Computer-Assisted Methods of Stemmatic Analysis

Robert O'Hara and Peter Robinson

In this essay, we review the methods of computer-assisted stemmatic analysis available to the *Canterbury Tales* Project.[1] Our belief that these techniques will permit us to arrive at a more exact reconstruction of the history of the *Canterbury Tales* than could Manly and Rickert (1940) is vital to our decision to undertake this work. There are two major strands to these techniques. The first, cladistic analysis, is used to gain a rapid overview of the broad relations among the manuscripts. The second, database analysis, is used to refine conclusions about the exact relationships of particular manuscripts and groups, on the basis of scrutiny of individual variants and their distribution. In addition to discussion of these techniques, we briefly report here the results of our testing of these tools on the Wife of Bath's Prologue manuscripts, among other materials.

Cladistic analysis

The collation of the manuscripts of a large medieval vernacular tradition yields enormous amounts of information concerning the agreements and disagreements among the manuscripts, even after regularization of spelling. Collation of transcripts of the Wife of Bath's Prologue manuscripts by the computer collation program *Collate*, during preliminary studies for the *Canterbury Tales* Project, supplied around 13,000 separate substantive variant readings among the forty-six manuscripts collated. With each variant occurring in an average of fifteen manuscripts, this gives about two hundred thousand separate items of information to be examined. Multiply this by all the manuscripts, then by all the parts of the *Canterbury Tales*, and we have a quantity of data far beyond the capacity of manual sorting techniques. Indeed, it appears that the inability of Manly and Rickert to devise any means of coping with this flood of information lies behind their failure to arrive at a genetic reconstruction of the manuscript tradition useful for editorial purposes.[2]

The difficulty of reconstructing manuscript stemmata, and the highly structured character of the data that result from collation, have suggested to a number of authors that computer-assisted techniques might prove valuable in pointing quickly to possible relationships which could then be thoroughly examined by other means.[3] The most successful and appropriate of these methods is cladistic analysis (from the Greek *clados* 'branch'). This technique has been developed over the last thirty years by researchers in the field of systematics, the branch of evolutionary biology which specializes in the reconstruction of the evolutionary tree of life. Cladistic analysis attempts to reconstruct the history of objects which are related in a tree of ancestry and descent, by study of the characteristics they share and do not share. Because the concepts and methods of cladistic analysis are explicitly historical in character (O'Hara 1988) they can be readily adapted to the reconstruction of manuscript stemmata.

An outline of cladistic analysis

Systematic biologists have been investigating the diversity of life for more than 200 years, and have thus far described several million species. When we examine this diversity, we see a great variety of differences among organisms: differences in size, colour, growth, ecological preferences, external and internal structure, molecular makeup, and so on. These differences have arisen and accumulated through the long course of reproduction and divergence that makes up the evolutionary past. They have arisen, in Darwin's concise phrase, through 'descent with modification' (1859). Suppose we wish to reconstruct this history of descent and modification in some branch of the evolutionary tree, the species of woodpeckers, say, or bats, or bird's-nest fungi. First we must search for what systematists call *characters*, that is, differences among the species under study that can divide them into two or more groups, and from which evolutionary events can be inferred. In the case of woodpeckers, for example, we will find that some of the 210 known species have four toes, while others have only three toes. It might seem that a character such as this, with two *states* (four-toed and three-toed), would give us evidence that there are two branches in the sought-for evolutionary tree: a four-toed branch and a three-toed branch. Reflection will show, however, that one of these states is likely to be the *ancestral* or primitive state, present in the ancestor of the whole group originally, and potentially retained unmodified anywhere. If, for example, the ancestor of all woodpeckers was four-toed (as we believe to be the case), and the three-toed state (called the *derived* state) arose and was passed on in one branch of the woodpecker tree, the collection of four-toed species would not themselves constitute a whole branch of the tree. The branch defined by the derived, three-toed state, would be *nested within* the whole tree, which elsewhere would exhibit the ancestral four-toed condition. Distinguishing the ancestral from the derived states of characters is called 'polarity determination' in the terminology of cladistics, and it is an important step, because it is only the *derived* states of characters that identify branches of the evolutionary tree.

