four runs were performed with different orderings of the species. Each run may result in more than one solution. For further analysis, the set containing the

shortest trees was selected. If sets with equally parsimonious trees resulted, the largest one was used.

Hennig86 became available to me after I had finished the analysis with the other programs. Nevertheless it was used to calculate minimum-length trees. However it always proved to find trees equal in length with the trees found by PAUP (on one occasion only, PAUP found a shorter solution than Hennig86), although the number of equally parsimonious trees found would often differ. So the results of Hennig86 were not further analyzed, assuming that no great differences would have been found from the PAUP results.

All trees from the selected sets from PAUP and PHYLIP were each compared with the true tree separately; that is, the *CFI* was calculated for all resulting consensus trees. For further analysis, the mean *CFI* was used, as calculated over all consensus trees resulting from the same set.

CAFCA may also produce many trees, which, however, need not all be of equal length. CAFCA provides six selection criteria to choose for the set of best trees. For the analysis, I used the average values of *CFI* calculated for the consensus trees of the true tree with the set of trees with minimum length (Criterion 4), for the set of trees with maximum support (Criterion 3), and for the set containing all trees generated, separately.

Each of the 188 original true trees had to be compared with the trees generated by the various methods tested. The four phenetic methods yielded four consensus trees, Wagner78 one, PAUP and PHYLIP one 'average' consensus tree each, and CAFCA three 'average' consensus trees, adding up to a total of $10 \times 188 = 1880$ tree comparisons. CAFCA, however, occasionally crashed during cladogram evaluation. This was due to memory problems, especially when many cladograms (number of cliques > 110) were generated. As a consequence, the total number of tree comparisons over all methods was 1818 instead of 1880.

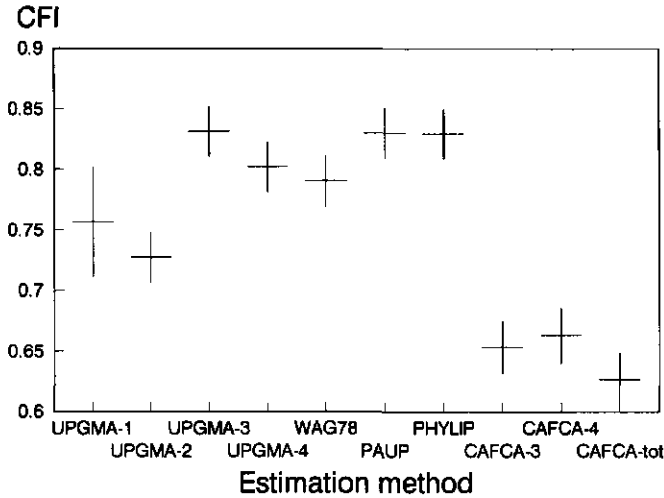


Figure 4.1 Adequacy of various estimation methods. The adequacy is measured by the consensus fork index (CFI). Phylogenies are estimated using UPGMA based on euclidian distances of unstandardized characters (UPGMA-1); UPGMA based on euclidian distances of standardized characters (UPGMA-2); UPGMA based on product moment

correlations of unstandardized characters (UPGMA-3); UPGMA based on product moment correlations of standardized characters (UPGMA-4); Wagner analysis using the Wagner78 program (WAG78); Wagner analysis using the PAUP package (PAUP); Wagner analysis using the MIX program from the PHYLIP package (PHYLIP); compatibility analysis using the CAFCA package, and using tree selection criterion 3 (see text for explanation) (CAFCA-3); compatibility analysis using the CAFCA package, and using tree selection criterion 4 (CAFCA-4); compatibility analysis using the CAFCA package, and using all trees. The horizontal bars indicate the mean CFI values; vertical bars represent the 95 % standard error bars ($n = 1818$).

Results

Overall adequacy

The results of a rough evaluation are presented in figure 4.1. These results are based on all 1818 comparisons. Figure 4.1 clearly illustrates that there are significant overall differences in performance between methods of estimation. UPGMA based on product moment correlations of unstandardized characters (UPGMA-3), PAUP and PHYLIP did not differ among each other and produced the best results. The UPGMA analysis based on unstandardized characters (UPGMA-1 and UPGMA-3) provided better results than when based on standardized characters (UPGMA-2 and UPGMA-4). Wagner78 produced results comparable with UPGMA of euclidian distances, based on

unstandardized characters (UPGMA-1). The trees as generated by CAFCA were the least accurate. There were no significant overall differences in performance between minimum-length-CAFCA trees (Criterion 4) and highest-support-CAFCA trees (Criterion 3). The trees selected by either of these two criteria, tended to be more accurate than the 'average' tree calculated from all trees generated.

Effects of input parameters and method on CFI

When the probability of extinction, the probability of a reversed character state change or the number fraction of incompatible characters was increased, the adequacy of the estimation methods decreased. The effects of method, P_{ex} , P_{ana} and $F_{compchar}$ on *CFI* were analyzed by means of a multivariate analysis of variance. All contributions were significant ($P < 0.0001$). The analysis was repeated for arcsine-transformed *CFI*, leading to exactly the same results. The results of an analysis of variance of *CFI* conducted for Wagner78, the best phenetic technique (UPGMA-3), PAUP, and CAFCA using the fourth and default selection criterion (CAFCA-4), are shown in table 4.1. All contributions are significant. The effect of method is greatest and the effect of P_{ex} is of least importance.

Table 4.1 Multivariate analysis of *CFI* between true trees and estimated trees for four numerical taxonomic methods (Wagner78, UPGMA-3, PAUP and CAFCA-4).

Source of variation	Sum of squares	d.f	Mean square	F ratio	Sign. level
Main effects	7.86	19	0.41	26.82	0.0000
method	3.47	3	1.16	75.08	0.0000
$F_{compchar}$	2.35	5	0.47	30.45	0.0000
P_{ana}	2.36	5	0.47	30.61	0.0000
P_{ex}	0.87	6	0.15	9.44	0.0000
Residual	10.94	709	0.02		
Total (corr.)	18.80	728			

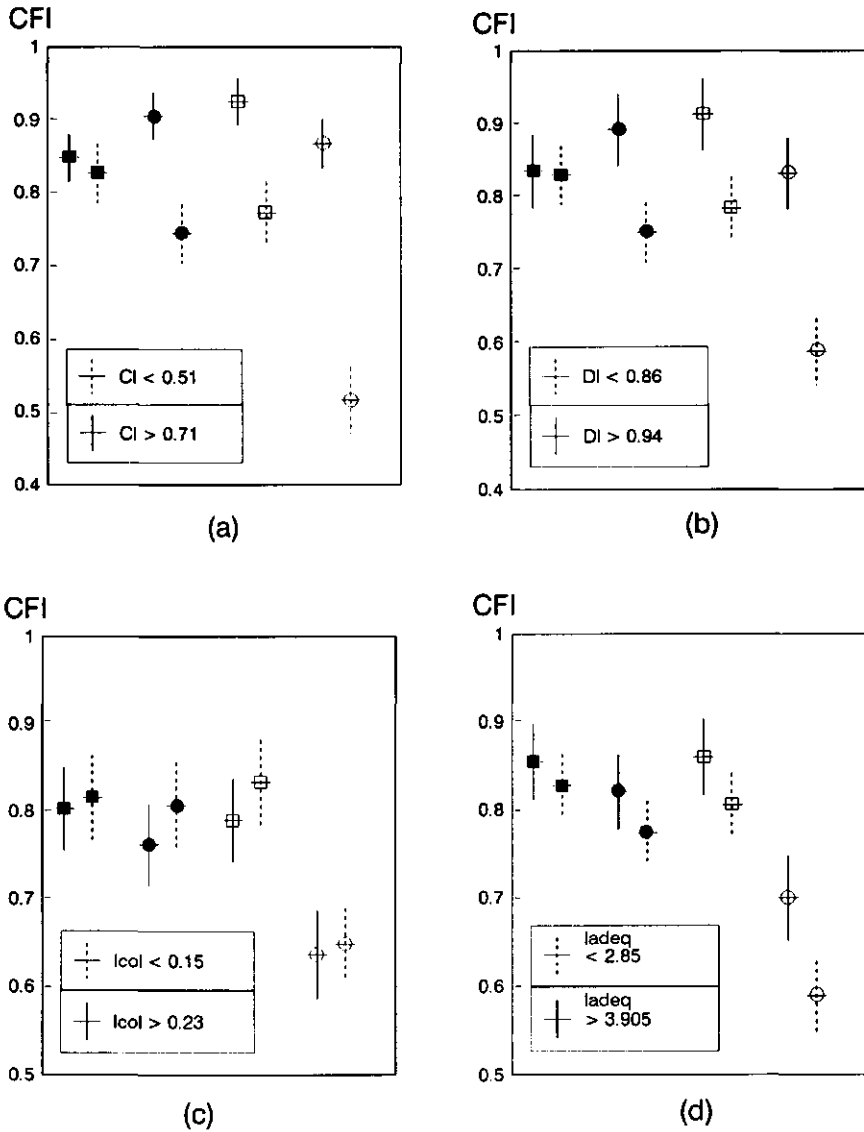
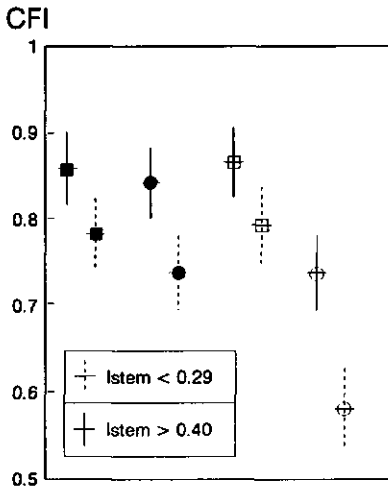
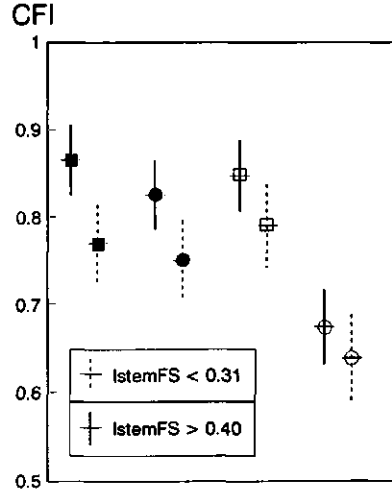


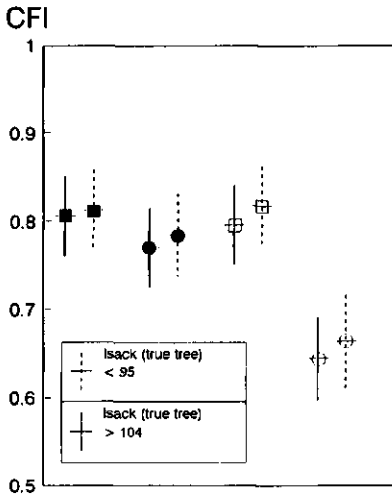
Figure 4.2a-h Effects of tree statistics on the adequacy (CFI) of four estimation methods. Solid squares: UPGMA-3; Solid circles: Wagner78; open squares: PAUP; open circles: CAFCA-4. Horizontal bars: mean CFI values; vertical bars: 95 % confidence interval for means; solid lines: CFI calculated for 25 % of the total ...



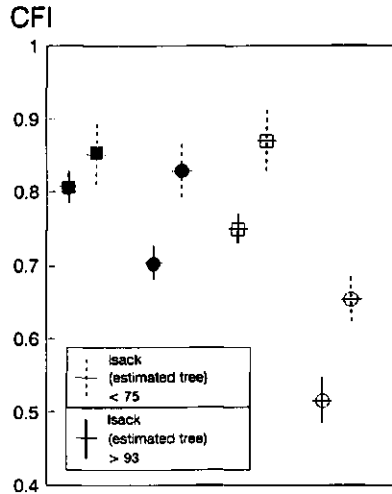
(e)



(f)



(g)



(h)

(figure 4.2, continued) ... number of trees with the highest values for the tree statistic concerned (uppermost quartile); dashed lines: CFI calculated for the 25 % of the total number of trees with the lowest values for the tree statistic concerned (lowermost quartile). Limiting quartile values are given in the figures.

Effects of tree properties on CFI

The sensitivity analysis of GENESIS (Heijerman, 1990) revealed that changing the input parameter values of $F_{compchar}$, P_{ana} , and P_{ex} affected some of the true tree statistics: R_1 is affected by P_{ana} , DI by $F_{compchar}$, CI by both P_{ana} and $F_{compchar}$, and I_{stem} and I_{stemFS} are more or less affected by P_{ex} ; I_{col} is not affected by any of the input parameters.

All the UPGMA techniques behaved in much the same way, the differences being only in the absolute values of CFI , with UPGMA-3 (product moment of unstandardized characters) always producing the best results. PHYLIP was not further considered because the results were the same as for PAUP. CAFCA-3 (highest support) and CAFCA-tot (all trees) behaved similarly to CAFCA-4 (total number of steps), and the values of CFI for CAFCA-3 did not differ greatly from those obtained for CAFCA-4. So the effects of tree statistics on the accuracy of the estimation methods will be presented and discussed only for Wagner78, UPGMA-3, PAUP and CAFCA-4.

The following procedure was carried out to find if any overall effects of tree statistics exist at all. For each method, the mean CFI value of the 25% of the trees having the highest values for a particular tree statistic was compared with the mean CFI value of the 25% of the trees with the lowest values. The results are presented in figure 4.2(a - h).

Figure 4.2 shows that the accuracy of the three phylogenetic methods is related with the majority of the tree properties. There is a strong effect of CI and DI on the adequacy of Wagner78, PAUP and CAFCA-4. The (a)symmetry of the true tree as measured both by I_{col} and I_{sack} does not significantly affect the accuracy of any method. The stemminess, as measured by both indices, is slightly related with the performance in all four methods, the effect of I_{stem} being somewhat stronger than the effect of I_{stemFS} , especially for CAFCA-4. There is some effect of I_{adeq} , especially in CAFCA-4. The results for R_1 are not presented here, but they will be dealt with when the results of the separate experiments will be discussed.

Adequacy of taxonomic methods

Table 4.2 Values of R^2 (%) of the regression analysis of adequacy (CFI) on tree statistics, for experiment 1. Source of variation: $F_{compchar}$. Regression lines for CFI on CI , DI , I_{stem} , I_{stemFS} , I_{sack} (true tree) and CI (estimated tree) are presented in figures 4.3, 4.7 and 4.8.

Tree statistics	Methods			
	UPGMA-3	Wagner78	PAUP	CAFCA-4
	Sample size			
	74	74	67	62
CI	4	40	40	66
DI	2	34	30	53
I_{adeq}	0	6	0	6
I_{col}	0	1	0	2
I_{stem}	20	45	34	45
I_{stemFS}	14	17	10	6
I_{sack}	0	1	1	1
I_{sack} (estimated tree)	12	8	10	7
CI (estimated tree)	-	-	38	76

The symmetry, as measured by I_{sack} , has also been calculated for the estimated trees. Figure 4.2h shows that there is some effect on the accuracy of all four methods.

To analyze the effects of tree properties on the adequacy of the reconstruction methods in more detail, a linear regression of CFI on each tree statistic was performed. The analysis was carried out for the three experiments separately.

In experiment 1, variation in tree statistics was introduced by varying the fraction of compatible characters ($F_{compchar}$) from 0 to 1 in steps of 0.2. Table 4.2 and figures 4.3, 4.7 and 4.8 show the results of the regression analysis of adequacy, expressed as CFI , on several tree statistics.

The indices I_{stem} , CI and DI appeared to be rather strongly correlated with adequacy, except for the UPGMA-3 method (fig. 4.3). The symmetry of the estimated tree, as calculated with I_{sack} , proved not to be a good predictor of accuracy (fig. 4.7a). The consistency index of the estimated tree also appeared to be well correlated with CFI , especially for CAFCA (fig. 4.8a).

Chapter 4

Table 4.3 Values for R^2 (%) of the regression analysis of adequacy (CFI) on tree statistics, for experiment 2. Source of variation: P_{ana} . Regression lines are presented in figures 4.4, 4.5, 4.7 and 4.8.

Tree statistics	Methods			
	UPGMA-3	Wagner78	PAUP	CAFCA-4
	Sample size			
	64	64	65	61
CI	8	51	53	67
DI	3	43	46	55
R_f	8	40	46	46
I_{adeq}	7	31	35	51
I_{col}	1	4	4	3
I_{stem}	15	0	0	0
I_{stemFS}	39	4	2	0
I_{sack}	2	6	7	6
$I_{sack}(\text{estimated tree})$	19	1	25	8
$CI(\text{estimated tree})$	-	-	47	85

Table 4.4 Values for R^2 (%) of the regression analysis of adequacy (CFI) on tree statistics, for experiment 3. Source of variation: P_{ex} . Regression lines are presented in figures 4.6, 4.7 and 4.8.

Tree statistics	Methods			
	UPGMA-3	Wagner78	PAUP	CAFCA-4
	Sample size			
	71	71	71	65
CI	4	13	18	13
DI	1	0	2	3
I_{adeq}	1	0	2	3
I_{col}	1	4	6	7
I_{stem}	3	5	7	0
I_{stemFS}	0	1	1	1
I_{sack}	1	4	7	8
$I_{sack}(\text{estimated tree})$	5	29	16	3
$CI(\text{estimated tree})$	-	-	13	38

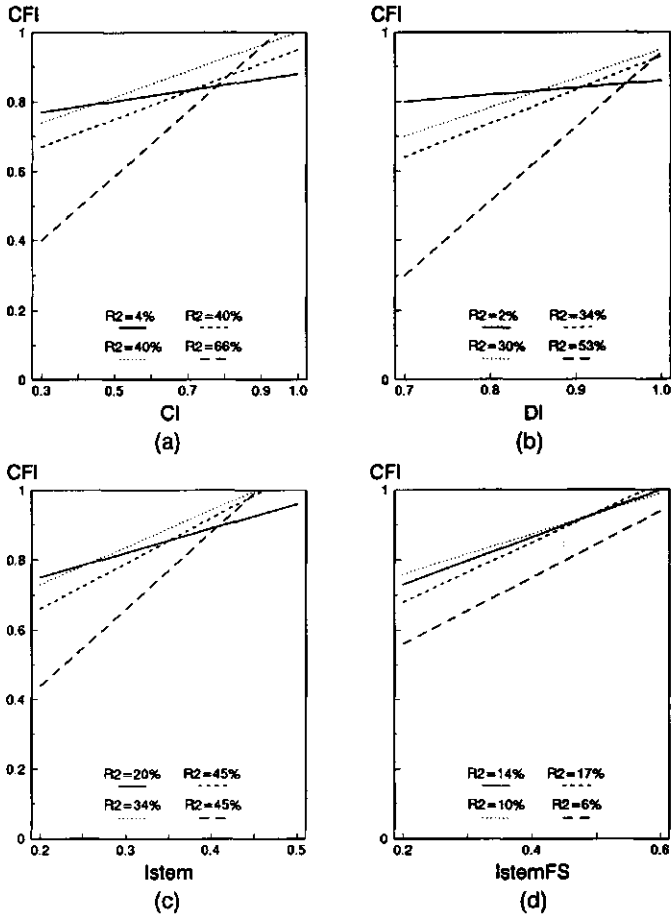


Figure 4.3a-d Regression of adequacy (*CFI*) on tree statistics in experiment 1 (source of variation: $F_{compchar}$). Estimated regression lines are drawn for UPGMA-3 (solid lines), Wagner78 (dashed), PAUP (dotted) and CAFCA-4 (long dashed), corresponding values of R^2 are given in the figures.

In experiment 2, the source of variation was the retrogression parameter (P_{ana}). Table 4.3 gives the R^2 values and figures 4.4, 4.5 and 4.7 show the regression of *CFI* on the main tree statistics. *CI*, *DI*, *R1* and also I_{adeq} seem to be good predictors of the adequacy for the phylogenetic methods (fig. 4.4). The results of the UPGMA analysis, however, were correlated with the stemminess, and especially with I_{stemFS} (fig. 4.5). As in experiment 1, the consistency index of the estimates of PAUP and CAFCA-4 (fig. 4.8b) was a good predictor of adequacy.

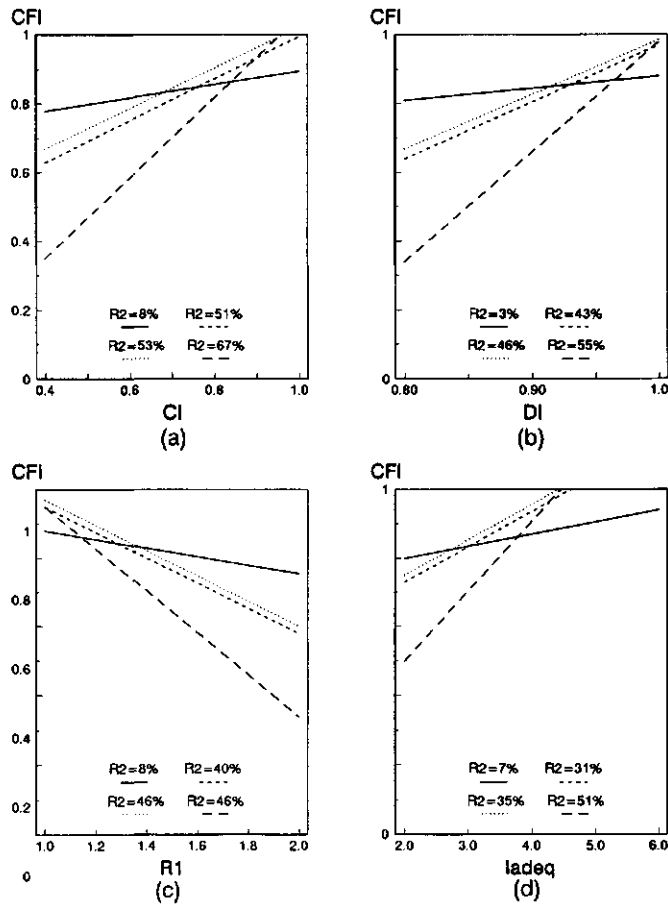


Figure 4.4a-d Regression of adequacy (*CFI*) on tree statistics in experiment 2 (source of variation: P_{ana}). Estimated regression lines are drawn for UPGMA-3 (solid lines), Wagner78 (dashed), PAUP (dotted) and CAFCA-4 (long dashed), corresponding values of R^2 are given in the figures.

In experiment 3, the source of variation was the probability of extinction (P_{ex}). Table 4.4 presents R^2 for the regression analysis of adequacy on various tree statistics, and some regression lines are shown in figures 4.6, 4.7 and 4.8. The index *CI* is slightly correlated with *CFI* in the three phylogenetic methods (fig. 4.6) and all other true tree statistics are very poor estimators of adequacy.