Once a collection of characters has been described for a group under study, and the polarity of those characters has been determined, the characters are, in a sense, 'added up' to yield an estimate of the phylogeny as a whole, one that accounts for the observed distribution of character states among the descendants in the simplest manner. When the number of characters and the number of taxa (organisms under study) is large, and when some of the characters conflict with one another owing to evolutionary convergence, this can be a difficult task, as the number of possible trees that must be evaluated for their fit to the data becomes enormous.[4] It is here that computer programs are of assistance, and several tree evaluation programs have been written and are in wide use in the systematics community (see Mayr and Ashlock 1991, 320–1 for a recent listing). These programs typically search through the range of possible trees, and determine the minimum number of character state changes that could have occurred on each tree, given the particular data supplied. The tree or trees on which the fewest changes overall are required (the 'shortest' or 'most parsimonious' tree or trees) are taken to be the best estimates of the true history of taxa under study. (For comprehensive introductions to cladistic analysis see Sober 1988; Swofford and Olsen 1990; Brooks and McLennan 1991; and Maddison and Maddison 1992.)

It should be apparent from this outline that there is a fundamental identity between cladistic systematics and stemmatics. Each discipline seeks to explain the existence of a varied collection of objects (manuscripts for stemmaticists, organisms for systematists) that has resulted from a sequence of branching descents over time from a common ancestor. Accordingly, the object of cladistic analysis is nearly identical with the object of stemmatics: the reconstruction of a tree of descent based on comparative observations of the descendants themselves. The cladistic principle that only the derived states of characters identify tree branches has long been a principle of stemmatics: it was spelt out by Lejay in 1888 (reported in Kenney 1974, 135), and has been repeated by many others (West 1973, 32–3; Kane 1984, 209; cf. Quentin 1926, 61–96). Additional phenomena addressed by cladistic theory include 'one-way variation' (irreversible characters) and 'sequential variation' (multistate ordered characters), both of which are familiar to textual scholars who have long known of particular types of errors (such as omissions) which prohibit restoration of the original, as well as compound errors which must occur in a particular sequence (Maas 1958, 4; Cameron 1987).

Cladistic analysis of Robinson's Svipdagsmál *material*

The soundest justification for any method is that it works. A convincing demonstration that cladistic methods provide a powerful tool for the reconstruction of the history of manuscript traditions was provided by the cladistic analysis of Robinson's collation output for the Old Norse *Svipdagsmál* sequence performed by O'Hara in 1991 (Robinson and O'Hara 1992; forthcoming). This sequence consists of the poems *Gróugaldr* and *Fjölsvinnsmál*, together about 1500 words in length, and is extant in forty-six manuscripts written between 1650 and 1830. Figure I on p. 56 shows the stemma of the manuscripts made by Robinson using traditional philological methods, notably collection of external evidence about the relationships among the manuscripts. The sixteen manuscripts that are linked by arrows are those for which there is clear external evidence (typically scribal statements in the manuscript) that they are related as given. This external evidence provided the opportunity to judge decisively the validity or invalidity of the cladistic approach.

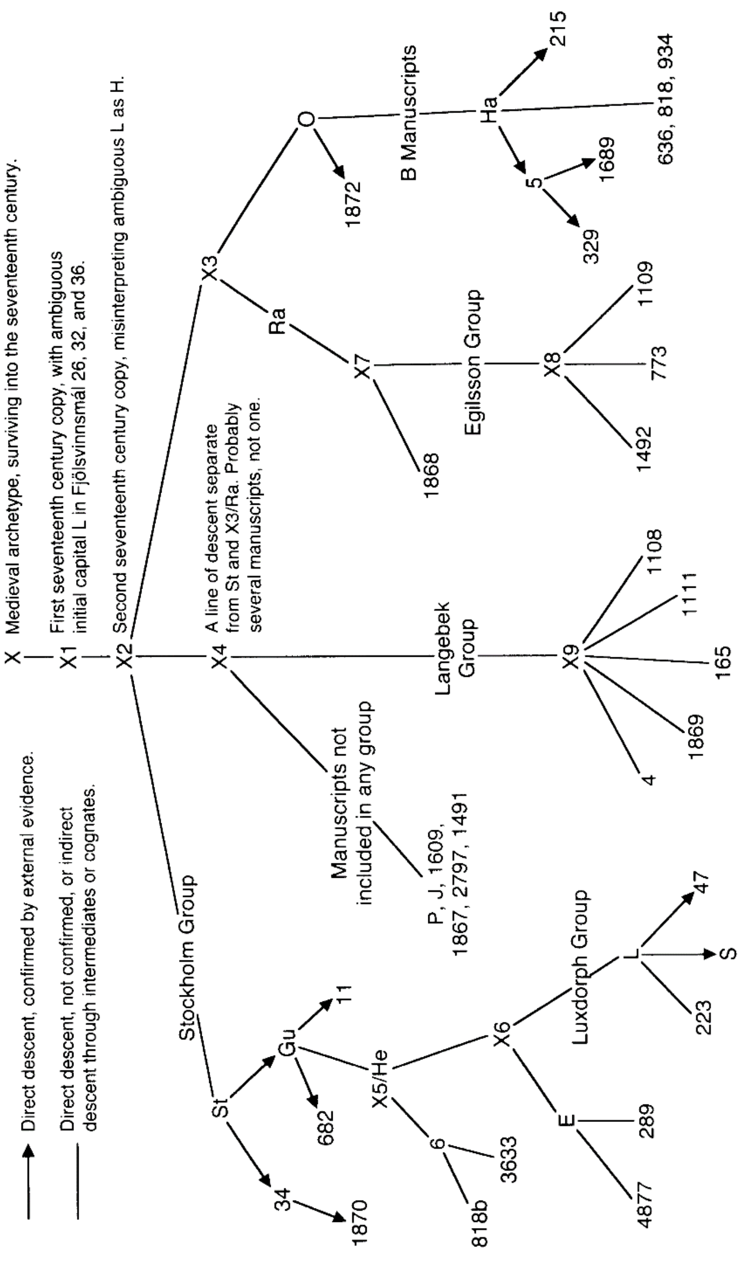

Direct descent, confirmed by external evidence.
Direct descent, not confirmed, or indirect descent through intermediates or cognates.
X Medieval archetype, surviving into the seventeenth century.
X1 First seventeenth century copy, with ambiguous initial capital L in Fjölsvinnsmál 26, 32, and 36.
X2 Second seventeenth century copy, misinterpreting ambiguous L as H.
X4 A line of descent separate from St and X3/Ra. Probably several manuscripts, not one.
X3
O
1872
B Manuscripts
Ha
215
5
1689
329
636, 818, 934
Ra
X7
1868
Egilsson Group
X8
1492
773
1109
Langebek Group
X9
1108
1111
165
1869
4
Manuscripts not included in any group
P, J, 1609, 1867, 2797, 1491
Stockholm Group
St
34
1870
Gu
11
682
X5/He
6
3633
818b
X6
Luxdorph Group
L
47
S
223
E
289
4877

Figure I: Stemma of the manuscripts of *Svipdagsmál*

Relationships of the *Svipdagsmál* manuscripts, after Robinson (1991). Branch lengths and branching angles are arbitrary, and branches may be rotated about nodes arbitrarily. Arrows indicate relationships confirmed by external evidence. X–X9 are hypothetical ancestors. Ra may be identical with X3 rather than a copy of it, and He may be either a copy of X5 or identical with X5.