Adequacy of taxonomic methods

Table 4.5 Values for R^2 (%) of the regression analysis of adequacy (CI) on tree statistics for all experiments taken together. Source of variation: $F_{compchar}$, P_{ana} , P_{ex} .

Tree statistics	Methods			
	UPGMA-3	Wagner78	PAUP	CAFCA-4
	Sample size			
	188	188	184	169
CI	8	28	30	45
DI	0	11	13	16
R_1	0	0	2	2
I_{adeq}	0	1	2	1
I_{col}	0	2	3	1
I_{stem}	0	12	7	8
I_{stemFS}	0	4	3	0
I_{sack}	9	2	15	2
$I_{sack}(estimated\ tree)$	11	16	17	0
$CI(estimated\ tree)$	-	-	22	49

However, as in experiments 1 and 2, the consistency index (fig. 4.8c) of the estimated tree was strongly correlated with the adequacy of the PAUP and especially the CAFCA-4 results.

Finally table 4.5 shows R^2 for the regression analysis with the data of the three experiments together. The adequacy of the UPGMA-3 method did not relate significantly to any of the true tree statistics. The adequacy of the phylogenetic methods were rather strongly correlated with CI , and to a lesser extent also with DI . The adequacy of both PAUP and CAFCA-4 was also positively correlated with the consistency index of the estimated tree. For both CAFCA and PAUP, the CI of the true tree was positively correlated with the CI of the estimated tree. The R_2 for the regressions of $CI(true\ tree)$ on $CI(estimated\ tree)$ for the CAFCA and PAUP results were 77.71 and 87.42 %, respectively.

In figure 4.9, $CI(estimated\ tree)$ is plotted versus $CI(true\ tree)$ for the PAUP results; the CI of the estimated tree was always higher than, or equal to the CI of the corresponding true tree. In contrast, 16 % of the trees estimated by CAFCA, had a lower CI than their corresponding true trees (fig. 4.10). This difference, however, partly resulted from the different ways in which characters were treated by CAFCA and PAUP.

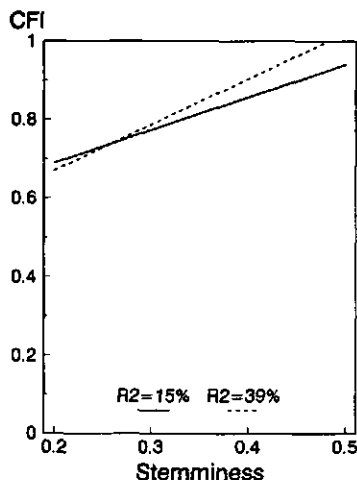


Figure 4.5 Regression of adequacy (CFI) on stemminess as measured by I_{stem} (solid line) and I_{stemFS} (dashed line) in experiment 2 (source of variation: P_{ana}). Estimated regression lines are drawn for UPGMA-3, corresponding values of R^2 are given in the figure.

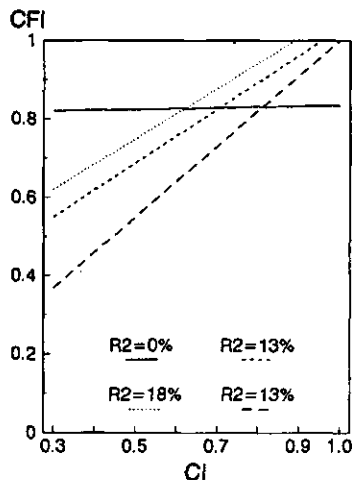


Figure 4.6 Regression of adequacy (CFI) on the consistency index (CI) in experiment 3 (source of variation: P_{ex}). Estimated regression lines are drawn for UPGMA-3 (solid lines), Wagner78 (dashed), PAUP (dotted) and CAFCA-4 (long dashed), corresponding values of R^2 are given in the figure.

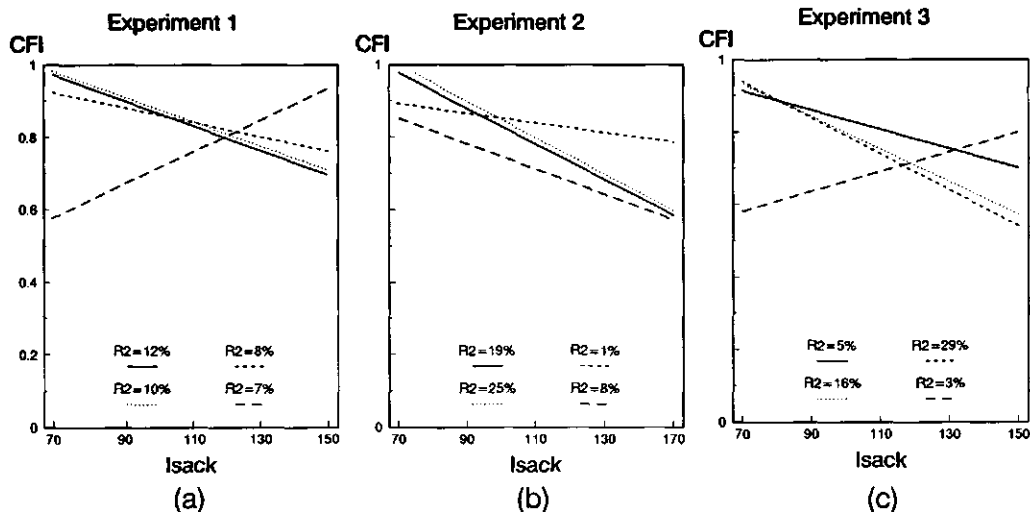


Figure 4.7a-c Regression of adequacy (CFI) on the symmetry of the estimated tree, as measured with I_{sack} in experiments 1, 2 and 3. Estimated regression lines are drawn for UPGMA-3 (solid lines), Wagner78 (dashed), PAUP (dotted) and CAFCA-4 (long dashed), corresponding values for R^2 are given in the figures.

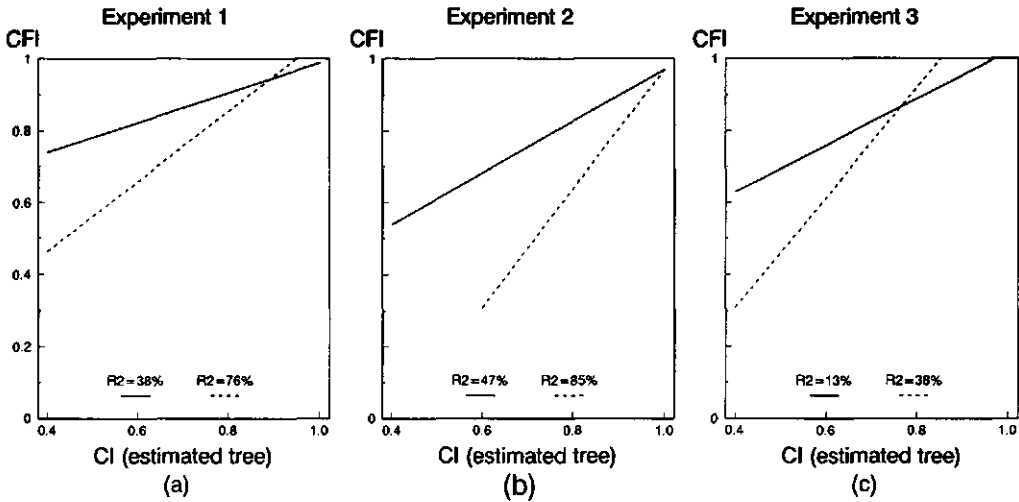


Figure 4.8a-c Regression of adequacy (*CFI*) on the consistency index of the estimated tree, as measured by *CI*; results for experiment 1, 2 and 3 taken together. Estimated regression lines are drawn for PAUP (solid lines) and CAFCA-4 (dashed), corresponding R^2 values are given in the figures.

Other aspects of efficiency

The adequacy of the estimation of the true phylogenetic tree, of course, is the main aspect of efficiency of a method of phylogenetic estimation. However some other aspects of efficiency may also be considered.

The parsimony methods tested were not guaranteed to always find the shortest tree(s). Those methods can be considered most efficient that find the greatest number of trees, given these do not differ in tree length. The mean lengths of the most parsimonious trees produced by PAUP, PHYLIP (MIX) and Wagner78, equalled 183.5, 183.6 and 185.1 steps respectively ($n = 188$). In five runs (3%) only, PAUP found a shorter solution, by one or two evolutionary steps, than did PHYLIP, and in one occasion PHYLIP found a solution which was one step shorter than the solution found by PAUP. PAUP however, was superior to PHYLIP, in that it generated more equally parsimonious trees. In 95 out of 182 cases (52%), PAUP found more trees that were equally parsimonious than did PHYLIP. The average number of

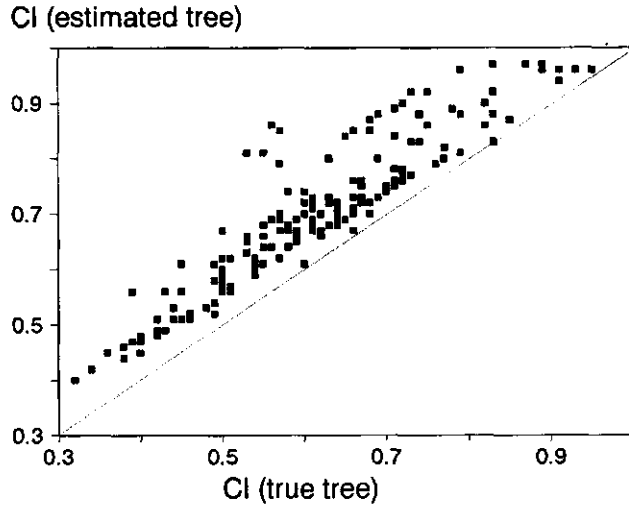


Figure 4.9 Plot of $CI(\text{estimated tree})$ versus $CI(\text{true tree})$ for the PAUP results ($n = 184$).

equally parsimonious trees found by PAUP and PHYLIP was 13.09 and 7.06, respectively ($n = 182$). PAUP was run with the default value (50) for MAXTREE, which sets the maximum number of equally parsimonious trees to be retained, giving PAUP a slight disadvantage. As for the number of equally parsimonious trees, the results of PAUP and PHYLIP cannot easily be compared with those of Hennig86. This is because the latter may also produce multifurcating trees, whereas PAUP and PHYLIP will always give fully resolved solutions, thus possibly leading to internal tree segments of zero length.

If we compare the lengths of the trees estimated by Wagner78 with those estimated by PAUP and PHYLIP, we find that in 68 cases (36%) Wagner78 found a (one) tree of the same length as did PAUP. In all other cases, PAUP found more parsimonious solutions, the greatest difference being 8 evolutionary steps. Also in 68 cases, Wagner78 found trees of equal length with the solutions found by PHYLIP. In one case, Wagner78 found a tree of shorter length (2 steps) than did PHYLIP; in the other cases, PHYLIP found more parsimonious solutions, the greatest difference again being 8 steps. All trees

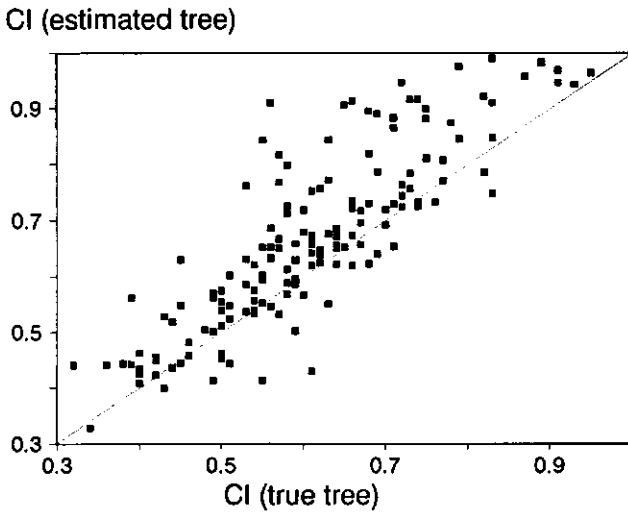


Figure 4.10 Plot of $CI(\text{estimated tree})$ versus $CI(\text{true tree})$ for the CAFA-4 results ($n = 169$).

found by Hennig86, PAUP or PHYLIP, and almost all Wagner78 trees, were more parsimonious than the corresponding true trees; that is, they were shorter. If $CI(\text{true tree})$ equalled 1, the phylogenetic estimation methods always found the true tree with the correct number of steps. However, the true topology occasionally was found by PAUP and PHYLIP (and Hennig86), without $CI(\text{true tree})$ being equal to 1. Among a set of equally parsimonious trees there may be a single tree with the same topology as the corresponding true tree ($CFI = 1$), but with a different number of evolutionary steps, and with $CI < 1$.

To investigate whether a relation exists between the number of equally parsimonious trees estimated by PAUP and PHYLIP and the consistency index of the estimated trees, a correlation analysis was performed. The results of the analysis clearly indicated that such a relation does not exist, the correlation coefficients of CI with the lengths of the PAUP and the PHYLIP trees being 0.0063 ($R^2 = 0.00\%$) and 0.0294 ($R^2 = 0.09\%$) respectively ($n = 186$).

Execution speed may be regarded another aspect of efficiency (see Platnick, 1987, 1989; Luckow & Pimentel, 1985). UPGMA, Wagner78 and PHYLIP were run on a main-frame computer (VAX 8600) and the execution times of their runs were not recorded. However, to give some impression, UPGMA as performed by the CLUSTAN package and Wagner78 were very fast (UPGMA took less than one second CPU-time per single run, Wagner78 less than 0.5 seconds per run), and that PHYLIP was very slow (more than one minute per run). PAUP, CAFCA and Hennig86 were run on a 20-MHz, 80386-based system, equipped with a 80387 math coprocessor, a 80-megabyte hard disk and 640-kilobytes of RAM. The mean execution times of 55 runs of Hennig86, PAUP (the 4 runs with different options taken together!), and CAFCA were 14.07, 113.19 and 1458.93 seconds, respectively, which amounts roughly to 1 : 8 : 100.

To compare and to evaluate the results of different phenetic procedures, Sokal & Sneath (1963) introduced the cophenetic correlation coefficient as an optimality criterion. The cophenetic correlation coefficient measures the amount of information in the similarity matrix, that is reproduced in the dendrogram. A high cophenetic correlation indicates a close agreement between the similarity matrix and the corresponding dendrogram.

A regression analysis of the consensus values (*CFI*) and the cophenetic correlation coefficients (r_{cs}) of the UPGMA phenograms was performed to test whether a relation exists between adequacy and r_{cs} value. The R^2 values presented in table 4.6, indicate that there is no correlation between the two.

Conclusions and discussion

Overall adequacy

As a general conclusion on the adequacy of the numerical taxonomic methods, the parsimony methods PAUP and PHYLIP (and Hennig86) and the UPGMA clustering based on product moment correlations of unstandardized characters produce equally

Adequacy of taxonomic methods

Table 4.6 Average values of *CFI* and the cophenetic correlation coefficients (r_{cs}), standard deviations and minimum and maximum values of r_{cs} and the values of R^2 of regressions of *CFI* on r_{cs} for the four UPGMA methods.

	Methods			
	UPGMA-1	UPGMA-2	UPGMA-3	UPGMA-4
	Euclidian distance		Product moment	
	non-stand	stand	non-stand	stand
Sample size	188	188	188	188
Average <i>CFI</i>	0.73	0.73	0.83	0.80
Average r_{cs}	0.920	0.934	0.960	0.963
Std deviation r_{cs}	0.048	0.039	0.033	0.025
Minimum r_{cs}	0.781	0.826	0.813	0.871
Maximum r_{cs}	0.994	0.992	0.998	0.993
R^2 (%)	3	1	6	9

accurate estimates of the true phylogenetic trees. These methods perform better than do the other UPGMA techniques, CAFCA, and Wagner78, but the differences from the latter method are not really significant.

These results generally conform to the results of Fiala & Sokal (1985). They found only small differences in overall accuracy between UPGMA, using Manhattan distances and Wagner parsimony (Wagner78). The average *CFI* value and the standard error for all their Wagner78 tree estimates, equals 0.728 and 0.005, respectively. In this study, an average value for *CFI* of 0.791 was found, with a standard error of 0.01. The average *CFI* value and standard error for the UPGMA estimates of Fiala & Sokal are 0.733 and 0.005, respectively, the values for UPGMA-3 are 0.832 and 0.008, respectively. The average value of *CFI* for UPGMA using euclidian distances of standardized characters, is 0.73 (see table 4.6).

The compatibility method as provided in the CAFCA package, is clearly inferior to all other phylogenetic methods and phenetic methods investigated.

Sensitivity to tree statistics

The results presented above clearly indicate that there are differences between reconstruction methods in their sensitivity to a number of tree statistics. The sensitivity to tree statistics of the accuracy of the methods tested, shows a consistent pattern. The results of UPGMA clustering are not significantly correlated with any of the statistics of the true tree. This means that the performance of UPGMA is not affected by the (amount of) information about the true evolutionary process present in the data matrix. The accuracy of the trees estimated by Wagner78, PAUP and CAFCA, however, is clearly affected by some properties of the true phylogenetic tree. PAUP and Wagner78 are equally sensitive, whereas the adequacy of CAFCA shows the strongest correlation with the tree statistics.

The consistency index calculated for the true phylogenetic tree proved to be the most important true tree predictor of adequacy. Particular the CAFCA results depend very much on the amount of homoplasy in the original data. If the amount of homoplasy is low ($0.8 < CI < 1.0$), all phylogenetic methods of estimation perform equally well, and somewhat better than does phenetic clustering. In the total absence of homoplasy ($CI = 1.0$), all phylogenetic methods are guaranteed to find the true topology.

The overall effects of the other true tree statistics on the adequacy of the estimation methods seem of minor importance. So in contrast to the findings of Fiala & Sokal (1985), stemminess seems not to be a good estimator of accuracy either. The symmetry of the true tree, whether measured by I_{col} or I_{sack} , seems no good predictor of adequacy. The indices I_{col} and I_{sack} are highly correlated, their correlation coefficient being 0.975. A similar result was found by Shao (1983).

The symmetry of the estimated tree, as measured by I_{sack} (fig. 4.7), turned out not to be a good estimator of accuracy of the estimates. There was some correlation between the symmetry of the true tree and the estimated trees, the correlation coefficient of $I_{sack}(true\ tree)$ with $I_{sack}(estimated\ tree)$ being 0.10 for the CAFCA

results, 0.40 for the Wagner78 results, 0.61 for the PAUP results and 0.71 for the UPGMA-3 results.

The consistency index of the estimated tree turns out to be correlated with the adequacy of the CAFCA and PAUP estimations. Especially for CAFCA, this index proved to be a reasonable estimator of adequacy (fig. 4.8). The consistency index of the estimated tree can be computed in practice and can therefore also be used as a predictor of accuracy of the estimation. Moreover, the consistency index of the true tree is highly correlated with the index calculated for the trees estimated by CAFCA and PAUP.

The trees estimated by PAUP were always "better", that is, shorter and with a higher *CI*, than were the corresponding true trees. Also the trees estimated by CAFCA often had higher *CI* values than the true trees. PAUP and PHYLIP (and Hennig86) always produced trees that were shorter than, or equal in length with the corresponding true trees.

As it is not possible to compute true tree statistics in practice, these cannot be used to evaluate the quality of a method of estimation. The consistency index of the estimated tree to some extent determines the adequacy of the estimation methods. The usefulness of this index lies in the fact that it really can be computed and therefore be used, to some extent, to evaluate the adequacy of phylogenetic estimation methods.

Recommendations

UPGMA using product moment correlations, and the parsimony methods of PAUP, PHYLIP (and Hennig86) are found to be superior to Wagner78 and CAFCA, in adequacy to estimate the true phylogenetic tree. The deficiency in phylogenetic character analysis of the CAFCA package, however, is in contrast to its apparent superior behaviour in the analysis of problems in historical biogeography (van Welzen & Zandee, in preparation).

If evolution is not completely parsimonious, that is, if there is a sufficient amount of homoplasy among the characters analyzed, than the phylogenetic estimation methods tend to find trees that

are too short and too parsimonious. So, when there are reasons to assume that the amount of homoplasy among the characters selected is not too high, one should use a parsimony method in favour of UPGMA clustering. If we do not feel so confident about the original character matrix, we may wish to favour UPGMA clustering of product moment correlations, since this method is not very sensitive to *CI* and to other tree properties, except perhaps for the symmetry of the estimated tree.

If we would find, using a parsimony method, trees with a relatively low consistency index (e.g. $CI < 0.7$) then there would be a serious reason not to accept the estimated tree as a reasonable supported hypothesis. Archie (1989) used 28 original data sets in randomization tests for phylogenetic information in systematic data. These data sets, of course, differ in the number of taxa and in the number of (coded) characters. For 18 out of the 28 (64%) corresponding estimated trees (PAUP), *CI* was less than 0.7. In his comparison of microcomputer parsimony programs, Platnick (1989) makes use of 60 data sets from the literature, 44 (73%) of which have a *CI* less than 0.7.

Because there are only minor differences in adequacy, as expressed by *CFI*, between PAUP and PHYLIP (MIX) (and Hennig86), one cannot recommend any of these parsimony programs as the best. However these programs do differ in the number of equally parsimonious trees that are found, as well as in execution speed. Several detailed studies comparing the efficiency and effectiveness, defined in terms of number of equally parsimonious trees and execution speed, of parsimony programs have been published (e.g. Platnick, 1987, 1989; Luckow & Pimentel 1985). Therefore I can support Platnick's (1989:160) conclusion: "Hennig86 ... should now become the tool of choice for practising systematists."

A critical note is necessary about the use of execution speed as a factor determining the efficiency of cladogram-estimation methods. In the majority of cases, the (worldwide) collecting of members of a monophyletic taxon and the subsequent character analysis will be a time-consuming activity. Therefore it seems to me that it is unimportant, whether the computer program used to

estimate phylogenetic relationships, takes several seconds, minutes, hours or even days, for the analysis. Hence, execution speed cannot really be considered an aspect of efficiency of numerical taxonomic methods.

Since the consistency index of the true tree appeared to be the best predictor of adequacy among the true tree statistics tested, it follows that in practice character analysis should be carried out with great care. It is of paramount importance to select characters that are informative about the evolutionary history of the taxon under study, and to avoid as much as is possible the use of characters that show homoplasy (non-homology). If we lack confidence about the quality of certain characters (uncertain polarity determination, suspected adaptive value), we might better not use them in the analysis at all. However we need to be able to explain these discordant characters; that is, there must be specific reasons why we believe that certain characters have evolved independently of each other. In fact we are dealing here with the delicate issue of character weighting (e.g. Wheeler, 1986; Sharkey, 1989).