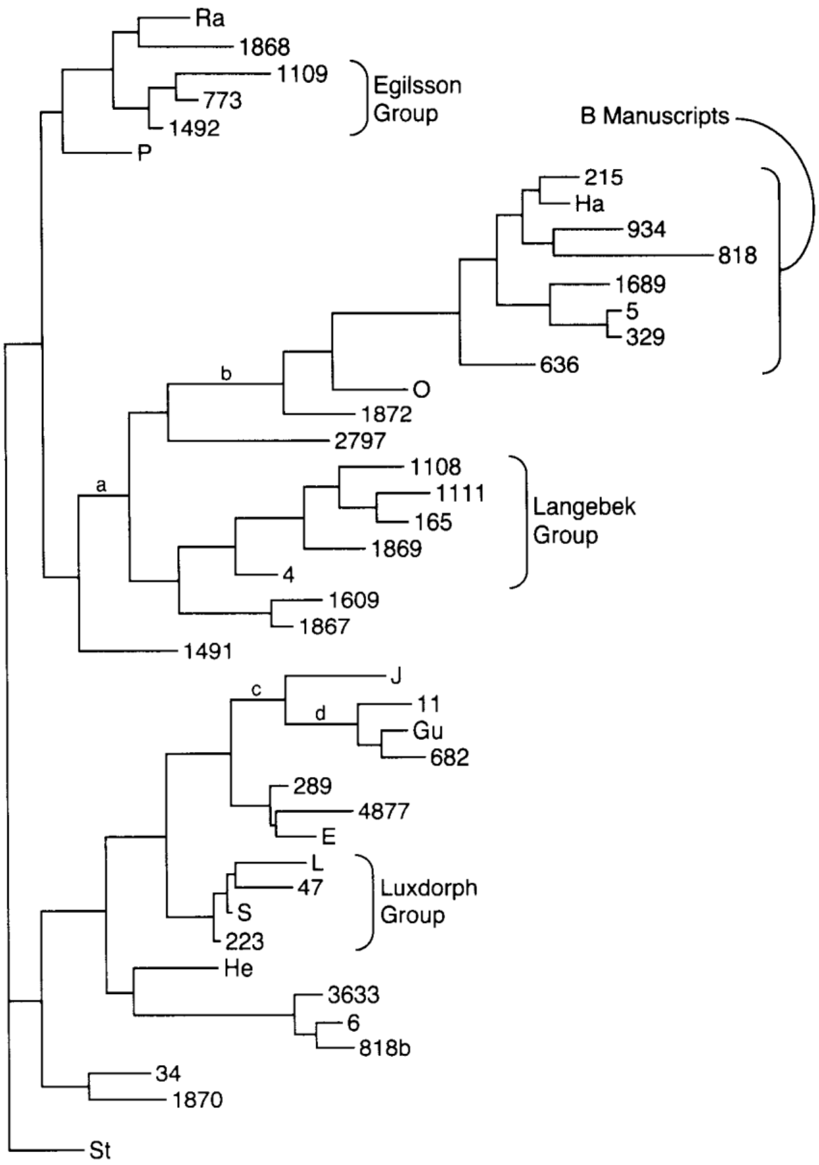


Figure II: Cladogram of the manuscripts of *Svipdagsmál*

Estimate of the history of the *Svipdagsmál* manuscripts generated by the cladistics program *PAUP*. Some of the major groupings of manuscripts common to this tree and to Robinson's stemma (Figure I) are indicated. Horizontal branch lengths are proportional to the number of character state changes along each branch. Vertical branch lengths are arbitrary, and branches may be rotated about nodes arbitrarily. See note 7 for additional details.

Our cladistic analysis of the *Svipdagsmál* sequence brought together three elements not previously joined in any study:[5]

- firstly, all the data from a complete collation of an entire manuscript tradition;
- secondly, a powerful and flexible cladistics program—*PAUP*, 'Phylogenetic Analysis Using Parsimony' (Swofford 1991);
- thirdly, a wealth of external evidence about how the manuscripts are related, evidence which could be used to test the results of the cladistic analysis.[6]

This last element, the external evidence, was particularly crucial. Previous attempts at numerical analysis have rarely been able to do more than replicate earlier non-numerically derived conclusions (for example, Moorman 1982, duplicating Manly and Rickert's 1940 results). If the earlier conclusions were themselves unsound then the later numerical efforts may not carry conviction. The collection of data subjected to cladistic analysis was in the form of a table representing all the agreements and disagreements in the *Svipdagsmál* manuscripts. No weighting of any kind was applied to the data, and no readings or groups of readings were excluded from the analysis even though there was clear evidence of substantial contamination and coincident variation.

The table in Figure II on p. 57 gives the family tree, or cladogram, for the manuscripts created by *PAUP* in its run over the raw collation data.[7] Comparison of this with the stemma in Figure I shows the accuracy with which *PAUP* replicated the outlines of Robinson's stemma. Firstly, the sixteen manuscripts which external evidence showed as directly related to one another: each of these manuscripts is placed very close to its known relative, usually adjacent. Note for example the sequence St copied to 34 to 1870, or the three manuscripts 1689, 5, and 329, all written by one scribe: these are placed directly adjacent in the cladogram just as they are in Robinson's stemma. Secondly, there are major groupings of manuscripts having relationships with one another and with key individual manuscripts. The cladistic analysis identified all these correctly. For example the Egilsson group manuscripts are placed very close to Ra in the cladogram, with one of them, 1868, separated by just one node. Without *PAUP*, establishing the closeness of these manuscripts to Ra took considerable effort. Another example can be seen in the B manuscripts, the group on the right in Robinson's stemma. After much effort without *PAUP*, Robinson had decided that the B manuscripts all descended from O, with another manuscript, 1872, descending on a different branch from O. That is very much how *PAUP*'s analysis places them in the cladogram, with O and the B manuscripts appearing as coordinate branches (called 'sister clades' in systematics), and with 1872 sister to O and the B manuscripts taken together. Consider too the three manuscripts 818b, 3633, and 6: Figure II shows these forming a subgroup of their own, within the larger Stockholm group, and that is just how they appear in the cladogram.

Robinson's study of the manuscripts St and Ra, summarized in the stemma in Figure I, revealed their fundamental importance in the evolution of the *Svipdagsmál* tradition. Some two-thirds of all the manuscripts, thirty-one of the total forty-six, appear to have descended either from St or Ra, or a manuscript (the hypothetical X3) very close to Ra. Thus, although St and Ra are very similar in absolute terms, in evolutionary terms they are far apart: they represent the twin roots from which most of the manuscripts derive. The cladistic analysis manages to show this: Figure II places five nodes between St and Ra, and from these five nodes all the other manuscripts descend.

Limitations of cladistic analysis: contamination

Although the cladistic analysis of the *Svipdagsmál* material was successful in discriminating all the major manuscript groups and in fixing the relations of some of these groups to others and to individual manuscripts, it was not correct in all details. Its greatest difficulty was caused by manuscript contamination, the deliberate importation of readings from one manuscript into another that is not its copy. Contamination takes place 'horizontally' across a stemma, rather than 'vertically' from ancestor to descendant, and cladistic analysis effectively assumes that instances of horizontal transmission will be outnumbered by instances of vertical transmission. This is broadly true of the mass of variants in most manuscript traditions, hence our general success with the *Svipdagsmál* material. But there may be subgroups of variants in subgroups of manuscripts that have been much influenced by horizontal transmission. There are, for example, a large number of variants found as marginalia in several groups of *Svipdagsmál* manuscripts which appear to have been borrowed from the text of other distinct groups, and the inclusion of these variants led to some deformation in the tree produced by the cladistic analysis. As a case in point, because of large scale contamination of the Langebek manuscripts by B manuscript readings, the Langebek manuscripts appear far closer to the B manuscripts in the cladogram in Figure II than they should. This incorrect placement of the Langebek manuscripts had other, potentially serious, consequences. Robinson's analysis of the manuscripts suggested that the B group had descended from Ra, or a manuscript very close to Ra, probably via manuscript O. The evidence for this is a set of some twenty-six readings found in Ra, also in O, and thence characteristically in the B manuscripts. In order to accommodate the Langebek manuscripts (none of which have any of these twenty-six readings) somewhere between Ra and O in the cladogram, *PAUP* had to suppose that these twenty-six readings were first removed along the branch marked *a* (hence their absence from the manuscripts below that point, including the Langebek group), and then restored along the branch marked *b* (hence their presence in the manuscripts below that point, including O and the B manuscripts). This obscures the most likely flow of readings and makes Ra, O, and the B manuscripts appear rather more distantly related than they actually are.

Limitations of cladistic analysis: coincident variation

There is a similar problem with manuscript J and the three manuscripts Gu, 11, and 682. There is strong external evidence for the close relationship of Gu, 11, and 682, with 11 and 682 both being copies of Gu, and Gu itself being a copy of St. Accordingly, the cladistic analysis succeeds in placing these three manuscripts directly adjacent to one another in the cladogram in Figure II. Robinson's account of the manuscripts has J descending quite separately from either St or Ra, and appearing among a group of manuscripts shown at the centre of the table which lack any strong affiliations with any other manuscripts. But the cladistic analysis has grouped J with the three manuscripts Gu, 11, and 682; not only this, it has moved Gu, 11, and 682 much farther away from their direct ancestor St than is correct. Examination of the variants that *PAUP* judged to have been introduced along the branch marked *c* in the cladogram revealed the reason for this error. In the second poem of the *Svipdagsmál* sequence a question formula is repeated eighteen times over. Most manuscripts abbreviate this formula in one way or other, some giving the initial letter of each word, some just giving the first one or two words, and so on. Four manuscripts alone spell out every word of the whole question sequence on each repetition: they are the four manuscripts J, Gu, 11, and 682. Clearly, 11 and 682 have simply inherited this from Gu. Clearly too, in view of the lack of any other evidence linking J and the three manuscripts Gu, 11, and 682—J has only six of the twenty-eight variants which characterize the manuscripts descended from St, while Gu, 11, and 682 have respectively twenty-five, sixteen, and twenty-five—it is simple accident that the scribe of J happened to spell out every instance of the formula just as the scribes of the other three manuscripts did. But this accident has caused the group Gu, 11, and 682 to be placed next to J and much further away from their direct ancestor St than is correct. Once more, this distorts the flow of readings: it requires us to suppose that most of the St variants present in Gu, 11, and 682 were removed before point *c*, and then restored along the branch marked *d*.