Methods differ in their implicit assumptions about the process of evolution, and therefore different methods, analyzing the same character data set, will and are expected to yield different hypotheses about phylogenetic relationships. However different taxa may have evolved under different and largely unknown conditions, with different and unknown consequences for the actual process of evolution. In view of the results of the simulation experiments presented in this paper, one cannot expect a single estimation method, corresponding to a single set of assumptions, to be the best method for all taxa to be examined. Therefore Sober (1988:13) shows too much optimism in stating that "the way to resolve this methodological debate in systematics is to identify clearly the process presuppositions of the main competing methods".

Zusammenfassung

Über die Eignung numerisch-taxonomischer Methoden: Eine vergleichende Untersuchung anhand von Computer-Simulationen

Das Simulationsmodell GENESIS wurde entwickelt, um die Brauchbarkeit verschiedener phylogenetischer und phenetischer Methoden relativ zueinander zu vergleichen. Das Modell wurde in der *Zeitschrift für zoologische Systematik und Evolutionsforschung* (26/1988 und 28/1990) schon früher beschrieben. Es wurde so entworfen, daß es künstliche Datensätze von Gruppen von Arten erzeugt, deren Phylogenie bekannt ist. Diese Datensätze wurden dann mit Hilfe verschiedener Methoden (UPGMA-Cluster-Methode, Wagner-Parsimonie-Analyse und Komponenten-Kompatibilitäts-Analyse) der numerischen Taxonomie einem Test unterworfen. Die Ergebnisse der Analysen konnten dann mit dem realen Stammbaum verglichen werden. Die Güte der Übereinstimmung zwischen dem wahren und dem rekonstruierten Stammbaum wurde als Maß für die Brauchbarkeit der Methode verwendet. Durch Abänderung der Eingabeparameter von GENESIS wurden Ausgabewerte für verschiedene evolutive Abläufe gewonnen, so daß die Güte der Methoden unter diversen evolutiven Bedingungen geschätzt werden konnte. Die Gesamtunterschiede zwischen den verschiedenen Wagner-Parsimonie-Programmen wie PAUP, PHYLIP (MIX) und Hennig86 und dem UPGMA-Cluster-Programm waren ziemlich gering. Insgesamt waren diese Programme etwas geeigneter als die Parsimonie-Methode Wagner78 und die Gruppen-Kompatibilitäts-Methode CAFCA. Die Güte der Stammbäume, die mit Wagner78, PAUP, PHYLIP und CAFCA gewonnen wurden, hängt von einigen besonderen Eigenschaften der Bäume ab, unter denen der Konstistenz-Index besonders wichtig erscheint.

Acknowledgments

I am grateful to Drs R. Zandee, A. de Winter, P. de Vrijer and R. Daamen for comments on earlier versions of the manuscript. Mr J. C. Rigg kindly corrected the English text

References

- Archie, J. W., 1989. A randomization test for phylogenetic information in systematic data. — *Systematic Zoology* 38: 239-252.
- Farris, J. S., 1972. Estimating phylogenetic trees from distance matrices. — *American Naturalist* 106: 645-668.
- Farris, J. S., 1977. On the phenetic approach to vertebrate classification. In: *Major patterns in vertebrate evolution* (M. K. Hecht, P. C. Goody & B. M. Hecht, eds): 823-850. Plenum Press, New York.
- Farris, J. S., 1978. *Wagner78*; manual, documentation and a FORTRAN IV Wagner program.
- Farris, J. S., 1988. *Hennig86, version 1.5*; Hennig86 reference manual and program.
- Felsenstein, J. S., 1987. *PHYLIP; Package for inferring phylogenies, version 3.0*; manual, documentation and several PASCAL programs. University of California Herbarium, Berkeley, California.
- Fiala, K. L. & R. R. Sokal, 1985. Factors determining the accuracy of cladogram estimation: evaluation using computer simulation. — *Evolution* 39: 609-622.
- Funk, V. A., 1983. The value of natural classification. In: *Numerical Taxonomy* (J. Felsenstein, ed.): 18-21. Springer-Verlag, Berlin.
- Geesink, R. & M. Zandee. Phylogeny and information theory: Redundancy Index (*R_I*) as an optimality criterion for phylogenetic trees. (In preparation)
- Heijerman, Th., 1988. GENESIS: a simulation model of phylogeny. 1. The origin and early evolution of character state vectors. — *Zeitschrift für zoologische Systematik und Evolutionsforschung* 26: 409-424.
- Heijerman, TH., 1990. GENESIS: a simulation model of phylogeny. 2. A sensitivity analysis. — *Zeitschrift für zoologische Systematik und Evolutionsforschung* 28: 81-93.
- Kim, J. & M. A. Burgman, 1988. Accuracy of phylogenetic estimation methods under unequal evolutionary rates. — *Evolution* 42: 596-602.
- Kluge, A. G. & J. S. Farris, 1969. Quantitative phyletics and the evolution of anurans. — *Systematic Zoology* 18: 1-32.
- Luckow, M. & R. A. Pimentel, 1985. An empirical comparison of numerical Wagner computer programs. — *Cladistics* 1: 47-66.
- Moss, A. W., 1983. Taxa, taxonomists, and taxonomy. In: *Numerical Taxonomy* (J. Felsenstein, ed.): 72-75. Springer-Verlag, Berlin.
- Platnick, N. I., 1987. An empirical comparison of microcomputer parsimony programs. — *Cladistics* 3: 121-144.
- Platnick, N. I., 1989. An empirical comparison of microcomputer parsimony programs, II. — *Cladistics* 5: 145-161.

- Rohlf, F. J., 1982. Consensus indices for comparing classifications. — *Mathematical Biosciences* 59: 131-144.
- Rohlf, F. J. & M. C. Wooten, 1988. Evaluation of the restricted maximum-likelihood method for estimating phylogenetic trees using simulated allele-frequency data. — *Evolution* 42: 581-595.
- Sackin, M. J., 1972. "Good" and "bad" phenograms. — *Systematic Zoology* 21: 225-226.
- Shao, K., 1983. *Consensus methods in numerical taxonomy*. Ph.D. Dissertation, State University of New York, Stony Brook.
- Sharkey, M. J., 1989. A hypothesis-independent method of character weighting for cladistic analysis. — *Cladistics* 5: 63-86.
- Schuh, R. T. & J. T. Polhemus, 1980. Analysis of taxonomic congruence among morphological, ecological and biogeographic data sets for the Leptopodomorpha (Hemiptera). — *Systematic Zoology* 29: 1-26.
- Sneath, P. H. A., 1983. Philosophy and method in biological classification. In: *Numerical Taxonomy* (J. Felsenstein, ed.): 22-37. Springer-Verlag, Berlin.
- Sneath, P. H. A. & R. R. Sokal, 1973. *Numerical taxonomy. The principles and practice of numerical classification*. W. H. Freeman & Co., San Francisco.
- Sober, E., 1988. *Reconstructing the past. Parsimony, Evolution, and Inference*. The MIT Press, Cambridge, Massachusetts.
- Sokal, R. R., 1983. A phylogenetic analysis of the Caminalcules. I. The data base. — *Systematic Zoology* 32: 159-184.
- Sokal, R. R., 1983. A phylogenetic analysis of the Caminalcules. II. Estimating the true cladogram. — *Systematic Zoology* 32: 185-201.
- Sokal, R. R. & F. J. Rohlf, 1981. Taxonomic congruence in the Leptopodomorpha re-examined. — *Systematic Zoology* 30: 309-325.
- Sokal, R. R. & P. H. A. Sneath, 1963. *Principles of numerical taxonomy*. W. H. Freeman & Co., San Francisco.
- Swofford, D. L., 1985. *PAUP Version 2.4: Phylogenetic Analysis Using Parsimony*; manual, documentation and program. Illinois Natural History Survey, Champaign, Illinois.
- Welzen, P. C. van & M. Zandee. Parsimony versus group/component compatibility in character analysis and historical biogeography. (submitted to *Cladistics*)
- Wheeler, Q. D., 1986. Character weighting and cladistic analysis. — *Systematic Zoology* 35: 102-109.
- Wishart, D., 1986. *CLUSTAN Version 3.2. A cluster analysis package*. Edinburgh University, Program Library Unit.
- Zandee, M., 1988. *C.A.F.C.A.: A Collection of APL Functions for Cladistic Analysis, PC version 1.9*; manual, documentation and program.

5

Adequacy of numerical taxonomic methods

Further experiments using simulated data ¹

"What confidence can we vest in taxonomy, if convergent evolution is such a powerful faker of deceptive resemblances?" Dawkins (1986:269)

Abstract

The adequacy of various phenetic and phylogenetic estimation methods was evaluated using simulated data sets. Two parsimony programs were used to construct maximum parsimony trees (Wagner78 and Hennig86). The CAFCA program was used to perform group-compatibility analysis. Four UPGMA clustering strategies were employed. The simulation model GENESIS was used to generate data sets under different evolutionary conditions. The effects of input parameters and tree properties on the accuracy of the estimated trees were evaluated. UPGMA based on product moment correlations of unstandardized characters appeared to perform best, under all evolutionary conditions tested. The effect of input parameters on the accuracy was not very significant. Among the tree statistics the stemminess of the true tree appeared to be the most important estimator of accuracy.

Keywords: Numerical taxonomy — Phylogeny — Simulation model
— Cladogram estimation

¹ Published as: Heijerman, Th., 1992. Adequacy of numerical taxonomic methods. Further experiments using simulated data. — *Zeitschrift für zoologische Systematik und Evolutionsforschung* 31: 81-97.

Introduction

It is well known that different numerical taxonomic methods produce different results which obviously cannot all be correct. Hence, there is an obvious requirement to test the relative adequacy of the different techniques under various evolutionary conditions. As phylogenies of real organisms are not known, Rohlf et al. (1990:1671) asserted that "Estimates of accuracy (of estimated phylogenies) ... require the use of simulated data, the only kind for which true phylogenies are known". Simulation studies to examine the relative merits of phylogenetic and phenetic methods were indeed carried out before (see Rohlf et al. 1990, and also Heijerman (1992) for short surveys of previous work). In a previous paper (Heijerman, 1992) I presented some results of experiments using data sets of artificial species with known phylogenies, generated by the simulation model GENESIS (Heijerman, 1988, 1990). In this study it was shown that the overall differences in adequacy between Wagner parsimony as performed by PAUP, PHYLIP (MIX) and Hennig86, and UPGMA clustering with product moment correlations of non-standardized characters, were rather small. Wagner parsimony with Wagner78 and group-compatibility with CAFCA were less adequate. Among the tree properties that influence the adequacy, the consistency index of the true tree proved to be the most important one. Rohlf et al. (1990) found maximum parsimony to perform better than UPGMA and the accuracy of the estimated trees was mainly influenced by the 'evolutionary context model' (phyletic, speciation and punctuational).

In the present study results will be presented of further experiments using data sets generated by GENESIS. These data sets were produced under different and more complicated evolutionary conditions than the sets used in the previous study. The relative efficiencies of several phylogenetic and phenetic methods were examined, and it was investigated how the accuracy of estimation methods depends on properties of the evolutionary process as expressed by various tree statistics.

Material and Methods

General design

The simulation model GENESIS can produce character data sets of recent species with known phylogenies. These data sets can be generated under differing evolutionary conditions. GENESIS can be made to simulate evolution according to four major evolutionary scenarios (see next section). The model itself, as well as the four scenarios, were described in detail in Heijerman (1988). The results of a sensitivity analysis were presented in Heijerman (1990). Only a short description of the four scenarios is given here.

In Scenarios 1 and 2, speciation and extinction probabilities are constant over time and the number of extant species increases exponentially with time ('radiation' version of GENESIS). In Scenarios 3 and 4, the number of species fluctuates around an equilibrium number of recent species ('equilibrium' version). In Scenarios 1 and 3, the rate of character evolution is the same in both daughter lineages ('gradualistic' version). In Scenarios 2 and 4 this rate of character change is unequal in the two daughter lineages ('punctuated' version).

The results of experiments conducted with data sets produced by simulations according to Scenario 1 (radiation/gradual) were presented in an earlier paper (Heijerman, 1992). In the current paper the three other scenarios will be treated.

Within each of the scenarios 'experiments' can be carried out. In an 'experiment' the effect is studied of changing the value of one input parameter while keeping the other input parameters constant. By selecting the proper parameter values, GENESIS can be made to produce ideal data sets on which phylogenetic estimation methods are expected to perform excellently. By allowing back-mutations, homoplasous character changes etc., the amount of *noise* contained by the data matrix is increased and as a consequence one may expect the estimation methods to perform increasingly badly.

Chapter 5

Table 5.1 Input parameters of all four scenarios. For each experiment the alternative values of the variable parameter are given. The right part of the table shows the values of the input parameters that remained constant. A "-" means: not applicable in this experiment. In some experiments the probability of character state change (P_+) was adjusted (adj.) to equalize the number of character state changes within the experiment.

exp	variable parameter	P_{ex}	P_+	P_-	P_p	$F_{compchar}$	P_d
Scenario 1, radiation/gradual							
1	$F_{compchar}$ (0/.2/.4/.6/.8/1)	0	0.03	0	-	.	-
2	P_{ana} (0/.1/.2/.3/.4/.5)	0	0.03	.	-	1.0	-
3	P_{ex} (0/.01/.02/.03/.04/.05/.06)	.	0.03	0	-	0.6	-
Scenario 2, radiation/punctuated							
1	P_p (1/4/8/12/16/20)	-	adj.	0	.	0.6	-
2	P_{ex} (.01/.02/.03/.04/.05/.06)	.	0.017	0	20	0.6	-
3	P_{clado} (.10/.20/.30/.40/.50)	0	0.107	.	20	0.6	-
Scenario 3, equilibrium/gradual							
1	P_d (1/3/5/7/9/11)	-	0.05	0	-	0	.
2	P_{ana} (0/.1/.2/.3/.4/.5)	-	0.05	.	-	0.6	5
3	$F_{compchar}$ (0/.2/.4/.6/.8/.9)	-	0.05	0	-	.	5
Scenario 4, equilibrium/punctuated							
1	P_d (1/3/5/7/9/11)	-	0.06	0	20	0.6	.
2	P_p (1/4/8/12/16/20)	-	adj.	0	.	0.6	5
3	P_{clado} (0/.10/.20/.30/.40/.50)	-	0.06	.	20	0.6	5

The data sets and the corresponding true trees were characterized using a number of statistics, and were used as input for several phylogenetic and phenetic estimation methods. The resulting estimated (phylogenetic and phenetic) trees, were compared with the corresponding true trees. The adequacy of the estimation methods was measured by the consensus of the true trees with the estimated ones, and it was investigated if and how the various tree properties affect the adequacy of the methods.

Data sets for analysis

Within the scenarios experiments were carried out by varying the values of the input parameters of GENESIS. At least 10 simulations were run for every single set of input parameter

values. All simulations were run to produce 20 recent species, each with 50 characters. In all experiments the probability of speciation, P_{split} was set at 0.10. In Scenarios 3 and 4 (equilibrium) the equilibrium number of species, N_{equi} was set at 20, and t_{min} , the minimum number of timesteps that must have been passed until the simulation is allowed to stop, was set at 100. The values of the input parameters in the various scenarios are listed in table 5.1.

In Scenario 2 (radiation/punctuated) the values of the following input parameters must be supplied by the user: the probability of extinction, P_{ex} ; the character change probability, P_{clado+} ; the probability of a reversed character state change, P_{clado-} ; the value of the 'punctuation'-parameter, P_p , which determines how much faster evolution proceeds, expressed as the number of character state changes per time unit, in one daughter lineage compared to the other; the number fraction of characters that are fully consistent with the tree, $F_{compchar}$. Three experiments were conducted within this scenario, varying P_p (experiment 1), P_{ex} (experiment 2) and P_{clado-} (experiment 3) respectively. If necessary, the value of P_{clado+} was adjusted to equalize the total number of evolutionary changes on the trees.

Input parameters of Scenario 3 were: the damping parameter, P_d ; P_{ana+} , which is the character change probability; P_{ana-} , the retrogression parameter; N_{equi} , which is the equilibrium number of species; $F_{compchar}$. In the three experiments the effects were studied of varying the values of the damping parameter, P_d (experiment 1), P_{ana-} (experiment 2) and $F_{compchar}$ (experiment 3) respectively. In experiment 3 the value of P_{ana+} had to be adapted to equalize the total number of evolutionary steps.

In Scenario 4 the parameters P_d ; P_p ; P_{clado+} ; P_{clado-} ; N_{equi} ; $F_{compchar}$ and t_{min} must be specified. In the experiments the values of P_d (experiment 1), P_p (experiment 2) and P_{clado-} (experiment 3) were varied.

Tree statistics

A number of tree statistics were calculated to characterize the results of each simulation. The reversal index, R_1 , as defined by Sokal (1983:170) measures the amount of reversals and repeats in character state changes. Sokal (1983:171) also defined the dendritic index (DI), which measures the amount of shared evolution. DI values range from 0 to 1, in the absence of parallelisms and reversals. The consistency index (CI), (Kluge & Farris, 1969; see also Sokal 1983:172,173) was also calculated. The index CI also ranges from 1 (no homoplasy) to 0 (maximum homoplasy). However, Farris (1989) pointed out that the consistency index can be no less than m/g , where m is the minimum amount of change on the tree and g denotes the greatest amount of change (number of steps in an unresolved bush). To measure the adequacy of the characters in resolving the cladogram, I_{adeq} was calculated. I_{adeq} is defined as the ratio $N_{pseu} / (2S - 3)$, where N_{pseu} stands for the number of binary coded characters and S denotes the number of species (cf. Sokal, 1983, and Sokal & Shao, 1985). As a measure of the symmetry of the tree, an index, I_{col} , proposed by Colless (1982:103) was used. This index was named Colless2 in Sokal (1983), S/b in Rohlf et al., (1990) and $I_C(1)$ in Shao & Sokal (1990), where it is referred to as an index that measures imbalance rather than tree asymmetry. I_{col} may range from 0 (perfect symmetry c.q. least imbalanced) to 1 (perfect asymmetry c.q. most imbalanced). As Shao & Sokal (1990) pointed out, I_{col} was defined for binary trees only, and they therefore modified the index for non-binary trees. In the present study the index as originally defined by Colless (1982) was used, thus inevitably disregarding any multifurcations in a tree. Two measures of stemminess were calculated: I_{stemFS} and I_{stem} . The first index refers to the stemminess as originally defined by Fiala & Sokal (1985), whereas in I_{stem} the length of segments of a tree is measured by the number of evolutionary steps, rather than in time units (Heijerman, 1988). Rohlf et al. (1990) modified the stemminess index as originally defined. Their newly defined index, which is referred to as the non-cumulative stemminess index (St_N ;

Rohlf et al. 1990:1672) would "lower the correlation between stemminess and tree imbalance" and, "unlike the cumulative stemminess index of Fiala & Sokal (1985), it does not give higher weights to groups nearer the tips of the tree". In order to compare with the results of Heijerman (1992), I_{stem} and I_{stemFS} were employed again, instead of St_N . Also Sackin's index of symmetry, I_{sack} , was employed (Sackin, 1972; see also Shao 1983: SI_a). Shao & Sokal (1990) argued that this index also measures imbalance rather than (a)symmetry and they used the notation $I_S(1)$ to refer to it. I_{sack} was computed for the true trees, as well as for the estimated trees.

The cophenetic correlation coefficient, r_{cs} , (Sokal & Sneath, 1962), which may be used as an optimality criterion for phenetic procedures, was calculated for the phenograms.

Estimation methods

UPGMA (Unweighted Pair-Group Methods using Arithmetic Averages; Sneath & Sokal, 1973) analysis was performed, using the CLUSTAN package version 3.2 (Wishart, 1982) to construct phenograms, based on matrices of squared euclidian distances and product moment correlations and calculated both from standardized as well as from non-standardized data sets.

Results for Scenario 1 (Heijerman, 1992) showed that differences in adequacy between Wagner parsimony as performed by PAUP, version 2.4 (Swofford, 1985), PHYLIP (MIX), version 3.0 (Felsenstein, 1987) and Hennig86 were very small. Benchmark comparisons showed that Hennig86 performs "best" in terms of execution speed and the number of equally parsimonious trees that are found (e.g. Platnick, 1987, 1989; Luckow & Pimentel, 1985; Sanderson, 1990; Heijerman, 1992). Therefore minimum length trees were constructed only using the 'simple' Distance Wagner procedure of Wagner78 (Farris, 1978) and Hennig86 Version 1.5 (Farris, 1988). The estimation methods were applied using the same procedures and options as described in Heijerman (1992).

Wagner78 was run using the HOM option only. The true ancestor (all characters are in state "0") was added as the first taxon in the data matrix. As the results appeared to be dependent on the order of species, Wagner78 was run ten times with different orderings of the species in the input matrix (see also Heijerman, 1992). From the ten solutions one of the shortest trees was selected for further analysis. Hennig86 was used with the mhennig* and bb options, which was found by Platnick (1989) to perform best on the kind of data sets as used in this study. All characters were treated as ordered characters (default option).

Compatibility analysis was performed using CAFCA (a Collection of APL Functions for Cladistic Analysis, Zandee, 1988). A primary analysis was run, using the default cladon option PMS (partial monothetic sets), and treating all characters as ordered. CAFCA offers six selection criteria for cladograms (Zandee, 1988:14): total number of homoplasous events (criterion 1); total number of single origins (criterion 2); homoplasy - support (criterion 3); total number of state changes (criterion 4); the redundancy index (criterion 5) and the consistency index (criterion 6). CAFCA was tested using sets of trees selected by both criterion 3 (CAFCA-3) and 4 (CAFCA-4), and by using all trees generated (CAFCA-tot).

Comparison of trees

To compare trees Colless' consensus fork index (*CFI*) was used (Colless, 1980). *CFI* measures the difference between the strict consensus tree and a bush, and is defined as the number of branching points (ignoring the basal one), normalized by dividing by $n - 2$, where n is the number of OTUs. *CFI* is identical with the consensus index (CI_c) of Rohlf, 1982 (see also Shao, 1983). *CFI* ranges from 1, if the strict consensus tree between the true tree and the estimated tree is fully resolved, and 0, if the consensus tree is a bush. In fact, *CFI* is used as a quality measure of the adequacy of the estimation methods.

UPGMA produced a single phenogram for each of the four analyses. The accuracy of each phenogram was measured by its

consensus (*CFI*) with the true tree. Hennig86 often generated multiple equally parsimonious trees. A maximum of 50 trees was used for further analysis. The consensus between all trees produced by Hennig86 and the true tree was computed, and for further analysis the mean *CFI* was used. Also CAFCA often generated more than 1 tree. Again a maximum of 50 trees was used, and for the current analysis the mean *CFI* was computed separately for the set containing all trees generated, for the set containing the trees selected by criterion 3 (homoplasy - support) and for the set with the trees selected according to criterion 4 (total number of state changes).