Cladistics and bifid trees

A further difficulty with the use of cladistics programs with manuscript data is that these programs tend to produce bifid trees: in the cladogram in Figure II, for example, every branch which divides always divides in two. Textual critics who recall Joseph Bédier's scathing denunciation (1928) of the tendency of textual editors to create bifid stemmata and only bifid stemmata will be dubious. It is simply not true that in the history of a manuscript tradition each exemplar is copied twice and just twice. M.D. Reeve (1986) reports eighteen textual traditions of classical Latin texts in which an archetype and more than two descendants survive. In six of these, he finds stemmata with more than two branches and no certain cases of bipartite stemmata among the other twelve. In the *Svipdagsmál* tradition there appear to have been at least three separate copyings of each of the manuscripts Gu and L, and the *Svipdagsmál* stemma as a whole grows from three distinct basal branches (see Figure I). In creating the

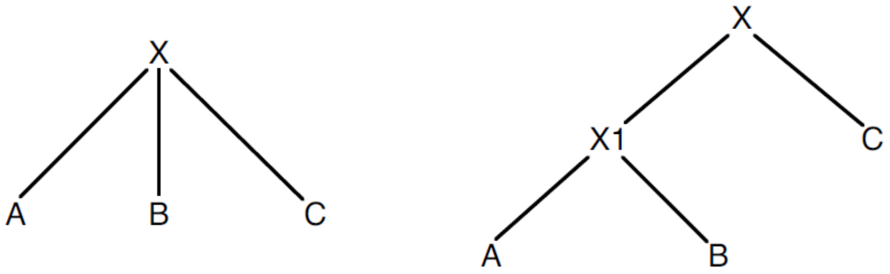


Figure III. Postulating hypothetical ancestors.

bifid tree shown in Figure II, *PAUP* was interpreting the data in a very strict sense. If three independent copies are made from a single ancestral manuscript, unless all three copies agree with each other in exactly the same number of introduced readings not present in their common ancestor, strict interpretation of the data will force the conclusion that two of these manuscripts are more closely related to one another (have a greater number of introduced readings or derived states in common) than either is to the third manuscript. Mere chance will see to it that some two of the three manuscripts will agree on a greater number of introduced readings than will either of these two with the third. In these circumstances, strict cladistic analysis will presume the existence of an intermediate ancestor for the two manuscripts sharing the greatest number of introduced readings. Thus, for three manuscripts A, B, and C all copied from a single ancestor X (Figure III, left), but with A and B having by chance a greater number of introduced readings in common than either has with C, strict cladistic analysis will generate the tree shown on the right in Figure III, hypothesizing a hyparchetype X1 as the ancestor of A and B but not of C. The textual critic must decide when this procedure is justified and when it is not. It is not argued that the application of numerical cladistic techniques obviates the need for critical thought.[8] Quite the contrary in fact: numerical cladistic analyses can provide estimates of manuscript histories very quickly, so that thought may be applied to the details of those histories with greater efficiency.

Cladistic analysis and the theory of stemmatics: ancestral readings and unrooted trees

One of the paradoxes of strict recension as formulated by Maas and others is that before one can reconstruct the stemma for a collection of manuscripts one must first determine which readings are original (ancestral) and which are introduced (derived). The obvious weakness in this procedure is that it may be impossible to determine, out of a group of variants, which variants are which. Here is one of those circles of impossibility loved by medieval story-tellers: one needs a stemma of the manuscripts before one can establish the original readings, but one cannot construct a stemma until one has established the original readings. Indeed, if one is able to establish all the original readings in

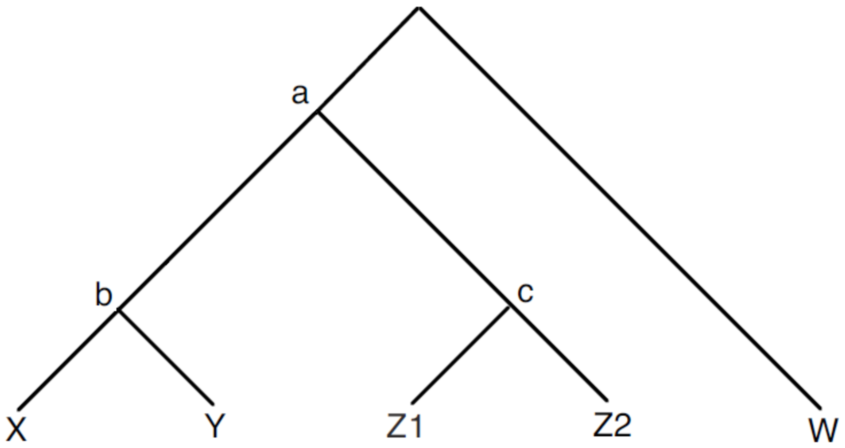


Figure IV. Stemma of five extant manuscripts (X, Y, Z1, Z2, W) rooted between W and the hypothetical archetype a.

advance, why bother, as E. Talbot Donaldson urbanely remarks, with the tedium of collation and analysis; one might as well just print the reconstructed archetype and have done with it (1970, 107; Kane and Donaldson 1975, 17 fn. 10). If it is necessary to identify all the ancestral variants before stemma construction can begin then reconstructing the history of the *Canterbury Tales* manuscripts would be impossible.

Traditional recension offers no way around these twin roadblocks, that there be just one authorial ‘original’ and that every reading in this one original be firmly identified. It may be the most remarkable contribution of cladistic analysis that it has thought through these problems and offers a way around them. Traditional stemmatics tries to identify which readings are ancestral at only one point of the manuscript family tree: the single root or ultimate ancestor. Cladistic analysis, by contrast, and in keeping with the goals of systematic biologists, aims not simply to reconstruct the ultimate ancestor but the whole history of the tradition, including the attributes of each ancestor throughout the tree. This has a remarkable and most powerful consequence. It allows systematists who are uncertain about which of their character states are ancestral and which are derived to regard their trees initially as ‘unrooted’, and then to orient those trees in any particular direction based on whatever partial information on the ancestral conditions they may have available. As a result of this practice all groupings (clades) will be defined by introduced variants no matter which way the tree is rooted, and there is no need to specify beforehand just what variants are ancestral to the tree as a whole.

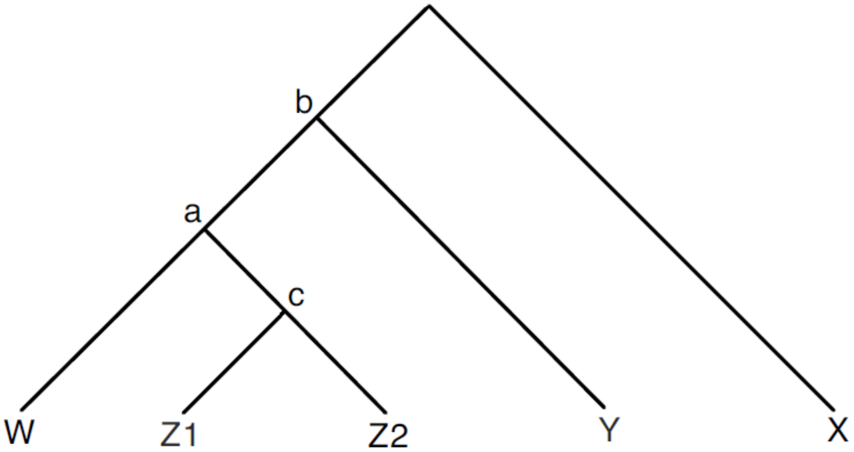


Figure V. Stemma of the same five manuscripts, rerooted between X and b.

Consider, for example, the hypothetical tree in Figure IV. Imagine that this tree is made of flexible wire and can be bent into a variety of positions. In Figure IV it is bent (rooted) such that the readings shared by 'W' and 'a' are taken to be ancestral, with the particular variants introduced in 'a' being passed on to 'b' and 'c' and their descendants. But without altering the topology of the tree in any way, or the distributions of the variants on it, we can change our judgement of the ancestral condition and reroot the tree such that the variants shared by 'X' and 'b' are ancestral, as in Figure V. Observe that the relations of the manuscripts to each other and to the nodes remain quite unchanged: manuscript 'W' is still separated from all the other manuscripts by the variants introduced at node 'a'; manuscripts 'X' and 'Y' are still linked by node 'b' and separated from manuscripts 'Z1' and 'Z2' by nodes 'a', 'b', and 'c'; 'Z1' and 'Z2' are still linked by node 'c'. In short, it does not matter initially what manuscript we decide is ancestral to the whole group. The relations of the manuscripts to each other within the tree are fixed, and do not alter however we root it. This may also provide a means of dealing with multiple archetypes. Effectively, each node within the tree is a hypothesized archetype. Most such archetypes will be the result of scribal intervention, but an intermediate archetype might also be the result of authorial revision. In the case of the second tree above, the author could have produced three different versions of the text. The first is 'b', and this is copied into manuscripts 'X' and 'Y'. The next is 'a', and this is copied into manuscript 'W'. The next is 'c', and this is copied into manuscripts 'Z1' and 'Z2'.

This is the method so far explored and proposed for the *Canterbury Tales* Project. First, one makes an unrooted tree on the basis of a table of the manuscripts and their variants, deferring judgement on just what variant readings are ancestral to the whole tree until one has this unrooted tree. Then, one can decide which of the branches of the tree lies closest to the archetype and root the whole tree near this branch. Especially, one can scrutinize the variants introduced at each hypothetical node. Where there appears a possibility that a particular group of variants introduced at a particular node might be authorial, this group can be isolated and studied in further detail.