Results

Comparison of the overall adequacy of the estimation methods

The overall adequacy of a method was computed as the mean consensus value over all tree comparisons within the separate experiments. Table 5.2 shows the mean *CFI* values between estimated trees and the true trees for each of the three evolutionary scenarios. There are significant overall differences between methods. UPGMA using product moment correlations of unstandardized characters (UPGMA-3) was found to perform consistently better than all other estimation methods. UPGMA using euclidian distances of unstandardized characters (UPGMA-1) and UPGMA based on product moment correlations of standardized characters (UPGMA-4) performed second best. In all but one of the experiments the trees produced by CAFCA have the lowest accuracy values. The adequacy of the remaining UPGMA strategies, Hennig86 and Wagner78 differed between the various experiments, but especially within scenarios, these differences are very small. The differences between CAFCA-3, CAFCA-4 and CAFCA-tot were similarly small, although CAFCA-4 (minimum number of state changes) tended to perform slightly better. The overall accuracy of trees generated within Scenario 2 seems to be somewhat higher compared with the other two

Chapter 5

Table 5.2 Mean values and standard deviations (between brackets) of the strict consensus indices (*CFI*) as indicators of the adequacy of the estimation methods evaluated, given for each experiment separately and for the three experiments within a scenario taken together (exp. 123). *n* denotes the number of tree comparisons.

estimation method	source of variation			
	exp. 1	exp. 2	exp. 3	exp. 123
	Scenario 2, radiation/punctuated			
	$P_p, n \geq 59$	$P_{ex'}, n \geq 74$	$P_{clado'}, n \geq 59$	$n \geq 172$
UPGMA-1	0.64 (0.14)	0.61 (0.12)	0.60 (0.10)	0.62 (0.12)
UPGMA-2	0.62 (0.14)	0.57 (0.14)	0.61 (0.11)	0.61 (0.13)
UPGMA-3	0.88 (0.07)	0.81 (0.11)	0.78 (0.10)	0.82 (0.11)
UPGMA-4	0.79 (0.19)	0.70 (0.14)	0.77 (0.10)	0.75 (0.12)
Wagner78	0.68 (0.15)	0.57 (0.10)	0.63 (0.13)	0.63 (0.13)
Hennig86	0.71 (0.15)	0.56 (0.11)	0.64 (0.12)	0.64 (0.14)
CAFCA-3	0.67 (0.15)	0.48 (0.20)	0.49 (0.15)	0.53 (0.20)
CAFCA-4	0.67 (0.15)	0.48 (0.19)	0.50 (0.14)	0.54 (0.19)
CAFCA-tot	0.64 (0.15)	0.46 (0.19)	0.47 (0.14)	0.51 (0.19)
	Scenario 3, equilibrium/gradual			
	$P_d, n \geq 66$	$P_{ana-}, n \geq 60$	$F_{compchar}, n \geq 66$	$n \geq 192$
UPGMA-1	0.72 (0.13)	0.63 (0.12)	0.68 (0.15)	0.68 (0.14)
UPGMA-2	0.64 (0.14)	0.58 (0.14)	0.64 (0.12)	0.62 (0.13)
UPGMA-3	0.74 (0.12)	0.66 (0.13)	0.71 (0.15)	0.71 (0.14)
UPGMA-4	0.67 (0.13)	0.62 (0.15)	0.67 (0.14)	0.65 (0.14)
Wagner78	0.66 (0.14)	0.51 (0.14)	0.58 (0.16)	0.59 (0.16)
Hennig86	0.69 (0.12)	0.56 (0.14)	0.63 (0.15)	0.63 (0.15)
CAFCA-3	0.43 (0.13)	0.30 (0.11)	0.38 (0.13)	0.37 (0.13)
CAFCA-4	0.45 (0.14)	0.31 (0.11)	0.39 (0.15)	0.39 (0.14)
CAFCA-tot	0.40 (0.13)	0.29 (0.10)	0.36 (0.13)	0.35 (0.13)
	Scenario 4, equilibrium/punctuated			
	$P_d, n \geq 60$	$P_p, n \geq 60$	$P_{clado'}, n \geq 60$	$n \geq 180$
UPGMA-1	0.61 (0.13)	0.66 (0.11)	0.64 (0.11)	0.64 (0.12)
UPGMA-2	0.57 (0.15)	0.68 (0.14)	0.63 (0.14)	0.62 (0.15)
UPGMA-3	0.69 (0.12)	0.73 (0.11)	0.69 (0.12)	0.70 (0.12)
UPGMA-4	0.63 (0.13)	0.69 (0.13)	0.65 (0.14)	0.65 (0.13)
Wagner78	0.54 (0.14)	0.60 (0.13)	0.62 (0.13)	0.59 (0.14)
Hennig86	0.58 (0.12)	0.64 (0.12)	0.65 (0.12)	0.62 (0.12)
CAFCA-3	0.32 (0.32)	0.35 (0.13)	0.35 (0.12)	0.34 (0.13)
CAFCA-4	0.34 (0.13)	0.36 (0.13)	0.35 (0.13)	0.35 (0.13)
CAFCA-tot	0.30 (0.12)	0.33 (0.12)	0.32 (0.11)	0.32 (0.12)

scenarios. However, no efforts were made to make experiments between scenarios comparable with each other. This could have been done by equalizing the total number of character state changes on the true tree between scenarios. As a consequence, runs within the equilibrium scenarios show higher values for I_{adeq} compared with the radiation scenarios. Differences in the results between scenarios may therefore not easily be explained.

Effects of input parameters and scenario on tree statistics

The effects of changes in the values of input parameters on tree statistics were analyzed by an analysis of variance. The results seem largely in agreement with the results of the sensitivity analysis (Heijerman, 1990) and will not be presented here in detail.

In Scenario 2, experiment 3, the retrogression probability effected R_1 ; R_1 increased if more reversals were made to occur. Increasing the probability of reversals and extinction resulted in somewhat lower CI -values in experiments 2 and 3. In experiment 1 there was only a small, though statistically significant, effect of P_p on I_{stem} . In Scenario 3, $F_{compchar}$ appeared to strongly affect both CI and DI , but increasing $F_{compchar}$ also caused I_{adeq} to increase in experiment 3. Also P_{ana-} affected I_{adeq} and of course R_1 . DI is affected, in experiment 1, by P_d . In Scenario 4 the only significant effect was of P_{clado-} on R_1 , in experiment 3.

Summarizing, only two input parameters, the retrogression probability and the number fraction of compatible characters, showed interaction with one or more tree statistics, indicating that variation in most tree statistics mainly resulted from mere chance. In the following sections the effects of both input parameters and properties and/or statistics of both the estimated and the true trees on the accuracy of estimated trees will be investigated.

Table 5.3 presents the mean values and the corresponding coefficients of variation of the true tree statistics for the three experiments taken separately and pooled.

Table 5.3 Mean values of tree statistics and coefficients of variation (between brackets).

tree statistic	source of variation							
	exp. 1		exp. 2		exp. 3		exp. 123	
	Scenario 2, radiation/punctuated							
	P_{pr} , $n=62$		P_{ex} , $n=82$		P_{clado} , $n=69$		$n=193$	
R_1	1	(0)	1	(0)	1.26	(13.2)	1.09	(14.5)
DI	0.84	(5.5)	0.85	(6.5)	0.81	(5.5)	0.84	(6.3)
CI	0.59	(6.8)	0.53	(12.2)	0.53	(7.6)	0.54	(10.7)
I_{col}	0.21	(30.2)	0.20	(37.7)	0.20	(32.5)	0.20	(34.1)
I_{adeq}	3.44	(10.1)	3.62	(15.8)	3.16	(9.9)	3.41	(14.4)
I_{stem}	0.21	(12.8)	0.22	(15.2)	0.20	(13.9)	0.21	(14.8)
I_{stemFS}	0.33	(18.2)	0.36	(20.1)	0.34	(21.1)	0.35	(19.6)
I_{sack}	101.6	(7.5)	100.9	(8.6)	101.1	(7.4)	101.2	(7.9)
	Scenario 3, equilibrium/gradual							
	P_{dr} , $n=66$		P_{ans} , $n=61$		$F_{compchar}$, $n=68$		$n=195$	
R_1	1	(0)	1.14	(7.3)	1	(0)	1.04	(19.2)
DI	0.88	(3.7)	0.91	(2.9)	0.92	(4.1)	0.90	(4.2)
CI	0.59	(9.0)	0.64	(8.5)	0.63	(13.6)	0.62	(11.4)
I_{col}	0.23	(33.4)	0.20	(33.2)	0.21	(32.6)	0.21	(33.5)
I_{adeq}	2.04	(10.0)	1.76	(6.3)	2.26	(19.1)	2.03	(17.2)
I_{stem}	0.35	(23.0)	0.35	(20.9)	0.38	(19.0)	0.36	(21.2)
I_{stemFS}	0.42	(14.1)	0.42	(11.7)	0.43	(11.8)	0.43	(12.6)
I_{sack}	105.2	(8.7)	101.0	(8.0)	102.7	(7.9)	103.0	(8.4)
	Scenario 4, equilibrium/punctuated							
	P_{dr} , $n=60$		P_{pr} , $n=60$		P_{clado} , $n=60$		$n=180$	
R_1	1	(0)	1	(0)	1.24	(17.1)	1.08	(15.5)
DI	0.88	(4.1)	0.88	(3.7)	0.87	(4.7)	0.88	(4.2)
CI	0.55	(8.1)	0.54	(9.2)	0.52	(9.7)	0.54	(9.3)
I_{col}	0.22	(34.3)	0.22	(26.9)	0.21	(33.7)	0.22	(31.7)
I_{adeq}	2.20	(9.5)	2.30	(10.6)	2.24	(8.4)	2.25	(9.7)
I_{stem}	0.28	(20.8)	0.31	(18.9)	0.30	(18.2)	0.30	(19.8)
I_{stemFS}	0.43	(12.8)	0.42	(12.5)	0.43	(13.7)	0.43	(13.0)
I_{sack}	103.6	(8.5)	103.1	(7.0)	102.6	(8.3)	103.1	(7.9)

The indices I_{stem} , I_{stemFS} and especially I_{col} appear to show high values for the coefficient of variation. Since these topology measures were not significantly affected by any of the input parameters, this variation results from chance only.

Adequacy of taxonomic methods; further experiments

Table 5.4 Effects of input parameters on the accuracy of estimated trees. Accuracy is measured by *CFI*. Relative significance of effects is tested by analysis of variance. *F*-values are presented. The last column presents R^2 values of the model used. An asterisk indicates significance probability values < 0.01 .

Scenario 2, radiation/punctuated $n \geq 172$					
estimation method	P_p	P_{ex}	P_{clado}	R^2 -Model	
UPGMA-1	7.1 *	2.4	1.3	0.21	*
UPGMA-2	6.2	1.1	1.1	0.21	*
UPGMA-3	1.2	1.4	2.0	0.25	*
UPGMA-4	1.4	3.7 *	0.8	0.26	*
Wagner78	17.1 *	1.6	2.7	0.43	*
Hennig86	18.8 *	2.2	2.5	0.53	*
CAFCA-3	5.8 *	8.6 *	3.1	0.51	*
CAFCA-4	6.7 *	9.0 *	3.0	0.52	*
CAFCA-tot	6.2 *	9.2 *	3.5	0.52	*

Scenario 3, equilibrium/gradual $n \geq 192$					
	P_d	P_{ana}	F_{compar}	R^2 -Model	
UPGMA-1	3.6 *	3.1	7.3 *	0.28	*
UPGMA-2	1.1	2.0	2.4	0.13	
UPGMA-3	3.0	2.9	8.3 *	0.28	*
UPGMA-4	1.6	1.2	4.5 *	0.17	*
Wagner78	7.0 *	7.1 *	14.8 *	0.45	*
Hennig86	5.2 *	5.6 *	14.8 *	0.42	*
CAFCA-3	3.8 *	8.1 *	10.0 *	0.38	*
CAFCA-4	4.4 *	8.4 *	12.6 *	0.41	*
CAFCA-tot	2.8	6.4 *	7.9 *	0.33	*

Scenario 4, equilibrium/punctuated $n \geq 180$					
	P_d	P_p	P_{clado}	R^2 -Model	
UPGMA-1	2.1	2.7	1.3	0.16	
UPGMA-2	3.2 *	4.4 *	1.1	0.21	*
UPGMA-3	2.6	2.0	2.9	0.18	*
UPGMA-4	1.6	3.4 *	3.2 *	0.20	*
Wagner78	1.4	1.2 *	1.3	0.11	
Hennig86	2.7	3.1	1.7	0.17	*
CAFCA-3	1.9	3.0 *	1.0	0.14	
CAFCA-4	1.4	2.6	1.7	0.14	
CAFCA-tot	1.5	2.9	1.6	0.14	

Effects of input parameters on accuracy

The relative effects of the input parameters on the adequacy of estimation methods was investigated by analysis of variance. The results for the three scenarios are presented in table 5.4. One

might expect that increasing the probability of reversed character state changes, the probability of extinction or the value of the punctuation parameter, as well as decreasing the value of the number fraction of compatible characters, would lower the accuracy of the estimates.

The retrogression parameter (P_{clado-}), in Scenario 2, does not interact with the adequacy of any of the estimation methods. The extinction probability mainly affects the accuracy of the trees produced by CAFCA; the higher the extinction probability the less accurate the estimations. The major effect is due to the 'punctuation' parameter, P_p , which clearly interacts with the accuracy of the trees produced by Wagner78 and Hennig86, but also with the UPGMA-1 and CAFCA trees.

The results for Scenario 3 show that increasing the value of the damping parameter, P_d , decreases the accuracy of the Wagner78 and Hennig86 trees, and to a lesser extent also the accuracy of the CAFCA estimates. The retrogression parameter, P_{ana-} and $F_{compchar}$ interact with the adequacy of all phylogenetic estimation methods. This indicates that trees with a higher amount of homoplasy and/or reversals are more difficult to estimate accurately.

In Scenario 4 there are only small effects, though some of them are statistically significant.

The results, in Scenarios 2 and 3, of the UPGMA analysis are hardly affected by the input parameters, while major interactions are in the CAFCA and especially in the Wagner78 and Hennig86 results, as can be concluded from the R^2 values presented in table 5.4. In general, the effects of input parameters on the adequacy of the estimation methods seem rather small.

Effects of tree statistics on accuracy

As DI and CI are both measures of the amount of homoplasy in the data one would expect these indices to be positively correlated with CFI . As high values of $R1$ indicate a high amount of reversals, this index may be expected to be negatively correlated with CFI . High values for stemminess, described by

Adequacy of taxonomic methods; further experiments

Table 5.5 Effects of tree statistics on the accuracy of estimated trees, as measured by *CFI*, in Scenario 2 (radiation/punctuated). R^2 -values (%) of the regression analysis of *CFI* on tree statistics are given. An asterisk denotes values with significance probabilities < 0.01 . $F_{compchar}$ equals 0.6 in all experiments.

estimation method	Tree statistics							
	R_1	DI	CI	I_{col}	I_{adeq}	I_{stem}	I_{stemFS}	I_{sack}
	Exp. 1, source of variation: $P_p, P_{ex} = 0, P_{clado} = 0$							
UPGMA-1	-	5	2	0	1	25 *	0	0
UPGMA-2	-	1	0	2	5	16 *	4	4
UPGMA-3	-	1	0	1	2	17 *	1	0
UPGMA-4	-	1	0	12 *	0	11 *	2	8
Wagner78	-	9	9	9	0	25 *	4	8
Hennig86	-	8	6	12 *	2	26 *	7	11
CAFCA-3	-	10	7	5	0	20 *	3	4
CAFCA-4	-	7	5	4	0	18 *	5	3
CAFCA-tot	-	7	7	5	0	17 *	3	3
Exp. 2, source of variation: $P_{ex}, P_p = 20, P_{clado} = 0$								
UPGMA-1	-	4	3	0	11 *	14 *	2	0
UPGMA-2	-	1	4	0	11 *	7	3	1
UPGMA-3	-	6	11 *	4	2	7	3	2
UPGMA-4	-	0	19 *	11 *	7	12 *	0	10 *
Wagner78	-	1	17 *	2	8	15 *	0	2
Hennig86	-	0	26 *	4	11 *	18 *	0	3
CAFCA-3	-	1	45 *	1	2	1	1	1
CAFCA-4	-	2	43 *	0	1	1	1	0
CAFCA-tot	-	2	43 *	0	2	0	2	0
Exp. 3: source of variation: $P_{clado}, P_p = 20, P_{ex} = 0$								
UPGMA-1	1	0	0	4	2	7	1	5
UPGMA-2	1	1	2	0	6	8	0	0
UPGMA-3	14 *	0	10 *	5	6	4	2	7
UPGMA-4	1	0	2	17 *	1	3	0	18 *
Wagner78	3	2	1	1	1	5	2	2
Hennig86	1	0	0	0	0	13 *	12 *	0
CAFCA-3	17 *	4	34 *	4	17 *	7	0	2
CAFCA-4	18 *	10	34 *	4	21	6	0	2
CAFCA-tot	23 *	8	38 *	5	21	5	0	2

Rohlf et al. (1990) as a "measure of the average distinctness of all the taxonomic subsets on a tree", and of course high I_{adeq} -values, may be expected to result in high *CFI*-values. I_{col} and I_{sack} , measures of asymmetry or imbalance, can be expected to affect accuracy in a negative way.

The results of the experiments of Scenario 2 are presented in table 5.5. In experiment 1 we find a significant positive correlation of the stemminess as measured by I_{stem} with CFI in all methods. In two cases also the symmetry as measured by I_{col} appears to affect CFI . From experiment 2 we may conclude that the adequacy of the majority of the methods increases more or less with CI , while for some methods we find a positive correlation with I_{stem} . The results of experiment 3 are not very consistent. The accuracy of the CAFCA estimates seem most sensitive to some of the tree statistics.

The results for Scenario 3 are presented in table 5.6. The results of experiment 1 show that all tree topology measures are correlated with the accuracy of the estimates of all methods, while there is no effect of CI and DI . I_{stem} seems to be the most important factor determining the accuracy of the estimation methods in experiments 2 and 3. The majority of the other tree statistics show no correlation with CFI whatsoever.

Table 5.7 shows the results of the regression analysis for the experiments within Scenario 4. A major result is that in all experiments a significant positive correlation was found between the stemminess as measured by I_{stem} , but not by I_{stemFS} , on the accuracy of the results of all estimation methods. In experiment 3, I_{adeq} appears to interact with the accuracy of some methods, especially CAFCA.

Effects of properties of estimated trees on the adequacy of estimation methods.

The results of the experiments that were conducted within Scenario 1 showed that in some cases there is some correlation between the symmetry of the estimated tree, I_{sack} , and the accuracy of the estimates (R^2 values ranging from 1 up to 29 %) and also between the consistency index of the estimated tree and the accuracy (R^2 values ranging from 13 up to 85 %) (Heijerman, 1992).

Adequacy of taxonomic methods; further experiments

Table 5.6 Effects of tree statistics on the accuracy of estimated trees, as measured by *CFI*, in Scenario 3 (equilibrium/gradual). R^2 -values (%) of the regression analysis of *CFI* on tree statistics are given. An asterisk denotes values with significance probabilities < 0.01 .

estimation method	Tree statistics							
	R_1	DI	CI	I_{col}	I_{adeq}	I_{stem}	I_{stemFS}	I_{sack}
Exp. 1, source of variation: $P_d, P_{ana} = 0, F_{compchar} = 0$								
UPGMA-1	-	1	1	16 *	1	35 *	19 *	15 *
UPGMA-2	-	0	3	7	7	28 *	17 *	7 *
UPGMA-3	-	1	1	17 *	1	30 *	19 *	15 *
UPGMA-4	-	0	1	33 *	0	29 *	20 *	32 *
Wagner78	-	3	0	16 *	3	29 *	25 *	15 *
Hennig86	-	1	1	11 *	2	29 *	12 *	10 *
CAFCA-3	-	0	0	12 *	0	15 *	18 *	11 *
CAFCA-4	-	0	1	12 *	0	19 *	23 *	11 *
CAFCA-tot	-	1	0	10 *	0	24 *	24 *	9
Exp. 2, source of variation: $P_{ana}, P_d = 5, F_{compchar} = 0.6$								
UPGMA-1	0	6	4	3	5	33 *	3	3
UPGMA-2	0	4	1	3	6	42 *	13 *	3
UPGMA-3	0	4	2	5	8	34 *	1	4
UPGMA-4	0	5	2	13 *	5	37 *	4	16 *
Wagner78	0	8	2	6	9	45 *	7	5
Hennig86	0	9	4	6	3	32 *	2	6
CAFCA-3	3	0	1	4	10	22 *	20 *	3
CAFCA-4	4	1	0	5	11 *	28 *	24 *	4
CAFCA-tot	4	0	0	6	10	28 *	19 *	7
Exp. 3, source of variation: $F_{compchar}, P_d = 5, P_{ana} = 0$								
UPGMA-1	-	6	2	8	2	7	4	7
UPGMA-2	-	3	2	2	0	20 *	4	1
UPGMA-3	-	13 *	7	3	2	6	3	2
UPGMA-4	-	6	3	12 *	2	11 *	3	12 *
Wagner78	-	5	2	10 *	1	9	3	11 *
Hennig86	-	7	3	9	1	7	4	8
CAFCA-3	-	0	2	5	9	13 *	3	4
CAFCA-4	-	0	2	8	7	11 *	4	7
CAFCA-tot	-	1	4	7	8	11 *	4	6

In Scenarios 2, 3 and 4, the symmetry of the estimated tree, $I_{adeq}(\text{estimated tree})$, is positively correlated with the accuracy of the estimates in a number of cases, but there seems to be no consistent pattern (table 5.8).

Chapter 5

Table 5.7 Effects of tree statistics on the accuracy of estimated trees, as measured by *CFI*, in Scenario 4 (equilibrium/punctuated). R^2 -values (%) of the regression analysis of *CFI* on tree statistics are given. An asterisk denotes values with significance probabilities < 0.01 . $F_{compchar}$ equals 0.6 in all experiments.

estimation method	Tree statistics							
	R_1	DI	CI	l_{col}	l_{adeq}	l_{stem}	l_{stemFS}	l_{sack}
Exp. 1, source of variation: $P_d, P_p = 20, P_{clado} = 0$								
UPGMA-1	-	0	2	18 *	14 *	31 *	1	14 *
UPGMA-2	-	1	3	8	12 *	31 *	0	6
UPGMA-3	-	7	5	16 *	11 *	35 *	2	16 *
UPGMA-4	-	1	2	17 *	7	31 *	1	15 *
Wagner78	-	1	3	2	1	15 *	0	3
Hennig86	-	2	2	2	5	14 *	0	2
CAFCA-3	-	4	1	2	6	11 *	2	1
CAFCA-4	-	4	0	1	7	14 *	3	1
CAFCA-tot	-	2	1	2	6	11 *	3	1
Exp. 2, source of variation: $P_p, P_d = 5, P_{clado} = 0$								
UPGMA-1	-	2	5	8	0	23 *	14 *	7
UPGMA-2	-	0	2	1	4	15 *	10	1
UPGMA-3	-	6	5	6	0	19 *	6	6
UPGMA-4	-	5	3	19 *	0	15 *	15 *	16 *
Wagner78	-	0	0	5	3	19 *	9	4
Hennig86	-	0	0	9	1	19 *	4	10
CAFCA-3	-	1	11	2	3	16 *	1	1
CAFCA-4	-	1	7	2	2	14 *	2	1
CAFCA-tot	-	0	7	1	2	13 *	1	1
Exp. 3, source of variation: $P_{clado}, P_d = 5, P_p = 20$								
UPGMA-1	4	0	8	9	3	15 *	2	11
UPGMA-2	7	1	0	2	15 *	19 *	1	3
UPGMA-3	7	0	8	14 *	18 *	17 *	5	14 *
UPGMA-4	6	2	2	11 *	9	25 *	2	13 *
Wagner78	2	0	0	1	8	12 *	0	1
Hennig86	0	1	0	3	10	15 *	0	2
CAFCA-3	0	0	1	4	25 *	19 *	3	3
CAFCA-4	0	0	1	5	21 *	18 *	5	5
CAFCA-tot	0	0	2	6	18 *	16 *	3	5

The accuracy of the estimates of Hennig86 and CAFCA-3 increased significantly with CI (estimated tree) in some of the experiments (table 5.9). As an alternative of the consistency index, Farris (1989) introduced the retention index (r), which was already in use in Hennig86. This index measures the amount of homoplasy related to the possible amount of homoplasy. The R^2

Adequacy of taxonomic methods; further experiments

Table 5.8 R^2 values of the regression analysis of the adequacy (CFI) on the symmetry of the estimated tree as measured by I_{sack} . An asterisk denotes values with significance probabilities < 0.01 .

	Scenario 2			Scenario 3			Scenario 4		
	exp.1 P_p $n \geq 59$	exp.2 P_{ex} $n \geq 74$	exp.3 P_{clado} $n \geq 59$	exp1. P_d $n \geq 66$	exp.2 P_{ana} $n \geq 60$	exp.3 $F_{compchar}$ $n = 66$	exp.1 P_d $n \geq 60$	exp.2 P_p $n = 60$	exp.3 P_{clado} $n = 60$
UPGMA-1	40 *	14 *	21 *	23 *	0	7	17 *	12 *	10
UPGMA-2	4	7	3	34 *	14 *	4	35 *	14 *	22 *
UPGMA-3	0	3	6	13 *	0	1	10	0	11
UPGMA-4	2	0	1	8	2	2	11 *	11	6
Wagner78	19 *	32 *	29 *	3	0	2	4	3	4
Hennig86	15 *	28 *	10	1	1	0	3	8	1
CAFCA-3	5	12 *	15 *	17 *	0	19 *	3	2	4
CAFCA-4	3	15 *	6	13 *	1	14 *	2	0	4
CAFCA-tot	3	12 *	7	25 *	6	22 *	7	4	1

Table 5.9 R^2 values of the regression analysis of the adequacy (CFI) on the consistency index of the estimated tree, $CI(\text{true tree})$. An asterisk denotes values with significance probabilities < 0.01 .

	Scenario 2			Scenario 3			Scenario 4		
	exp.1 P_p $n \geq 59$	exp.2 P_{ex} $n \geq 74$	exp.3 P_{clado} $n \geq 59$	exp1. P_d $n \geq 66$	exp.2 P_{ana} $n \geq 60$	exp.3 $F_{compchar}$ $n = 66$	exp.1 P_d $n \geq 60$	exp.2 P_p $n = 60$	exp.3 P_{clado} $n = 60$
Hennig86	0	15 *	0	5	2	14 *	9	5	2
CAFCA-3	9	53 *	34 *	13 *	14 *	8	19 *	33 *	16 *

values obtained by using the retention index were almost identical to those obtained for the consistency index. Table 5.10 presents the mean values of $CI(\text{estimated tree})$; in all experiments in all scenarios Hennig86 trees have higher CI -values than CAFCA-4 trees. There are small differences between and within scenarios for both methods. Mean values of CI of the trees as estimated by Hennig86 and CAFCA-3 in Scenarios 2 and 3, are higher than the mean values of CI of the corresponding true trees (compare tables 5.3 and 5.10).

From table 5.11 it may be concluded that the cophenetic correlation coefficient of the UPGMA estimates does not really effect their accuracy. Mean values of cophenetic correlation

Chapter 5

Table 5.10 Mean values of the consistency index of the estimated trees as measured by *CI*.

estimation method	source of variation			
	exp. 1	exp. 2	exp. 3	exp. 123
	Scenario 2, radiation/punctuated			
	P_{pr} , $n \geq 59$	P_{ext} , $n \geq 74$	P_{clado} , $n \geq 59$	$n \geq 172$
Hennig86	0.64 (0.03)	0.61 (0.06)	0.55 (0.06)	0.59 (0.07)
CAFCA-4	0.61 (0.05)	0.54 (0.07)	0.52 (0.06)	0.55 (0.07)
	Scenario 3, equilibrium/gradual			
	P_{dt} , $n \geq 66$	P_{ana} , $n \geq 60$	$F_{compchar}$, $n \geq 66$	$n \geq 192$
Hennig86	0.64 (0.06)	0.66 (0.06)	0.71 (0.10)	0.67 (0.08)
CAFCA-4	0.60 (0.00)	0.59 (0.09)	0.63 (0.11)	0.61 (0.09)
	Scenario 4, equilibrium/punctuated			
	P_{pr} , $n \geq 60$	P_{pr} , $n \geq 60$	P_{clado} , $n \geq 60$	$n \geq 180$
Hennig86	0.63 (0.04)	0.62 (0.05)	0.58 (0.04)	0.61 (0.05)
CAFCA-4	0.55 (0.06)	0.54 (0.08)	0.53 (0.05)	0.54 (0.07)

coefficients are given in table 5.12. There are no differences within scenarios. UPGMA-3 and UPGMA-4 (product moment) have higher coefficients than UPGMA-1 and UPGMA-2 (euclidian distance). No differences exist between UPGMA-3 (product moment, non-standardized) and UPGMA-4 (product moment, standardized). UPGMA-2 (euclidian distance, non-standardized) results show higher values than the UPGMA-1 (euclidian distance, standardized) results in Scenarios 2 and 4 (punctuated).

Conclusions and discussion

The recent study of Rohlf et al. (1990) showed that in general maximum parsimony methods produce more accurate estimates of the true tree than does UPGMA (average taxonomic distances, non-standardized characters). It was also found that tree topology is an important factor affecting accuracy. Also their evolutionary context models (\equiv Scenarios) were found to be important in

Adequacy of taxonomic methods; further experiments

Table 5.11 R^2 values of the regression analysis of the adequacy (CFI) on the cophenetic correlation coefficient of the estimated tree (r_{cs}). An asterisk denotes values with significance probabilities < 0.01 .

	Scenario 2			Scenario 3			Scenario 4		
	exp.1 P_p $n \geq 59$	exp.2 P_{ex} $n \geq 74$	exp.3 P_{clado} $n \geq 59$	exp1. P_d $n \geq 66$	exp.2 P_{ana} $n \geq 60$	exp.3 $F_{compchar}$ $n = 66$	exp.1 P_d $n \geq 60$	exp.2 P_p $n = 60$	exp.3 P_{clado} $n = 60$
UPGMA-1	1	1	2	11 *	2	0	1	8	3
UPGMA-2	10	3	5	1	4	0	0	1	0
UPGMA-3	2	14 *	1	2	1	3	1	8	1
UPGMA-4	7	17 *	4	4	3	1	6	0	2

Table 5.12 Mean values of the cophenetic correlation coefficient r_{cs} .

estimation method	source of variation			
	exp. 1	exp. 2	exp. 3	exp. 123
	Scenario 2, radiation/punctuated			
	$P_{p'}$, $n = 62$	$P_{ex'}$, $n = 82$	$P_{clado'}$, $n = 69$	$n = 191$
UPGMA-1	0.77 (0.06)	0.81 (0.08)	0.77 (0.05)	0.79 (0.07)
UPGMA-2	0.82 (0.08)	0.86 (0.08)	0.83 (0.06)	0.84 (0.07)
UPGMA-3	0.95 (0.03)	0.93 (0.05)	0.94 (0.02)	0.94 (0.04)
UPGMA-4	0.94 (0.02)	0.93 (0.03)	0.94 (0.02)	0.94 (0.03)
	Scenario 3, equilibrium/gradual			
	P_d , $n = 68$	P_{ana} , $n = 61$	$F_{compchar}$, $n = 66$	$n = 195$
UPGMA-1	0.95 (0.02)	0.93 (0.04)	0.94 (0.04)	0.94 (0.04)
UPGMA-2	0.93 (0.03)	0.93 (0.04)	0.95 (0.05)	0.94 (0.04)
UPGMA-3	0.97 (0.01)	0.96 (0.03)	0.96 (0.04)	0.96 (0.03)
UPGMA-4	0.96 (0.01)	0.96 (0.02)	0.96 (0.03)	0.96 (0.02)
	Scenario 4, equilibrium/punctuated			
	P_d , $n = 61$	$P_{p'}$, $n = 60$	P_{clado} , $n = 60$	$n = 181$
UPGMA-1	0.93 (0.04)	0.93 (0.04)	0.90 (0.06)	0.92 (0.05)
UPGMA-2	0.95 (0.04)	0.95 (0.03)	0.93 (0.06)	0.94 (0.05)
UPGMA-3	0.96 (0.03)	0.96 (0.02)	0.96 (0.03)	0.96 (0.03)
UPGMA-4	0.96 (0.03)	0.96 (0.02)	0.95 (0.02)	0.96 (0.02)

determining accuracy. The results of experiments with Scenario 1 (Heijerman, 1992), which is simpler than the scenarios studied in this paper, showed that there were only small differences between

maximum parsimony and UPGMA (product moment, non-standardized characters). It was also found that the consistency index, i.e. the amount of homoplasy, was the most important factor affecting accuracy.

The results presented here clearly show that UPGMA (product moment, unstandardized characters) performed better in all scenarios and in all experiments than did maximum parsimony and group-compatibility. There is a correlation between the adequacy of the estimation methods and some of the input parameters and/or true tree properties, the most important ones being the consistency index of the true tree (table 5.5) and the stemminess as measured by I_{stem} , but not by I_{stemFS} (tables 5.6 and 5.7). Also under some circumstances the consistency of the estimated tree and its symmetry appeared to affect the adequacy (tables 5.8 and 5.9).

However, it is difficult to compare the present results with those of Rohlf et al. (1990) because of differences in simulation model and experimental design. Sokal et al. (1984) showed that the number of characters in binary notation (n) used in relation to the number of OTUs (t) will affect the accuracy of estimations, and suggested that the ratio $n / (2t - 3)$, "may furnish a rough indication of the adequacy of the data set for cladistic estimation". In their studies Sokal et al. (1984) found that maximum parsimony methods did perform better than phenetic methods as the ratio ranges from 5.91 to 2.71. They also analyzed some data from the literature and concluded that parsimony methods perform better than phenetic methods when this ratio is high, although different threshold values were found. Since Rohlf et al. (1990) used 8 OTUs and 50 characters in their studies, the value of the ratio equals 3.8, based on multistate characters. They noted that after recoding to binary characters this ratio would be much larger. They therefore state that maximum parsimony methods were expected to yield more accurate trees than UPGMA. In the present study 20 OTUs were used each with 50 characters. This would result in the ratio being equal to $50 / 37 = 1.35$. If the calculations are based on binary coded characters, the ratio, named I_{adeq} in this study, becomes

3.41, 2.25 and 2.03 for Scenarios 2, 3 and 4, respectively. For Scenarios 3 and 4 these are relatively low values, and UPGMA did indeed perform better than did maximum parsimony. However, this was also true for Scenario 2. In Scenario 1, I_{adeq} based on binary characters and averaged over all three experiments, was found to equal 3.46 (Standard deviation = 0.79). In this scenario maximum parsimony methods did perform equally well as UPGMA (product moment of unstandardized characters) (Heijerman, 1992). The effects of the number of characters and I_{adeq} on the adequacy of the estimation methods are currently being studied in greater detail.

In the present study four phenetic methods were applied. In all 9 experiments UPGMA based on a product moment matrix of unstandardized characters appeared to perform best. In Scenarios 2 and 3, UPGMA based on product moment correlation of standardized characters was always second best. Clustering based on unstandardized characters always produced better results than when based on standardized characters, using the same similarity criterion. The same results were also found for Scenario 1 (Heijerman, 1992).

Although product moment correlations have been used in a number of phenetic studies (see e.g. Sokal & Sneath, 1963, for references), this coefficient is supposed to be inappropriate in the case of heterogeneity of character vectors (Sokal & Sneath l.c.). The number of states for the various characters used in the present study apparently do not vary too much. However, standardization of the characters did not improve the accuracy. Eades (1965) argued that "the theoretical basis of the correlation coefficient as a measure of resemblance is unsound ...", and that the use of the average taxonomic distance (d ; Sokal & Sneath, 1963) will lead to more satisfactory results. Eades (1965) concludes that the correlation coefficient is inappropriate as a measure of taxonomic resemblance, although he admits that the use of a large number of characters would lessen its disadvantages.

When comparing phenetic and phylogenetic methods, different authors have often used different clustering strategies.

Sokal et al. (1990) compared maximum parsimony and UPGMA clustering. They used average taxonomic distances (euclidian distance divided by \sqrt{n} , where n is the number of characters) as a similarity measure and average linkage clustering (GROUP) as the clustering technique. In general they did not standardize characters. However, they did some evaluations using standardized characters and found that accuracy was somewhat lower. They further argued that for their study, where all characters are in the same arbitrary units (like in the present study), the use of unstandardized characters seems appropriate. Fiala & Sokal (1985) tested UPGMA clustering based on a matrix of Manhattan distances. In their simulation study, Kim & Burgman (1988) applied UPGMA clustering based on average taxonomic distance of standardized (gene frequency) data. Rohlf & Wooten (1988) used UPGMA based on a variety of similarity and dissimilarity coefficients. They used average taxonomic distances as well as product moment correlations of both standardized and unstandardized gene frequency data sets, and a number of genetic distance coefficients. Inspection of their table 1 (Rohlf & Wooten 1988:588) reveals that UPGMA based on product moment correlations and UPGMA based on average taxonomic distance are about equally adequate. In view of the present results, it seems there is a need for additional work to evaluate the adequacy of the various phenetic strategies.

In their simulation experiments, Rohlf et al. (1990) computed trees with eight OTUs only. Therefore they were able to use exact algorithms and find the actual minimum length trees. They used the branch-and-bound option of the PAUP program and found multiple trees in a number of cases, in which a majority-rule consensus tree (Margush & McMorris, 1981) was constructed. Next, the strict consensus index, *C_{ic}* (\equiv *CFI*), between the majority-rule consensus of the MP (maximum parsimony) trees and the true cladogram was computed (Rohlf et al., 1990:1676). This procedure is different from the one followed in the present study. In case of multiple trees, *CFI* was calculated between every single tree generated and the true tree. The mean *CFI* was computed for the set of equally parsimonious trees and used in subsequent

analyses. As a consequence, the consensus tree of Rohlf et al. will on average have lower *CFI* (*C/c*) values. In fact, I feel that they underestimate the consensus between the estimations and the true trees.

Summarizing, the current results show that, under the evolutionary conditions tested:

- a) UPGMA, using product moment correlations of unstandardized characters, produces the most accurate estimations;
- b) UPGMA, using product moment correlations and euclidian distances of unstandardized characters, produce better estimations than any of the phylogenetic methods;
- c) Hennig86 performs better than Wagner78 in almost all cases, though equally well in a few;
- d) compatibility analysis as performed by CAFCA does produce the least accurate estimations;
- e) Although UPGMA, using product moment correlations and euclidian distances of unstandardized characters, was generally superior to the other methods, under certain evolutionary conditions (*i.e.* in certain scenarios) it still performed rather poorly.

Further work needs to be done to study, among others things, the effects of the number of characters in relation to the number of OTUs in the study on the adequacy of phenetic as well as phylogenetic methods. Because there are rather large differences between the various UPGMA strategies, it seems desirable to perform additional tests to evaluate their adequacy.

In agreement with Fiala & Sokal (1988) it must be concluded that none of the phylogenetic estimation methods tested are generally adequate. Therefore I also subscribe to the viewpoint of Rohlf et al. (1990:1671) that "... the great majority of estimated phylogenetic trees are likely to be quite inaccurate ..."

Zusammenfassung

Über die Eignung numerisch-taxonomischer Methoden: Weitere Untersuchungen mit simulierten Daten

Die Brauchbarkeit verschiedener phänetischer und phylogenetischer Berechnungsmethoden wurde an Hand simulierten Daten überprüft. Zwei verschiedene Parsimonie-Programme wurden benutzt, um Maximum-Parsimonie-Stammbäume zu konstruieren (Wagner78 und Hennig86). Das CAFCA-Programm wurde eingesetzt, um Kompatibilitäts-Analysen durchzuführen. Schließlich wurden mit Hilfe von UPGMA-Programmen vier verschiedene Cluster-Analysen gemacht. Zur Erzeugung von Datensätzen unter verschiedenen Evolutionsbedingungen wurde das Simulations-Modell GENESIS eingesetzt. Die Auswirkungen der Eingabeparameter und die der Stammbaumeigenschaften auf die Genauigkeit der Stammbaumkonstruktion wurden bewertet. Die UPGMA-Methode, die sich auf die Korrelation von Productmomenten von nicht-standardisierten Characteren stützt, erwies sich unter allen getesteten Evolutionsbedingungen als die beste Methode. Der Effekt der Eingabeparameter auf die Genauigkeit war relativ unbedeutend. Unter den Baum-Statistika erwies sich ein Topologie-Index der wahren Bäume als wichtigste Bezugsgröße für die Genauigkeitsabschätzung.

Acknowledgments

I want to thank Drs M. C. Gillham, R. J. Post, P. W. F. de Vrijer, A. J. de Winter and R. Zandee for comments on the manuscript.

References

- Colless, D. H., 1980. Congruence between morphometric and allozyme data for *Menidia* species: a reappraisal. — *Systematic Zoology* 29: 288-299.
- Colless, D. H., 1982. (Review of) Phylogenetics: the theory and practice of phylogenetic systematics. — *Systematic Zoology* 31: 100-104.
- Dawkins, R., 1986. *The blind watchmaker*. Harlow, Longman.
- Eades, D. S., 1965. The inappropriateness of the correlation coefficient as a measure of taxonomic resemblance. — *Systematic Zoology* 14: 98-100.

Adequacy of taxonomic methods; further experiments

- Farris, J. S., 1978. *Wagner78*; manual, documentation and a FORTRAN IV program.
- Farris, J. S., 1988. *Hennig86, version 1.5*; Hennig86 reference manual and program.
- Farris, J. S., 1989. The retention index and the rescaled consistency index. — *Cladistics* 5: 417-419.
- Felsenstein, J., 1987. *PHYLIP; Package for inferring phylogenies, version 3.0*; manual, documentation and several PASCAL programs. University of California Herbarium, Berkely, California.
- Fiala, K. L. & R. R. Sokal, 1985. Factors determining the accuracy of cladogram estimation: evaluation using computer simulation. — *Evolution* 39: 609-622.
- Heijerman, Th., 1988. GENESIS: a simulation model of phylogeny. 1. The origin and early evolution of character state vectors. — *Zeitschrift für zoologische Systematik und Evolutionsforschung* 26: 409-424.
- Heijerman, Th., 1990. GENESIS: a simulation model of phylogeny. 2. A sensitivity analysis. — *Zeitschrift für zoologische Systematik und Evolutionsforschung* 28: 81-93.
- Heijerman, Th. 1992. Adequacy of numerical taxonomic methods. A comparative study based on simulation experiments. — *Zeitschrift für zoologische Systematik und Evolutionsforschung* 30: 1-20.
- Kim, J. & M. A. Burgman, 1988. Accuracy of phylogenetic estimation methods under unequal evolutionary rates. — *Evolution* 42: 596-602.
- Kluge, A. G. & J. S. Farris, 1969. Quantitative phyletics and the evolution of anurans. — *Systematic Zoology* 18: 1-32.
- Luckow, M. & R. A. Pimentel, 1985. An empirical comparison of numerical Wagner computer programs. — *Cladistics* 1: 47-66.
- Margush, T. & F. R. McMorris, 1981. Consensus *n*-trees. — *Bulletin of Mathematical Biology* 43: 239-244.
- Platnick, N. I., 1987. An empirical comparison of microcomputer parsimony programs. — *Cladistics* 3: 121-144.
- Platnick, N. I., 1989. An empirical comparison of microcomputer parsimony programs, II. — *Cladistics* 5: 145-161.
- Rohlf, F. J., 1982. Consensus indices for comparing classifications. — *Mathematical Biosciences* 59: 131-144.
- Rohlf, F. J. & M. C. Wooten, 1988. Evaluation of the restricted maximum-likelihood method for estimating phylogenetic trees using allele-frequency data. — *Evolution* 42: 581-595.
- Rohlf, F. J., W. S. Chang, R. R. Sokal & J. Kim, 1990. Accuracy of estimated phylogenies: effects of tree topology and evolutionary model. — *Evolution* 44: 1671-1684.

- Sackin, M. J., 1972. "Good" and "bad" phenograms. — *Systematic Zoology* 21: 225-226.
- Sanderson, M. J., 1990. Flexible phylogeny reconstruction: a review of phylogenetic inference packages using parsimony. — *Systematic Zoology* 39: 414-420.
- Shao, K., 1983. *Consensus methods in numerical taxonomy*. Ph.D. Dissertation, State University of New York, Stony Brook.
- Shao, K. & R. R. Sokal, 1990. Tree balance. — *Systematic Zoology* 39: 266-276.
- Sokal, R. R., 1983. A phylogenetic analysis of the Caminalcules. I. The data base. — *Systematic Zoology* 32: 159-184.
- Sokal, R. R., K. L. Fiala & H. Hart, 1984. OTU stability and factors determining taxonomic stability: examples from the Caminalcules and the Leptopodomorpha. — *Systematic Zoology* 33: 387-407.
- Sokal, R. R. & K. Shao, 1985. Character stability in 39 data sets. — *Systematic Zoology* 34: 83-89.
- Sokal, R. R. & P. H. A. Sneath, 1963. *Principles of numerical taxonomy*. W. H. Freeman and Co., San Francisco.
- Swofford, D. L., 1985. *PAUP: Phylogenetic Analysis Using Parsimony, version 2.4*; manual, documentation and program. Illinois Natural History Survey, Champaign, Illinois.
- Wishart, D., 1986. *CLUSTAN Version 3.2. A cluster analysis package*. Edinburgh University, Program Library Unit.
- Zandee, M., 1988. *C.A.F.C.A.: a Collection of APL Functions for Cladistic Analysis, PC version 1.9*; manual, documentation and program.

6

Adequacy of numerical taxonomic methods

Why not be a pheneticist? ¹

"The school of distance measurers, or 'numerical taxonomists', has become a bit unfashionable lately. My view is that the unfashionableness is a temporary phase, as fashions often are, and that this kind of 'numerical taxonomy' is by no means easily to be written off. I expect a comeback."
Dawkins (1986:280)

Abstract

A comparative study was conducted of the adequacy of a number of phenetic procedures and two phylogenetic ones, based on simulation experiments. The phenetic clustering procedures examined included the single linkage method, the complete linkage method, the average linkage method, centroid sorting, the median or Gower's method, Ward's method, the Lance-Williams flexible BETA method and McQuitty's similarity analysis. The squared euclidean distance, the product moment correlation and the cosine coefficient were used as (dis)similarity criteria, calculated both on standardized and non-standardized data. The Hennig86 and Wagner78 programs were used to calculate most parsimonious trees. The relative efficiency of these numerical procedures was evaluated using data matrices simulated under a large variety of evolutionary conditions. McQuitty's similarity analysis and the

¹ Intended for publication as: Heijerman, Th. Adequacy of numerical taxonomic methods; why not be a pheneticist? — *Zeitschrift für zoologische Systematik und Evolutionsforschung*. Submitted.

average linkage method based on cosine- or product moment correlation matrices for unstandardized characters were found to perform consistently better than the other phenetic clustering procedures and maximum parsimony. However, once again it was shown that none of these methods was very accurate. The majority of published phylogenies should therefore only be considered best approximations of the true tree.

Key words: Numerical taxonomy — Phylogeny estimation — Simulation model — Phenetics — Maximum parsimony

Introduction

A large number of numerical taxonomic methods for the estimation of phylogenetic trees have been developed. So the question may arise how one should choose for the best method and also, how confident can we be about the results of the various methods. There have indeed been a number of studies evaluating the adequacy of numerical taxonomic methods. Most of these studies made use of artificial data sets mainly produced by computer simulation (e.g. Sokal, 1983; Fiala & Sokal, 1985; Kim & Burgman, 1988; Rohlf et al., 1990; Heijerman, 1992, 1993; Kim, 1993; Kim, Rohlf & Sokal, 1993). In other studies phylogenetic estimation methods were evaluated using data sets of the kind used in molecular systematics (e.g. Rohlf & Wooten, 1988; Sourdis & Nei, 1988; Nei, 1991; Huelsenbeck & Hillis, 1993). From some of these and other simulation experiments one may conclude that the large majority of published phylogenies will be rather inaccurate and therefore must be viewed sceptically. Also these experiments do not demonstrate unequivocally the superiority of numerical cladistic procedures over phenetic ones. Nevertheless, it now seems that most workers employ numerical cladistic methods as their preferred tools to construct phylogenetic hypotheses, while phenetic methods are only used for other taxonomic purposes, mainly identification. To quote Quicke (1993:85): "... phenetic clustering methods ... neither

provide reliable evidence of evolutionary relationships, nor form a sound basis for classification", and "... many of the taxonomic techniques developed by workers such as Sokal and Sneath are now largely defunct or redundant, ... (Quicke, 1993:84). Further to this Ax (1987:7) states: "The connections worked out by phenetics between different species and groups of species, remain pure similarity connections (phenetic relationships) - they have no necessary link with the closed descent communities that exist in Nature".

In Heijerman (1992, 1993) results were presented of experiments designed to evaluate several phenetic and cladistic methods: Wagner parsimony as performed by Wagner78 (Farris, 1978), PAUP (Swofford, 1985), PHYLIP (MIX) (Felsenstein, 1987), and Hennig86 (Farris, 1988), group-compatibility analysis as performed by CAFCA (Zandee, 1988) and four phenetic procedures: UPGMA clustering using euclidian distances and product moment correlations, calculated both from standardized as well as non-standardized characters, employing the CLUSTAN package (Wishart, 1986). The results showed that under a certain rather simple evolutionary scenario, the overall differences in adequacy between Wagner parsimony as performed by PAUP, PHYLIP (MIX) and Hennig86, and UPGMA clustering with product moment correlations of unstandardized characters, were rather small, and that these methods were more accurate than Wagner parsimony with Wagner78 and group-compatibility with CAFCA. It was also shown that under the more complex evolutionary conditions of my 1993-study, UPGMA using product moment correlations of unstandardized characters, performed better than parsimony, group-compatibility and the other three clustering procedures. Under all evolutionary conditions, UPGMA using product moment correlations of unstandardized characters, appeared to perform best. All simulation experiments showed the superiority of one of these four phenetic techniques over the others: UPGMA applied to a product moment correlation matrix of unstandardized characters always gave the best results, followed by UPGMA of product moment correlations of standardized characters, UPGMA of squared euclidian distances of

unstandardized characters and UPGMA of squared euclidian distances of standardized characters, generally in this order.

The above mentioned experiments were carried out using the simulation model GENESIS, which can produce sets of recent 'species' with known phylogenies and character state distributions under various evolutionary scenarios (for details about GENESIS, see Heijerman, 1988, 1990). These earlier experiments were primarily designed to measure the effects of input parameters and tree properties on the accuracy of estimated trees. In the present paper results are presented of experiments in which the relative adequacy of a much larger number of phenetic procedures was examined. A total of 36 phenetic techniques were evaluated, together with two phylogenetic ones, viz. Wagner parsimony as performed by both the 'primitive' Wagner78 program and the more 'advanced' Hennig86 program. The latter program is often recommended as the "tool of choice for practising systematists" (Platnick, 1989:160, see also Heijerman, 1992).

Material and methods

Data sets used for analysis

For details about the approach followed to produce character data sets with GENESIS, see Heijerman (1988, 1992 & 1993). GENESIS was used to produce character data sets of recent species with known phylogenies, generated under different evolutionary conditions. GENESIS can be made to simulate evolution according to four major evolutionary scenarios. In Scenarios 1 and 2, speciation and extinction probabilities are constant over time and the number of extant species increases exponentially with time ('radiation' version). In Scenarios 3 and 4, the number of species oscillates around an equilibrium number of recent species ('equilibrium' version). In Scenarios 1 and 3, the rate of character evolution is the same in both daughter lineages ('gradualistic' version). In Scenarios 2 and 4 this rate of character

change is unequal in the two daughter lineages ('punctuated' version).

Within each of the scenarios the values of a number of input parameters have to be supplied by the user. By selecting the proper parameter values, GENESIS can be made to produce *ideal* data sets on which phylogenetic estimation methods are expected to perform excellently. By allowing back-mutations, homoplasous character changes etcetera, the amount of *noise* contained by the data matrix can be increased and as a consequence one may expect the estimation methods to perform increasingly badly.

In this study we are only interested in the relative accuracy of the trees estimated by the different phenetic and phylogenetic methods. In the two previous studies (Heijerman, 1992, 1993) the effects of evolutionary scenario and tree properties on accuracy were investigated, and experiments were carried out in which the value of one input parameter was changed while keeping the other input parameters constant. In this study a somewhat different experimental design was employed.

Within each of the main scenarios a number of evolutionary conditions were simulated by using different values for the input parameters of GENESIS. A total of 50 simulations were run for every single set of input parameter values. All simulations were run to produce 20 recent species, each with 50 characters. In all experiments the probability of speciation, P_{split} , was set at 0.10. In Scenarios 3 and 4 ('equilibrium' version) the equilibrium number of species, N_{equi} , was set at 20, and t_{min} , the minimum number of timesteps that must have been passed until the simulation is allowed to stop, was set at 100.

The user must supply: the values of the character change probability, P_{ana+} ; the probability of a reversed character state change, P_{ana-} ; the probability of extinction, P_{ex} ; the number fraction of characters that are fully consistent with the tree, $F_{compchar}$ (Scenario 1); the character change probability, P_{clado+} ; the probability of a reversed character state change, P_{clado-} ; P_{ex} ; the value of the 'punctuation'-parameter, P_p (which determines how much faster evolution proceeds, expressed as the number of character state changes per time unit, in one daughter lineage

compared to the other); $F_{compchar}$ (Scenario 2): P_{ana+} ; P_{ana-} ; N_{equi} , which is the equilibrium number of species; the value for the damping parameter, P_d ; $F_{compchar}$ (Scenario 3): P_d ; P_p ; P_{clado+} ; P_{clado-} ; N_{equi} ; $F_{compchar}$; t_{min} (Scenario 4).

The choice for input parameter values was such as to guarantee a large variation in tree statistic values. The probability of character state change was adjusted in an effort to equalize the number of evolutionary steps within Scenarios. However, the adjustment was difficult in cases where $F_{compchar}$ equalled 0.8. Values of input parameter used within each of the scenarios are listed in appendix 6.1.

Within each scenario 12 evolutionary conditions were simulated, within which 50 simulation runs were performed, summing up to a total of $4 \times 12 \times 50 = 2400$ true trees.

Tree comparison and the measure of adequacy

Data sets and their corresponding true trees were characterized using a number of tree statistics (Heijerman, 1988). The data sets were used as input for the phylogenetic and phenetic estimation methods. The estimated trees were compared with the corresponding true trees. The adequacy of the estimation methods was measured by the consensus of the true trees with the estimated ones, using Colless' consensus fork index (CFI) (Colless, 1982; also CI_c , Rohlf, 1982). The index measures the difference between the strict consensus tree and a bush, and is defined as the number of branching points in the consensus tree (ignoring the basal one), normalized by dividing by $n - 2$, where n is the number of OTUs.

Estimation methods

Minimum length trees were constructed using the Distance Wagner procedure of Wagner78 (Farris, 1978) and Hennig86 Version 1.5 (Farris, 1988). Wagner78 was run using the HOM option, to request computation of the total homoplasy and the deviation ratio. The ancestor was added to the data matrix as a

Adequacy of taxonomic methods; why not be a pheneticist?

Table 6.1 Overview of phenetic procedures examined in this paper. All procedures were tested using standardized as well as unstandardized data.

(dis)similarity criterion	cluster criterion
Squared euclidian distance	Single linkage (Nearest neighbour)
..	Complete linkage (Furthest neighbour)
..	Average linkage (Group average)
..	Centroid sorting
..	Median (Gower's method)
..	Ward's method (error sum of squares)
..	Lance-Williams flexible BETA (BETA = -0.25)
..	McQuitty's similarity analysis
Product moment correlation	Single linkage
..	Complete linkage
..	Average linkage
..	Centroid sorting
..	McQuitty's similarity analysis
Cosine (Ochiai coefficient)	Single linkage
..	Complete linkage
..	Average linkage
..	Centroid sorting
..	McQuitty's similarity analysis

first taxon, with all characters having state "0". As the results of Wagner78 depend of the order of species in the input matrix (Heijerman, 1992, 1993) Wagner78 was run ten times with different orderings of species. From the ten solutions one of the most parsimonious trees was used for further analysis. Hennig86 was run using the MHENNIG* and BB options, as recommended by Platnick (1989), and treating all characters as ordered. Hennig86 would often produce multiple equally parsimonious trees, of which a maximum of 50 were used for further analysis. The consensus of all these trees with the true tree was calculated separately, and for further analysis the mean CFI was used.

For phenetic analysis the UPGMA cluster technique was employed using a variety of (dis)similarity indices, both on raw, unstandardized data as well as on standardized data, in combination with three cluster criteria. Table 6.1 gives an overview of the (dis)similarity indices and cluster criteria used.

Chapter 6

Table 6.2 Descriptive statistics for *CFI*. Sample size for each procedure: $n = 2400$; Procedures arranged in decreasing order of mean *CFI*; *cv* = coefficient of variation; Grouping is based on Range test; T-test ($\text{Alpha} = 0.05$, $\text{MSE} = 0.022266$, Critical value of $T = 1.96$, Least Significant Difference = 0.0084). For description of numerical procedures, see table 6.1. Stand (—/+) indicates whether unstandardized or standardized data were used. COS = cosine; PM = product moment; EUC = squared euclidian distances.

Numerical procedure		stand	mean	cv	T grouping
COS	McQuitty	—	0.743	8.276	*
COS	Average	—	0.742	8.269	*
PM	McQuitty	—	0.741	8.358	*
PM	Average	—	0.740	8.279	*
COS	Complete	—	0.725	9.210	*
PM	Centroid	—	0.725	8.438	*
COS	Centroid	—	0.723	8.571	*
PM	Complete	—	0.723	9.226	* *
PM	Single	—	0.715	8.710	* *
COS	Single	—	0.712	8.599	*
PM	McQuitty	+	0.696	9.312	*
PM	Average	+	0.696	9.072	*
PM	Single	+	0.689	9.297	* *
COS	McQuitty	+	0.684	11.380	*
COS	Average	+	0.683	11.485	*
PM	Complete	+	0.674	9.738	*
COS	Single	+	0.670	12.889	* *
COS	Complete	+	0.664	12.102	* *
EUC	Lance	—	0.663	15.045	* *
PM	Centroid	+	0.663	10.173	* *
EUC	McQuitty	—	0.662	16.215	* *
EUC	Average	—	0.662	16.222	* *
EUC	Ward	—	0.657	14.078	* *
COS	Centroid	+	0.650	12.240	* *
EUC	Lance	+	0.648	16.253	* *
EUC	Complete	—	0.647	17.102	* *
EUC	Ward	+	0.646	15.724	* *
Hennig86			0.641	17.418	* *
EUC	Median	—	0.640	17.347	* *
EUC	Centroid	—	0.637	17.836	* *
EUC	McQuitty	+	0.629	16.664	* *
EUC	Average	+	0.626	16.483	* *
Wagner78			0.625	17.115	* *
EUC	Complete	+	0.624	17.813	* *
EUC	Single	—	0.617	18.568	*
EUC	Median	+	0.593	18.242	*
EUC	Centroid	+	0.585	19.065	* *
EUC	Single	+	0.583	19.566	*

Results

Simulated trees

Forty-eight experiments were carried out resulting in 48 sets of 50 true trees. Each of these 2400 true trees was characterized by a number of descriptive statistics (for definitions and references, see Heijerman, 1988, 1992): the reversal index, R_1 ; the dendritic index, DI ; the consistency index, CI ; Colless index of symmetry, I_{col} ; adequacy of the data, I_{adeq} ; the stemminess, I_{stem} ; the Fiala & Sokal index of stemminess, I_{stemFS} ; and Sackins index of symmetry, I_{sack} . As a summary, in appendix 6.2 the mean values of these tree statistics are given, per experiment and for the total set of trees, showing that there is variation between experiments, which is larger in some statistics (CI , R_1) and smaller in others (I_{sack}). Also variation within experiments is larger for some statistics (I_{col}) and smaller in others (R_1 , DI , I_{sack}).

Adequacy of estimation methods

In table 6.2 all methods tested are listed together with their corresponding mean CFI values as measures of accuracy. The methods are arranged from best performing (top) to worst (bottom). A range test has been conducted to test for differences between methods. On the basis of the results, as presented in table 6.2, roughly 5 groups of methods can be distinguished. Clustering a correlation matrix (product moment or cosine) from unstandardized data applying McQuitty's similarity analysis or the average linkage clustering technique (UPGMA) clearly produce the best results. The second best group contains all remaining techniques (complete linkage, centroid sorting and single linkage) that use a correlation matrix from unstandardized characters. The third group contains McQuitty's similarity analysis and average linkage clustering of a correlation matrix, but based on standardized characters, plus single linkage clustering of a product moment matrix from standardized characters. The group with the three worst performing methods contains median, centroid

Table 6.3 Number of times (No) that a given method belonged to the group of best performing methods. Maximum = 48 (100%), N= 2400. For description of numerical procedures: see Table 6.1. Stand (—/+) indicates whether unstandardized or standardized data were used.

Numerical procedure		stand	No	%
COS	McQuitty	—	48	100
COS	Average	—	48	100
PM	McQuitty	—	48	100
PM	Average	—	48	100
PM	Centroid	—	48	100
COS	Complete	—	46	96
COS	Centroid	—	45	94
PM	Complete	—	42	88
COS	Single	—	42	88
PM	Single	—	41	85
PM	McQuitty	+	25	52
EUC	McQuitty	—	23	48
COS	McQuitty	+	23	48
PM	Average	+	22	46
COS	Average	+	21	44
EUC	Average	—	21	44
EUC	Lance	—	17	35
EUC	Complete	—	14	29
EUC	Ward	+	14	29
EUC	Lance	+	14	29
PM	Single	+	14	29
COS	Single	+	14	29
EUC	Ward	—	13	27
EUC	Centroid	—	11	23
PM	Centroid	+	10	21
EUC	Median	—	9	19
PM	Complete	+	8	17
COS	Complete	+	7	15
EUC	McQuitty	+	7	15
Hennig86			7	15
Wagner78			7	15
COS	Centroid	+	7	15
EUC	Complete	+	6	13
EUC	Average	+	5	10
EUC	Centroid	+	3	6
EUC	Single	—	2	4
EUC	Median	+	2	4
EUC	Single	+	0	0

clustering and single linkage clustering of a matrix of squared euclidean distances calculated for standardized characters. The majority of methods, that is the remaining 18 clustering methods

and the two phylogenetic methods, fall in the fourth group. From table 6.2 it can also be seen that the range of variation (cv) increases with decreasing accuracy.

Also for each of the 48 experiments separately, a range test was carried out. Table 6.3 presents the number of times that a given methods fell in the group of best performing ones. On the basis of these results the methods can be classified into three groups, the group with the best performing methods containing all clustering techniques (McQuitty similarity analysis, average linkage, centroid sorting, complete linkage and single linkage) that use a correlation matrix (cosine or product moment correlation) from unstandardized characters as a resemblance matrix. McQuitty and average linkage of a correlation matrix but based on standardized characters fall in the intermediate group of methods, together with McQuitty's analysis and average linkage of a matrix of squared euclidean distances from unstandardized characters. The remaining 22 techniques, inclusive of Hennig86 and Wagner78 end up in the third group.

As a summary of the results a classification is presented of methods, based on their *CFI*-values per experiment. Classifications were produced for all methods tested and for the methods based on unstandardized characters. As a clustering procedure UPGMA was used, based on a distance matrix of unstandardized *CFI*-values, but also a number of other procedures were followed, showing similar results. Methods that show a similar performance pattern over all 48 experiments, will be grouped together. Figure 6.1 shows the dendrogram for the phenetic methods based on unstandardized characters plus the two cladistic procedures. The first division leads to a group of phenetic correlation-procedures and a group of phenetic distance-procedures plus the two cladistic approaches. This latter cluster is next divided into phenetic and cladistic procedures respectively. The phenetic correlation-procedures are next divided into the group of best performing methods and the remaining correlation-procedures. This clustering corresponds very well with the clustering based on the range tests of table 6.2.

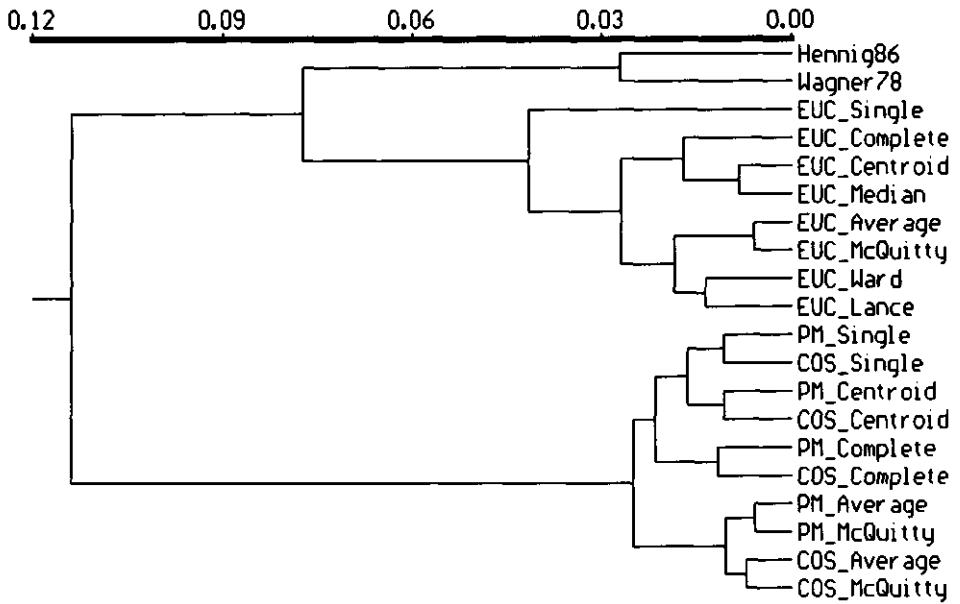


Figure 6.1 Classification of numerical taxonomic procedures, showing clusters of methods with similar performance patterns (UPGMA; average taxonomic distance; analysis based on unstandardized *CFI*-values).

Conclusions and discussion

Superior methods

From the results as presented in tables 6.2 and 6.3 it is obvious that the various methods tested are not similar in their ability to find trees that are as close to the true tree as possible. McQuitty's similarity analysis and average linkage on a correlation matrix (either using cosine or product-moment correlation) of unstandardized characters are superior to the remaining phenetic procedures and even to Hennig86 and Wagner78, in obtaining closest approximations of the correct tree. These results are consistent with earlier results from experiments with GENESIS (Heijerman, 1992, 1993) which showed that the average linkage method using a matrix with squared euclidean distances calculated

for unstandardized data proved to be superior to the other UPGMA clustering techniques tested and Wagner parsimony as performed by Wagner78 and Hennig86, using Scenarios 2, 3 and 4. Experiments conducted in the simpler Scenario 1 showed that UPGMA (average linkage, squared euclidean distance, unstandardized characters) performed equally good as did Hennig86. Fiala & Sokal (1985:620) concluded from a simulation analysis in which they tested Wagner parsimony, character compatibility and UPGMA clustering (using 20 OTUs and 25 multistate ordered characters) that "... differences among the methods are rather small, at least small enough to be completely overshadowed by the common deficiencies ..." Rohlf et al. (1990) compared UPGMA clustering and Wagner parsimony methods (using 8 OTUs and 50 multistate ordered characters): "Overall, maximum parsimony yielded more accurate trees than UPGMA - but that was expected for these simulations since many more characters than OTUs were used." The study of Kim et al. (1993) is an extension of the work of Rohlf et al. (1990) just mentioned, to include the neighbor-joining method. They concluded (in the abstract) that "... no one method was more accurate than the other two for all combinations of treatments." A number of simulation studies to evaluate relative efficiencies of estimation methods were carried out using molecular data or allele-frequency data (see Rohlf & Wooten, 1988; Nei, 1991; Huelsenbeck & Hillis, 1993). Rohlf & Wooten (1988) compared the restricted maximum-likelihood, Wagner parsimony and UPGMA (using 20 OTUs) and concluded that (pg. 587): "All of the methods examined were similar in their ability to estimate the true cladograms."

Although the degree of difference in adequacy of the estimation methods examined may be considered relatively small (mean *CFI* ranging from 0.583 to 0.743; see table 6.2), the present study nevertheless showed that these differences show a consistent trend and that methods can be clearly clustered into groups with differing ability to estimate the correct tree.

GENESIS generates data matrices with qualitative, ordered multistate characters. The standard phenetic procedure to be used for these kind of data sets seems to be UPGMA clustering of

either a correlation or a distance matrix, based on standardized characters (e.g. Sokal, 1983; Sokal, Fitch & Hart, 1984). It was shown in a number of studies (Sneath & Sokal, 1973, and references therein), that standardization may have distinct effects on correlations but not so on distances. Standardization is said to reduce "the atypicality of aberrant OTUs, particularly when correlations are employed". Sneath and Sokal (1973:178, 157), recommend both the product moment coefficient and distances as useful similarity coefficients, and for predominantly multistate characters they advise standardization. Although one could have expected the effect of standardization of characters in the data sets generated by GENESIS to be of little importance since all the characters are in the same arbitrary units, standardization clearly appeared to have a negative effect on the results.

How accurate?

The accuracy of the four superior procedures, measured as the mean *CFI*, is approximately 0.74. The poorest method yielded an average of 0.58 while Hennig86 and Wagner78 obtained an accuracy of 0.64 and 0.63 respectively. Earlier experiments with GENESIS resulted in accuracies of 0.69, 0.76, 0.65 and 0.68 for UPGMA (average linkage, squared euclidean distance, non-standardized characters), UPGMA (average, product moment, non-standardized characters), Wagner78 and Hennig86 respectively, averaged over all simulation conditions. For their Wagner results, Fiala & Sokal (1985) found a mean *CFI* (CI_c), calculated over all experimental treatments, of 0.728 and 0.733 for their UPGMA results. As a striking result of their study they state (pg. 620) that none of the methods examined is "especially good at reconstructing phylogenies accurately". Rohlf & Wooten (1988:587), comparing amongst others Wagner78 and UPGMA (average with a variety of similarity and dissimilarity coefficients) and also using strict consensus trees and CI_c (*CFI*), observed consensus scores generally in the range of 0.7 - 0.8, which they seem to consider relatively high values. Inspection of their table 1 shows however that there is a very large variation (range for

Adequacy of taxonomic methods; why not be a pheneticist?

Wagner78: 0.111 - 1.000; UPGMA (correlation): 0.056 - 1.000; UPGMA (average taxonomic distance): 0.056 - 1.000). Rohlf et al. (1990:1682) found an average C/c of 0.552 for maximum parsimony trees and 0.487 for UPGMA trees. Also these C/c values show a large variation (see their tables 2 and 3). The accuracy of trees produced by Kim et al. (1993) using the neighbor-joining method, averaged over all treatments, was 0.573. Both Kim et al. (1993) and Rohlf et al. (1990) consider these overall **low values** "somewhat discouraging" because these results indicate that "the great majority of estimated phylogenetic trees are likely to be quite inaccurate". Compared with these other studies, the value of 0.74 obtained for the four most accurate methods of this study, is a relatively high one. In the present study 20 OTUs were used, so a fully resolved consensus tree contains 18 internal nodes. A value of 0.74 indicates that the true tree and the estimated one share 74% information. So at best, approximately 13 out of the 18 subgroups would be correctly estimated. In my opinion we cannot be satisfied in correctly having obtained less than, say 15 out of 18 subgroups (in which case CFI would equal 0.833, which is also the cutoff point used by Rohlf et al., 1990). These results, and those of other authors, strongly indicate that estimated phylogenetic trees (cladograms or phenograms) cannot be considered good approximations of the true phylogeny.

Why not be a pheneticist?

Order in nature is produced by the process of evolution. We do not only want to reconstruct the evolutionary history of taxa, but we also want to construct classifications. It is my view that these classifications should reflect this natural order. So our classifications should correspond to evolutionary history, that is, they should be consistent with phylogenies in such a way that the taxa are arranged according to the branching pattern in the true tree. In cladistic classification all supra-specific taxa are set up as natural, that is monophyletic groups diagnosed by apomorphic character states. Phenetic procedures are not developed to

discover natural, monophyletic taxa, but groups based on phenetic relationship instead, that is based on overall similarity. This overall similarity may and will include similarity due to homoplasy (primitive similarity) and therefore some of the groups constructed may be of paraphyletic or polyphyletic nature. So it seems that from this theoretical point of view one would do best to be a cladist.

However, simply stated we are faced with two sources of *noise* that can obscure the signal we are looking for: divergence and homoplasy (convergence and parallelism). Divergence is the situation where ancestral relationship is closer than phenetic relationship (see also Pankhurst, 1991), whereas convergence results in closer phenetic than ancestral affinity. Divergence does not effect the topology of cladograms but only branch lengths. The topology of phenograms, however, may indeed be effected: A taxon with many autapomorphies will cluster further away from the other taxa, at the same time the taxa in which these autapomorphies are absent will cluster together. Similarity among the other members is increased which may result in the formulation of paraphyletic groups.

Homoplasy affects both cladograms and phenograms: the two distantly related taxa may be clustered together based on their convergent characters which will result in polyphyletic groups and they will show greater overall similarity and will therefore be grouped into the same phenetic cluster.

It is clear that in the absence of conflicting characters ($CI = 1.0$) cladistic methods will always be able to recover the true tree (unless of course all characters show homoplasy in the same taxa). Sneath (1988:266): "Unless all characters are perfectly nested the synapomorphic methods fail, because one must prejudge the issue by deciding which exceptions are the false synapomorphies. Perfectly nested data is exceptional unless the number of characters is small, or biased by censoring". So only if we have a perfect or near perfect data set we may expect cladistic estimation methods to perform better. However, homoplasy, viewed as incongruent data, seems to be commonplace. We almost certainly will have sampled some

characters that are in conflict with others. The smaller this number the larger our confidence in the resulting tree can be.

Simulation studies indeed showed that accuracy of estimation is affected by a number of tree properties (Fiala & Sokal, 1985; Rohlf et al., 1990; Kim et al., 1993). Earlier experiments with GENESIS showed that the consistency index of the true tree may be an important factor affecting the accuracy of maximum parsimony methods. Also tree topology, as measured by the stemminess (I_{stem} , but not I_{stemFS}) was shown to be an important factor (Heijerman, 1992, 1993). Maximum parsimony methods appeared to be more affected by tree properties than UPGMA techniques. In the almost or complete absence of conflicting characters (CI near to unity) and with a sufficient number of characters to allow for a fully bifurcating tree, maximum parsimony procedures will easily recover the true tree topology, whereas phenetic methods were shown not to be able to produce the correct tree, even not under these ideal circumstances. As soon as some character conflict was present in the data, phenetic methods would produce equally good or even better estimations of the true phylogeny.

There is no reason to assume that nature itself did proceed most economically. Therefore we may expect our data to contain at least some homoplasy. As a consequence the shortest possible trees will not necessarily be closest to the true one. GENESIS experiments also showed that in most cases the consistency of the true tree was lower than that of the estimates. So even if one favours cladistic methods, it is recommended to also consider nonminimal trees.

There seems to be no superior approach: If we base our analysis on perfect data we should be using cladistic methods, in many cases however phenograms may provide closer approximations to the true tree, though still not very good ones. Despite a large and growing number of estimation methods, available in a variety of sophisticated and user-friendly computer packages, the true phylogeny will indeed not be recovered for certainty for many groups of organisms.

Acknowledgements

I want to thank R. Post for his comments on a draft of this paper.

References

- Ax, P., 1987.** *The phylogenetic system: The systematization of organisms on the basis of their phylogenies.* John Wiley & Sons, Chichester.
- Colless, D. H., 1982.** Review of phylogenetics: The theory and practice of phylogenetic systematics. — *Systematic Zoology* 31: 100-104.
- Dawkins, R., 1986.** *The blind watchmaker.* Harlow, Longman.
- Farris, J. S., 1978.** *Wagner78*; manual, documentation and a FORTRAN IV Wagner program.
- Farris, J. S., 1988.** *Hennig86, version 1.5*; Hennig86 reference manual and program.
- Felsenstein, J., 1987.** *PHYLIP; Package for inferring phylogenies, version 3.0*; manual, documentation and several PASCAL programs. University of California Herbarium, Berkely, California.
- Fiala, K. L. & R. R. Sokal, 1985.** Factors determining the accuracy of cladogram estimation: evaluation using computer simulation. — *Evolution* 39: 609-622.
- Heijerman, Th., 1988.** GENESIS: a simulation model of phylogeny. 1. The origin and early evolution of character state vectors. — *Zeitschrift für zoologische Systematik und Evolutionsforschung* 26: 409-424.
- Heijerman, Th., 1990.** GENESIS: a simulation model of phylogeny. 2. A sensitivity analysis. — *Zeitschrift für zoologische Systematik und Evolutionsforschung* 28: 81-93.
- Heijerman, Th., 1992.** Adequacy of numerical taxonomic methods: A comparative study based on simulation experiments. — *Zeitschrift für zoologische Systematik und Evolutionsforschung* 30: 1-20.
- Heijerman, Th., 1993.** Adequacy of numerical taxonomic methods: Further experiments using simulated data. — *Zeitschrift für zoologische Systematik und Evolutionsforschung* 31: 81-97.
- Huelsenbeck, J. P. & D. M. Hillis, 1993.** Success of phylogenetic methods in the four-taxon case. — *Systematic Biology* 42: 247-264.
- Kim, J., 1993.** Improving the accuracy of phylogenetic estimation by combining different methods. — *Systematic Biology* 42: 331-340.
- Kim, J. & M. A. Burgman, 1988.** Accuracy of phylogenetic estimation methods under unequal evolutionary rates. — *Evolution* 42: 596-602.

Adequacy of taxonomic methods; why not be a pheneticist?

- Kim, J., F. J. Rohlf & R. R. Sokal, 1993. The accuracy of phylogenetic estimation using the neighbor-joining Method. — *Evolution* 47: 471-486.
- Nei, M., 1991. Relative efficiencies of different tree-making methods for molecular data. In: *Phylogenetic analysis of DNA sequences* (M. M. Miyamoto & J. Cracraft, eds): 90-128. Oxford University Press, Oxford, England.
- Pankhurst, R. J., 1991. *Practical taxonomic computing*. Cambridge University Press, Cambridge, United Kingdom.
- Platnick, N. I., 1989. An empirical comparison of microcomputer parsimony programs, II. — *Cladistics* 5: 145-161.
- Quicke, D. L. J., 1993. *Principles and techniques of contemporary taxonomy*. Chapman & Hall, Glasgow.
- Rohlf, F. J., 1982. Consensus indices for comparing classifications. — *Mathematical Biosciences* 59: 131-144.
- Rohlf, F. J. & M. C. Wooten, 1988. Evaluation of the restricted maximum-likelihood method for estimating phylogenetic trees using allele-frequency data. — *Evolution* 42: 581-595.
- Rohlf, F. J., W. S. Chang, R. R. Sokal & J. Kim, 1990. Accuracy of estimated phylogenies: effects of tree topology and evolutionary model. — *Evolution* 44: 1671-1684.
- Sneath, P. H. A., 1988. The phenetic and cladistic approaches. In: *Prospects in systematics* (D. L. Hawksworth, ed.): 252-273. Clarendon Press, Oxford.
- Sneath, P. H. A. & R. R. Sokal, 1973. *Numerical Taxonomy*. W. H. Freeman, San Francisco.
- Sokal, R. R., 1983. A phylogenetic analysis of the Caminalcules. I. The data base. — *Systematic Zoology* 32: 159-184.
- Sokal, R. R., K. L. Fiala & H. Hart, 1984. OTU stability and factors determining taxonomic stability: examples from the Caminalcules and the Leptopodomorpha. — *Systematic Zoology* 33: 387-407.
- Sourdis, J. & M. Nei, 1988. Relative efficiencies of the maximum parsimony and distance-matrix models in obtaining the correct phylogenetic tree. — *Molecular Biology and Evolution* 5: 298-311.
- Swofford, D. L., 1985. *PAUP: Phylogenetic Analysis Using Parsimony version 2.4*; manual, documentation and program. Illinois Natural History Survey, Champaign, Illinois.
- Wishart, D., 1986. *CLUSTAN version 3.2. A cluster analysis package*. Edinburgh University, Program Library Unit.
- Zandee, M., 1988. *C.A.F.C.A.: A collection of APL functions for cladistic analysis, PC version 1.9*; manual, documentation and program.

Appendix 6.1 Values of input parameters for each of the 48 evolutionary settings. Evolutionary pattern: grad = gradualism; punct = punctuated equilibrium. Species diversity: rad = radiation; equi = equilibrium. Rad/grad = Scenario 1; rad/punct = Scenario 2; equi/grad = Scenario 3; equi/punct = Scenario 4.

#	Scenario	P_{ex}	P_d	P_+	P_-	$F_{compchar}$
1	rad/grad	0	-	0.030	0	0
2	rad/grad	0.03	-	0.025	0	0
3	rad/grad	0	-	0.030	0.30	0
4	rad/grad	0.03	-	0.025	0.30	0
.....						
5	rad/punct	0	-	0.017	0	0
6	rad/punct	0.03	-	0.014	0	0
7	rad/punct	0	-	0.017	0.30	0
8	rad/punct	0.03	-	0.014	0.30	0
.....						
9	equi/grad	-	1	0.011	0	0
10	equi/grad	-	10	0.012	0	0
11	equi/grad	-	1	0.012	0.30	0
12	equi/grad	-	10	0.012	0.30	0
.....						
13	equi/punct	-	1	0.006	0	0
14	equi/punct	-	10	0.007	0	0
15	equi/punct	-	1	0.007	0.30	0
16	equi/punct	-	10	0.007	0.30	0
.....						
17	rad/grad	0	-	0.030	0	0.60
18	rad/grad	0.03	-	0.025	0	0.60
19	rad/grad	0	-	0.030	0.30	0.60
20	rad/grad	0.03	-	0.025	0.30	0.60
.....						
21	rad/punct	0	-	0.017	0	0.60
22	rad/punct	0.03	-	0.014	0	0.60
23	rad/punct	0	-	0.017	0.30	0.60
24	rad/punct	0.03	-	0.014	0.30	0.60
.....						
25	equi/grad	-	1	0.011	0	0.60
26	equi/grad	-	10	0.012	0	0.60
27	equi/grad	-	1	0.012	0.30	0.60
28	equi/grad	-	10	0.012	0.30	0.60
.....						
29	equi/punct	-	1	0.006	0	0.60
30	equi/punct	-	10	0.007	0	0.60
31	equi/punct	-	1	0.007	0.30	0.60
32	equi/punct	-	10	0.007	0.30	0.60
.....						
33	rad/grad	0	-	0.030	0	0.80
34	rad/grad	0.03	-	0.025	0	0.80
35	rad/grad	0	-	0.045	0.30	0.80
36	rad/grad	0.03	-	0.052	0.30	0.80
.....						
37	rad/punct	0	-	0.016	0	0.80
38	rad/punct	0.03	-	0.014	0	0.80
39	rad/punct	0	-	0.030	0.30	0.80
40	rad/punct	0.03	-	0.034	0.30	0.80
.....						
41	equi/grad	-	1	0.027	0	0.80
42	equi/grad	-	10	0.031	0	0.80
43	equi/grad	-	1	0.059	0.30	0.80
44	equi/grad	-	10	0.058	0.30	0.80
.....						
45	equi/punct	-	1	0.023	0	0.80
46	equi/punct	-	10	0.024	0	0.80
47	equi/punct	-	1	0.040	0.30	0.80
48	equi/punct	-	10	0.039	0.30	0.80

Appendix 6.2 Tree statistics. Means and coefficients of variation (cv) for each of the 48 evolutionary settings. Sample size for each setting: $n = 50$.

#	R_1		DI		CI		l_{col}		l_{adeq}		l_{stem}		l_{stemFS}		l_{sack}	
	mean	cv	mean	cv	mean	cv	mean	cv	mean	cv	mean	cv	mean	cv	mean	cv
1	1.00	0	0.78	8	0.40	10	0.21	34	3.30	17	0.26	20	0.33	18	101	8
2	1.00	0	0.81	8	0.43	13	0.23	42	3.52	25	0.28	19	0.35	20	104	11
3	1.42	9	0.79	7	0.41	10	0.21	31	3.56	12	0.27	13	0.35	14	102	8
4	1.43	10	0.82	7	0.42	11	0.20	42	3.41	14	0.29	17	0.35	15	101	9
5	1.00	0	0.71	8	0.36	9	0.19	35	3.21	8	0.17	16	0.35	16	100	7
6	1.00	0	0.76	8	0.38	10	0.22	30	3.36	14	0.19	18	0.35	14	103	8
7	1.30	5	0.72	6	0.40	5	0.18	34	3.61	5	0.17	17	0.34	19	99	7
8	1.39	7	0.78	7	0.40	5	0.21	36	3.75	8	0.19	19	0.38	18	102	9
9	1.00	0	0.88	4	0.50	11	0.21	31	3.47	13	0.34	18	0.42	14	102	8
10	1.00	0	0.84	4	0.44	10	0.22	31	3.57	7	0.35	15	0.43	16	103	8
11	1.52	9	0.87	4	0.45	9	0.21	35	3.56	10	0.36	19	0.44	16	102	9
12	1.46	6	0.83	4	0.42	8	0.21	34	3.61	9	0.35	15	0.42	15	103	9
13	1.00	0	0.88	5	0.50	11	0.22	29	3.64	14	0.26	17	0.43	16	104	7
14	1.00	0	0.83	6	0.44	10	0.22	33	3.48	12	0.28	16	0.40	16	103	8
15	1.47	11	0.86	5	0.45	9	0.21	31	3.46	11	0.27	16	0.44	13	103	7
16	1.39	9	0.82	5	0.44	7	0.21	32	3.57	11	0.27	14	0.41	13	101	7
17	1.00	0	0.88	7	0.63	9	0.20	34	3.72	24	0.29	15	0.33	18	101	8
18	1.00	0	0.90	5	0.63	12	0.20	39	3.20	15	0.31	20	0.36	17	101	10
19	1.43	11	0.86	6	0.50	9	0.21	34	2.90	13	0.29	17	0.33	15	102	8
20	1.40	11	0.88	4	0.51	11	0.22	33	2.70	12	0.31	15	0.35	14	103	9
21	1.00	0	0.84	5	0.60	7	0.20	35	3.46	12	0.19	16	0.34	17	101	8
22	1.00	0	0.86	7	0.57	10	0.20	41	3.21	16	0.20	21	0.35	17	102	10
23	1.35	6	0.81	5	0.51	5	0.21	39	2.96	8	0.19	19	0.33	15	102	9
24	1.36	11	0.84	5	0.49	8	0.21	34	2.76	10	0.20	22	0.36	16	102	8
25	1.00	0	0.93	3	0.61	14	0.20	37	2.49	12	0.36	19	0.44	12	102	10
26	1.00	0	0.90	3	0.56	11	0.21	33	2.54	10	0.34	20	0.43	12	103	8
27	1.44	12	0.91	3	0.50	9	0.20	32	2.30	8	0.36	18	0.44	14	101	7
28	1.40	8	0.88	3	0.48	8	0.22	32	2.35	9	0.34	21	0.41	17	103	9
29	1.00	0	0.90	5	0.57	13	0.19	29	2.42	13	0.28	23	0.43	14	100	6
30	1.00	0	0.87	4	0.52	9	0.22	30	2.44	12	0.28	19	0.43	15	103	8
31	1.46	12	0.88	4	0.48	10	0.21	32	2.35	8	0.27	21	0.44	15	103	8
32	1.39	9	0.86	5	0.47	10	0.21	36	2.39	8	0.27	19	0.41	15	103	9
33	1.00	0	0.94	3	0.78	5	0.20	33	3.69	21	0.30	17	0.33	15	102	8
34	1.00	0	0.94	3	0.77	6	0.20	32	3.24	20	0.33	17	0.38	15	102	8
35	1.56	9	0.87	5	0.50	9	0.21	36	3.59	13	0.30	18	0.32	18	102	9
36	1.63	9	0.88	4	0.44	11	0.21	30	3.52	15	0.33	16	0.35	16	102	7
37	1.00	0	0.90	4	0.76	5	0.19	38	3.55	12	0.20	21	0.33	19	100	9
38	1.00	0	0.92	4	0.71	10	0.20	32	3.26	20	0.22	19	0.35	16	101	7
39	1.52	9	0.82	5	0.49	6	0.19	34	3.66	13	0.20	15	0.34	17	101	8
40	1.65	11	0.84	6	0.41	14	0.18	31	3.51	21	0.21	17	0.37	15	99	7
41	1.00	0	0.93	3	0.56	11	0.20	30	3.20	15	0.36	19	0.43	16	102	7
42	1.00	0	0.91	3	0.52	14	0.22	34	3.50	16	0.35	16	0.41	14	103	9
43	1.91	9	0.89	4	0.31	12	0.21	37	3.52	14	0.35	16	0.43	14	102	10
44	1.95	9	0.88	3	0.31	12	0.23	34	3.85	17	0.35	19	0.40	17	104	8
45	1.00	0	0.87	5	0.46	17	0.21	32	3.53	22	0.26	20	0.42	16	102	8
46	1.00	0	0.86	5	0.44	17	0.21	36	3.67	21	0.28	18	0.41	14	102	9
47	1.97	15	0.84	6	0.28	16	0.23	35	3.29	15	0.25	18	0.42	17	105	9
48	2.02	15	0.81	7	0.27	20	0.21	28	3.39	16	0.27	15	0.40	14	103	7
1-48	1.27	26	0.86	8	0.49	26	0.21	34	3.26	20	0.28	28	0.38	19	102	8
min	1.00		0.60		0.17		0.05		1.73		0.10		0.19		88	
max	2.96		0.99		0.86		0.48		6.46		0.53		0.65		136	

7

Discussion

"We have no written pedigrees; we have to make out community of descent by resemblances of any kind." (Darwin, 1859)

Phylogenetic trees are a foundation for many studies in evolutionary biology, historical biogeography and comparative biology. Also in conservation evaluation there is an increasing demand for phylogenetic information (e.g. Krajewski, 1994 and articles in Forey et al., 1994). The results and conclusions of such studies can only be as sound as the phylogenetic information on which they are based. Therefore, it is most important to have an indication of the quality of the estimation methods that are used to produce phylogenetic hypotheses in the form of phylogenetic trees. We would need to know the true phylogeny in order to tell whether our estimations are close enough approximations of it. In general true phylogenies are unknown. Therefore, in the assessment of phylogenetic estimation methods, we are forced to use simulated phylogenies of artificial species. Computer simulations allow us to study the relative efficiencies of estimation methods under a variety of evolutionary conditions.

The current simulation study showed that, under the evolutionary conditions tested, the accuracy of phylogenetic estimations was rather low. Also this study revealed that phenetic clustering procedures cannot easily be written off as phylogenetic estimation methods. On the contrary, under certain conditions phenograms may supply even more accurate phylogenetic hypotheses than do cladograms. Still, Farris (1970) was quite right to conclude that "Evolutionary interpretations of dendrograms generated by phenetic clustering procedures should ... generally be viewed with scepticism". However, the same seems to apply for cladograms produced by cladistic tree making methods.

The present study did not take into account a number of possible biases that are certainly inherent to many phylogenetic analyses. These biases include the effects of missing characters and missing taxa as well as errors in coding and polarizing of characters. It would therefore be useful to carry out simulation experiments to find out how missing data and errors in coding and character argumentation would affect accuracy.

Some simulation studies examining the effects of missing taxa and characters, were carried out by Wheeler (1992), indicating that both affect accuracy. Rohlf & Wooten (1988) found accuracy to increase with the number of characters used, for all methods tested by them. Sokal (1983) earlier concluded from his study of the Caminalcules that, when the number of characters is reduced, the decrease of the accuracy in phenograms is smaller than in cladograms and, as a result, phenograms became more accurate estimates of the true tree than cladograms. Also Kim & Burgman (1988) found accuracy to decline with a reduction in the number of characters, but the decrease in accuracy was especially clear for phenetic clustering. All simulation experiments discussed in the current study were carried out using 50 characters. The resolving power of the data (adequacy, I_{adeq}) was calculated based on these characters after additive binary coding. Only in a few experiments I_{adeq} appeared to be correlated with accuracy. Some preliminary simulation experiments were carried out to study explicitly the effects of I_{adeq} on the accuracy of the estimates. In agreement with Sokal (1983) but not with Kim & Burgman (1988), it was found that cladistic methods were more sensitive to changes in I_{adeq} than phenetic methods. However, under the evolutionary conditions of these simulations, phenetic methods still performed better than cladistic methods. Further experiments would be required to examine in more detail the effects of adequacy on accuracy of phenetic versus cladistic estimates.

In a very interesting simulation study by Kim (1993), three tree making methods were evaluated, viz. UPGMA clustering, maximum parsimony and the neighbour joining method. Kim (1993) showed that "agreement among trees estimated by different methods lends greater credibility to the estimates". The average accuracy,

measured in terms of CI_c (= CFI), of all methods was as low as 0.5382. However, when all three methods agreed with one another, accuracy increased to 0.8880 (Kim, 1993: 333). Kim's experiments were based on 5,400 data sets of 8 taxa and 50 characters each. Only in 326 out of 5,400 cases, the three methods produced identical results; in the majority of cases (3,872 out of 5,400), all three methods gave different estimates with an average accuracy of 0.4687. Kim (1993) suggested to use the degree of agreement among the different methods as a measure of the reliability of the estimated tree, and for this he proposes to use the Methods Concordance Index (MCI), calculated as the average of the CI_c values of the possible comparisons. The index was further used in a character weighting procedure, by which the average accuracy appeared to increase.

The present study showed that the majority of phylogenetic estimates are likely to be quite inaccurate; yet they are the foundation for many studies in taxonomy and comparative evolutionary biology. Computer simulation experiments are useful to evaluate the various tree making methods and have already enlarged our insight in their performance. Further simulations are needed to examine, among other things, the effects of missing data and errors in character argumentation. Future simulations should also include a larger variety of evolutionary conditions as well as a larger number of tree making methods. The problem remains however, that we will want to know how much confidence we can have in the results of a particular phylogenetic analysis. A number of recommendations can be made.

- 1) All published estimations should be accompanied by one or more measures of fit between the trees and the data like the consistency index, the retention index (Farris, 1989) or the homoplasy excess ratio (Archie, 1989).

- 2) Confidence can be understood to refer to the tree as a whole. However, the confidence of each of the nested sets of monophyletic groups in a tree (nodes) may be assessed separately. A simple way of measuring the confidence in a specific node is in terms of the number of supporting characters. Another way of assessing

confidence is by applying data randomization techniques, Felsenstein's bootstrapping procedure (Felsenstein, 1985), or some other technique (e.g. Sanderson, 1989; Lindner, 1991; Davis, 1993; Hillis, 1995, and references therein).

3) Usually a phylogenetic analysis will result in several equally parsimonious trees. These will not differ in tree length but may well differ in topology. Preferably all of these should be presented, or, if only one or a few are chosen, it should be argued why they were preferred. The number of possible trees itself might be interpreted as indicative of reliability.

4) The shortest tree is not necessarily the correct one; it is therefore recommended to consider and present also non-minimal trees, especially if the differences in tree length are relatively small.

5) Kim's simulation experiments (Kim, 1993) demonstrated that one should employ many different methods instead of a single one; the agreement between the different estimations can be considered an indication of reliability.

6) The reliability of phylogenetic hypotheses may further be increased by analyzing different, 'independent' data sets derived from the same set of taxa, e.g. molecular data versus morphological data or data from adult specimens versus data of immature stages. A high degree of correspondence between the results from the different types of data validates reliability. See e.g. Patterson et al (1993), Omland (1994), Miyamoto & Fitch (1995); Hillis (1995) and references therein.

7) As homoplasy is the ultimate trickster of parsimony, as Stewart (1993) has put it, it might be worth considering downweighting or even excluding adaptive or environmentally dependent characters. Darwin (1859) already remarked that adaptive characters are not only valueless, but resemblances caused by adaptation to similar environmental conditions, will not reveal but rather tend to conceal blood-relationship. However, the only way to detect homoplasy and thus adaptive characters, is by character analysis on a phylogenetic tree!

Phylogenetic methods, both phenetic and cladistic, are indispensable tools in taxonomy. But, as was demonstrated in this study and in a

number of papers by other authors, the results of these methods can be misleading. I share the fear that is expressed by Thompson (1994) that "... newcomers to systematics will be hyped into believing that simply by feeding a limited selection of characters into a suitably programmed computer they will discover the evolutionary history of the group in question." The proper use of phylogenetic estimation methods requires therefore, that one is not only aware of their powers, but above all, of their weaknesses and potential pitfalls.

References

- Archie, J. W., 1989.** Homoplasy Excess Ratios: new indices for measuring levels of homoplasy in phylogenetic systematics and a critique of the Consistency Index. — *Systematic Zoology* 38: 253-269.
- Darwin, Ch., 1859.** *The origin of species by means of natural selection, or the preservation of favoured races in the struggle for life.* Reprinted: Penguin Books Ltd, Harmondsworth, England, 1968, edited with an introduction and bibliography by J. W. Burrow.
- Davis, J. I., 1993.** Character removal as a means for assessing stability of clades. — *Cladistics* 9: 201-210.
- Farris, J. S., 1989.** The retention index and the rescaled consistency index. — *Cladistics* 5: 417-419.
- Farris, J. S., A. G. Kluge & M. J. Eckardt, 1970.** A numerical approach to phylogenetic systematics. — *Systematic Zoology* 19: 172-191.
- Felsenstein, J. 1985.** Confidence limits on phylogenies: an approach using the bootstrap. — *Evolution* 39: 783-791.
- Forey, P. L., C. J. Humphries & R. I. Vane-Wright (eds), 1994.** *Systematics and conservation evaluation.* Clarendon Press, Oxford.
- Hillis, D. M., 1995.** Approaches for assessing phylogenetic accuracy. — *Systematic Biology* 44:3-16.
- Kim, J., 1993.** Improving the accuracy of phylogenetic estimation by combining different methods. — *Systematic Biology* 42: 331-340.
- Kim, J. & M. A. Burgman, 1988.** Accuracy of phylogenetic-estimation methods under unequal evolutionary rates. — *Evolution* 42: 596-602.
- Krajewski, C., 1994.** Phylogenetic measures of biodiversity: a comparison and critique. — *Biological conservation* 69: 33-39.
- Lindner, H. P., 1991.** Confidence limits in phylogenies: An example from the African Restionaceae. — *Taxon* 40: 253-266.

- Miyamoto, M. M. & W. M. Fitch, 1995. Testing species phylogenies and phylogenetic methods with congruence. — *Systematic Biology* 44: 64-76.
- Omland, K. E., 1994. Character congruence between a molecular and a morphological phylogeny for dabbling ducks (*Anas*). — *Systematic Biology* 4: 369-386.
- Patterson, C., D. M. Williams & C. J. Humphries, 1993. Congruence between molecular and morphological phylogenies. — *Annual Review of Ecology and Systematics* 24: 153-188.
- Rohlf, F. J. & M. C. Wooten, 1988. Evaluation of the restricted maximum-likelihood method for estimating phylogenetic trees using simulated allele-frequency data. — *Evolution* 42: 581-595.
- Sanderson, M. J., 1989. Confidence limits on phylogenies: the bootstrap revisited. — *Cladistics* 5: 113-129.
- Sokal, R. R., 1983. A phylogenetic analysis of the Caminalcules. IV. Congruence and character stability. — *Systematic Zoology* 32: 259-275.
- Stewart, C.-B., 1993. The powers and pitfalls of parsimony. — *Nature* 361: 603-607.
- Thompson, R. T., 1994. Lies, damned lies and cladistics. Letter to *Antenna* 18: 51-52.
- Wheeler, W. C., 1992. Extinction, sampling, and molecular phylogenetics. In: *Extinction and phylogeny* (M. J. Novacek & Q. D. Wheeler, eds): 205-115. Columbia University Press, New York.

8

Summary

A simulation model of phylogeny, called GENESIS, was developed to evaluate and to estimate the qualities of various numerical taxonomic procedures. The model produces sets of imaginary species with known character state distributions and with known phylogenies. The model can be made to produce these species and their phylogenies under different evolutionary conditions.

Within GENESIS, there are two mathematical models that describe the diversification of the number of taxa. The number of taxa increases exponentially (the 'radiation' option), or according to a logistic curve (the 'equilibrium' option). As far as character evolution is concerned, GENESIS allows for two options; in the 'gradualistic' version character state changes occur in equal rates in the two daughter lineages after a speciation event; in the 'punctualistic' version these rates can be made to differ. Combining these options, GENESIS basically offers four evolutionary scenarios. The exact evolutionary conditions within each of these scenarios can be controlled by the user who must specify the values of a number of input parameters. GENESIS produces species and their phylogenies in the form of character data sets and corresponding true trees. The output is characterized by a number of tree statistics. Within each of the main evolutionary scenarios experiments were carried out, in which some input parameters were subjected to change while the others were kept constant. For the precise experimental design one is referred to the relevant paragraphs in chapters 4 - 6.

A number of cladistic and phenetic tree making methods was evaluated. The PAUP, PHYLIP, Wagner78 and Hennig86 programs were used to produce most parsimonious trees. Group-compatibility was performed with CAFCA. Four UPGMA algorithms were used to construct phenograms: UPGMA using product squared euclidian distances of unstandardized characters (UPGMA-1); UPGMA using squared euclidian distances of unstandardized characters (UPGMA-

2); UPGMA using product moment correlations of unstandardized characters (UPGMA-3) and UPGMA using product moment correlations of standardized characters (UPGMA-4).

Experiments using the most simple evolutionary scenario (combining the 'radiation' and 'gradualistic' options) showed that overall differences in accuracy were small between Wagner parsimony (PAUP, PHYLIP, Hennig86) and UPGMA-3. Parsimony with Wagner78, UPGMA-1, UPGMA-2, UPGMA-4 and especially compatibility analysis with CAFCA were shown to be inferior to these methods. The efficiency of various methods to recover the true tree, viz. Wagner78, PAUP, PHYLIP and CAFCA, depended on several tree properties, the consistency indices of both the true tree and the estimated tree being the most important ones.

When more complicated evolutionary scenario's are considered, simulation experiments showed that UPGMA based on product moment correlations of unstandardized characters, clearly produced better results than the other phenetic or cladistic methods (Wagner78, Hennig86 and CAFCA). The efficiency now appeared to be affected most importantly by the stemminess of the true tree.

A large number of phenetic procedures together with parsimony analysis, as performed with Hennig86 and Wagner78, were evaluated under a great variety of evolutionary conditions. McQuitty's similarity analysis and the average linkage method, both based on cosine- or product moment correlations of unstandardized characters, were found to perform consistently better than maximum parsimony and the other phenetic procedures.

The average accuracy of UPGMA-3, over all experiments described in chapters 4 and 5, and as measured by the consensus fork index (*CFI*), was 0.76. Hennig86, PHYLIP (MIX) and PAUP produced similar results with an average *CFI* of 0.68. In chapter 6, the four superior phenetic methods (McQuitty's similarity analysis and the average linkage method, both based on cosine- or product moment correlations of unstandardized characters) had an average accuracy of 0.74. In these experiments accuracy of maximum parsimony as performed by Hennig86 was at 0.64. Also other authors generally observed equally low or even lower accuracy values (chapter 6).

Summary

Stemminess and congruence of the characters with the tree as measured by the consistency index of the true tree, were found to be correlated with accuracy (*CFI*) in some experiments. Although, as also other authors pointed out, there may be a good correlation between an index and accuracy, there is the problem that the true tree must be known in order to compute the index. Therefore these indices cannot really be used as estimators of accuracy. Nevertheless they can serve to indicate the major determinants of accuracy. The consistency index of the estimated tree can be calculated in practice. Therefore the *CI* of the estimated tree might be used as a predictor of accuracy, though not a very reliable one. Estimated trees with low *CI* values, say less than 0.7, are probably not good estimates of the true tree.

In the present study, overall low values of accuracy were obtained. This is in agreement with the findings of a number of other authors. All simulations in the this study were run to produce 20 'species'. If we use a cutoff point of $CFI = 0.833$, where 13 of the 18 subgroups would be correctly obtained, than it can safely be assumed that most published phylogenetic estimations are likely to be quite inaccurate. Therefore I support the view of authors that it is inappropriate to refer to phylogeny *estimation* methods as methods for phylogeny *reconstruction*.

9

Samenvatting

GENESIS, een evolutie-simulatiemodel, is ontwikkeld om de eigenschappen van verschillende numeriek taxonomische procedures te evalueren en te taxeren. Het model produceert groepen soorten waarvan zowel de kenmerktoestanden bekend zijn als de fylogenie. Het model is in staat om deze soorten en hun fylogenieën te genereren onder verschillende evolutionaire condities.

Binnen GENESIS zijn er twee mathematische modellen die de diversificatie van het aantal taxa beschrijven; het aantal taxa kan exponentieel toenemen of volgens een logistisch verloop. In verband met deze diversiteitpatronen, biedt GENESIS zowel de 'radiatie'- als de 'equilibrium'-optie. Ook wat de kenmerkevolutie betreft biedt GENESIS twee opties; in de 'gradualistische' versie treden kenmerktoestandsveranderingen na een speciatie in beide dochterlijnen met even grote frequentie op; in de 'punctualistische' versie kunnen deze snelheden van elkaar verschillend worden ingesteld. Door deze opties met elkaar te combineren, biedt GENESIS vier evolutionaire hoofdscenario's. De precieze evolutionaire condities binnen elk van deze scenario's kunnen worden bepaald door de gebruiker die de waarden moet opgeven van een aantal inputparameters. Het model produceert soorten en hun fylogenieën in de vorm van kenmerkmatrices en corresponderende ware stambomen. Modelresultaten worden gekarakteriseerd door een aantal beschrijvende statistieken. Binnen het kader van elk van de hoofdscenario's zijn experimenten uitgevoerd, waarbij de inputwaarden van enkele parameters werden gevarieerd, terwijl de andere constant werden gehouden. Voor het exacte experimentele ontwerp wordt men verwezen naar de relevante paragrafen in de hoofdstukken 4 - 6.

Er zijn een aantal cladistische en phenetische schattingsmethoden geëvalueerd. De programma's PAUP, PHYLIP, Wagner78 en Hennig86 zijn gebruikt voor een maximum-parsimonie-analyse, en

CAFCA voor een compatibiliteits-analyse. Vier UPGMA technieken zijn toegepast: UPGMA gebaseerd op de gekwadrateerde euclidische afstanden van ongestandaardiseerde kenmerken (UPGMA-1); UPGMA gebaseerd op de gekwadrateerde euclidische afstanden van gestandaardiseerde kenmerken (UPGMA-2); UPGMA gebaseerd op product moment correlaties van ongestandaardiseerde kenmerken (UPGMA-3) en UPGMA gebaseerd op product moment correlaties van gestandaardiseerde kenmerken (UPGMA-4).

Experimenten uitgevoerd binnen het meest eenvoudige evolutionaire scenario, een combinatie van de 'radiatie' en de 'gradualistische' optie, toonden aan dat er slechts kleine verschillen bestaan in nauwkeurigheid tussen de resultaten van Wagner parsimonie (PAUP, PHYLIP, Hennig86) en UPGMA-3. Ook werd aangetoond dat parsimonie met Wagner78, UPGMA-1, UPGMA-2, UPGMA-4, en met name de compatibiliteits-analyse met behulp van CAFCA, minder goed presteerden vergeleken met de eerder genoemde methoden. De efficiëntie waarmee verschillende methoden, met name Wagner78, PAUP, PHYLIP en CAFCA, de ware boom wisten te benaderen, hing af van verscheidene boom-eigenschappen, waarvan de consistentie-index (*C*) van zowel de ware boom als de geschatte boom de meest belangrijke waren.

Simulatie experimenten, uitgevoerd onder meer ingewikkelde scenario's, toonden aan dat UPGMA-3 duidelijk betere resultaten produceerde vergeleken met de overige fenetische en cladistische methoden (Wagner78, Hennig86 en CAFCA). De efficiëntie bleek nu met name gecorreleerd met de 'stemminess' van de ware boom.

In een volgende serie experimenten is een groter aantal fenetische methoden getest, en ook weer parsimonie-analyse met Hennig86 en Wagner78. Uit deze experimenten, uitgevoerd onder een grote verscheidenheid aan evolutionaire condities, bleek dat de similariteits-analyse volgens McQuitty en UPGMA, beiden gebaseerd op de cosinus als similariteitsmaat of de product moment correlaties van ongestandaardiseerde kenmerken, consistent tot betere resultaten leidden dan maximum parsimonie en de overige fenetische methoden.

De gemiddelde nauwkeurigheid van UPGMA-3, berekend over alle experimenten beschreven in de hoofdstukken 4 en 5, en gemeten door de 'consensus fork index' (*CFI*), bedroeg 0.76. De resultaten van Hennig86, PHYLIP (MIX) en PAUP waren onderling vergelijkbaar met een gemiddelde waarde voor *CFI* van 0.68. In het tweede deel (hoofdstuk 6) haalden de vier beste methoden, namelijk de similariteits analyse volgens McQuitty en UPGMA, beiden gebaseerd op cosinus of product moment correlaties van ongestandaardiseerde kenmerken, een gemiddelde nauwkeurigheid van 0.74. De nauwkeurigheid van maximum parsimonie, zoals uitgevoerd door Hennig86, was 0.64. Andere auteurs vonden eveneens dergelijke lage waarden of zelfs lagere (hoofdstuk 6).

In een aantal experimenten bleken de 'stemminess' en de consistentie-index van de ware boom gecorreleerd te zijn met de nauwkeurigheid (*CFI*). Ook al vindt men echter een duidelijke correlatie tussen een bepaalde index en de nauwkeurigheid, de ware boom moet bekend zijn voordat een dergelijke index berekend kan worden. Ook andere auteurs wezen hier al op. Dergelijke beschrijvende statistieken kunnen daardoor niet gebruikt worden als schatters van de nauwkeurigheid. Desalniettemin kunnen ze ons enig inzicht geven in de mogelijke factoren die van invloed zijn op deze nauwkeurigheid. De consistentie-index van de geschatte boom kan daarentegen wel worden berekend en in enkele experimenten vertoonde deze index een correlatie met de nauwkeurigheid van de boom. Daardoor kan de *CI* van de geschatte boom wel gebruikt worden om de nauwkeurigheid te voorspellen, hoewel de voorspelling niet zeer betrouwbaar zal zijn. Geschatte bomen met *CI* waarden lager dan de arbitraire grens van 0.7 zijn waarschijnlijk tamelijk slechte schatters.

In de huidige studie werden over het algemeen nogal lage waarden gevonden voor de nauwkeurigheid van de schattingen. Dit is in overeenstemming met de resultaten van een aantal andere auteurs. Alle experimenten in de deze studie zijn zodanig uitgevoerd dat er 20 'soorten' werden geproduceerd. Indien een omslagpunt gekozen wordt bij $CFI = 0.833$, waarbij 13 van de 18 subgroepen correct zijn geschat, dan mag worden aangenomen dat de meeste

gepubliceerde fylogenetische schattingen niet erg accuraat zijn. Daarom ben ik het eens met het standpunt van sommige auteurs, dat het misplaatst is om naar fylogenetische *schattings*-methoden te verwijzen als zijnde fylogenie-*reconstructie*-methoden.

Curriculum vitae

Theodoor Heijerman werd op 11 november 1951 geboren te Aalten en in deze plaats doorliep hij tevens de Christelijke Hogere Burger School. Hierna is hij enige tijd werkzaam geweest in de zwakzinnigen-verpleging, in afwachting van de beslissing op zijn "verzoek tot erkenning van zijn bezwaren tegen de vervulling van militaire dienst als ernstige gewetensbezwaren in de zin van artikel 2 van de Wet gewetensbezwaren militaire dienst". De vervangende dienst heeft hij vervolgens, als adjunct laborant, vervuld bij het toenmalige Instituut voor Oecologisch Onderzoek (IOO), te Arnhem, Schaarsbergen. In 1973 is hij begonnen met de studie biologie aan de toenmalige Landbouwhogeschool te Wageningen. Tijdens de doctoraalfase deed hij een hoofdvak Diertaxonomie, een hoofdvak Hydrobiologie en een bijvak Entomologie. In 1980 heeft hij zijn doctoraal examen gehaald (met lof). Na een betrekking van enkele maanden bij het Centraal Bureau Nederland van het European Invertebrate Survey te Leiden, kreeg hij een aanstelling bij de sectie Diertaxonomie (vakgroep Entomologie) van de Landbouwniversiteit Wageningen, waar hij nu in deeltijd als universitair docent werkzaam is. In 1965 prikte hij zijn eerste kever aan een speld.