Cladistic analysis and stemmatics: further research

Although the analyses of the *Svipdagsmál* tradition described in the previous sections as examples of cladistic analysis may seem substantial, nearly duplicating in a few minutes what it had taken several months to accomplish earlier, they were in fact very preliminary. Had the *Svipdagsmál* data been initially recorded with a view to their subsequent use in cladistic analysis, more information about the relations of different readings to one another could have been incorporated into the analysis, and the inclusion of this information would almost certainly have improved the result. It could have been specified in advance, for example, that certain readings represented omissions which were unlikely to be restored, or that other readings almost certainly arose in a particular sequence, and this information could have been taken into account by *PAUP* as it evaluated the fit of the data to different trees. The inclusion of such considerations is standard in cladistic analyses in evolutionary biology. There is some danger in including transformation assumptions of this sort in an initial analysis, of course, because it is possible to adjust the data put into the program to such an extent that whatever result one desires can be made to come out. Nevertheless, judicious incorporation of such assumptions, as long as they are clearly stated, is a reasonable procedure, and this will be explored in the *Canterbury Tales* Project. Particular attention will also be given to the problem of identifying contamination and coincident variation (see also pp. 66–9 below). The authors have met with David Swofford, the developer of the cladistic program *PAUP* used in these analyses, to discuss how *PAUP* might be extended to cope with these problems.

The systematics community has developed many other useful tools which the *Canterbury Tales* Project will review. Particularly promising is *MacClade*, developed by Wayne and David Maddison (1992). *MacClade* permits rapid access to the information behind the trees generated by *PAUP* and other cladistic programs, so that the exact degree of support for any one hypothesized manuscript family may be assessed. It also permits interactive reshaping of these trees—moving branches from one node to another, eliminating intermediate nodes, etc.—thereby allowing an investigator to develop a clearer sense of the strength of the evidential support for each of a variety of possible trees. The expertise developed during the *Canterbury Tales* Project and the links already made with the systematics community will also allow us to produce a User's Manual to be issued to all scholars interested in the application of cladistic analysis to manuscript traditions. The first version of this manual should be available by June 1994, to accompany the release of the 'Project

Edition' of *Collate 2* developed for the *Canterbury Tales* Project, incorporating the tools for translation of collation data into cladistic format.

Cladistic analysis in summary

We believe that cladistic analysis offers a powerful new tool for the exploration of the manuscript tradition of the *Canterbury Tales*:

- the theoretical basis of cladistics, assuming that a varied group of objects is the result of a sequence of branching descents over time, is as sound for manuscripts as it is for species;
- the cladistic approach to ancestral variants, seeking to identify not just the readings ancestral at the root of the tree but those ancestral at every node within the tree, is an advance over traditional stemmatic thinking and offers a means of coping with complex, multiversion texts;
- the thirty years of work in evolutionary biology on cladistic methods provides a sound body of software and published research upon which we may draw.

Especially, because it can handle (indeed, thrives upon) enormous quantities of data, cladistic analysis avoids the problems inherent in Manly and Rickert's manual attempts at stemmatic reconstruction. Because cladistics 'sees' all the data, it will not be drawn into misleading reconstructions on the basis of parts of the data only. Above all, cladistics gives a 'road-map' of the possible manuscript relations. It is not to be expected, because of the certain operation of contamination within the *Canterbury Tales* tradition, that this 'road-map' will be perfect in all details. But it will provide a starting point for further analysis by other means. It will allow analysis to begin exactly where Manly and Rickert had to finish. It took them so much effort to arrive at any sort of textual history that they were not able to go further and criticize, refine, and elaborate on that history. Cladistics will provide estimates of the history of the *Canterbury Tales* tradition with great rapidity and ease. Effort and time can then be expended on the necessary refinement, by other means, of the stemmata suggested by cladistic analysis.

Database analysis

The second analytic tool to be developed and used by the *Canterbury Tales* Project for the exploration of the manuscript tradition is database analysis of the corpus of variants. Such analysis is necessary for two reasons. Firstly, cladistic methods, powerful as they are, may only be able to reconstruct the general outlines of the relations of the manuscripts, and may not be able to determine all of the exact details. Secondly, we know from our work with the *Svipdagsmál* tradition that cladistic methods may be defective in their treatment of convergent variation (see p. 60 above).

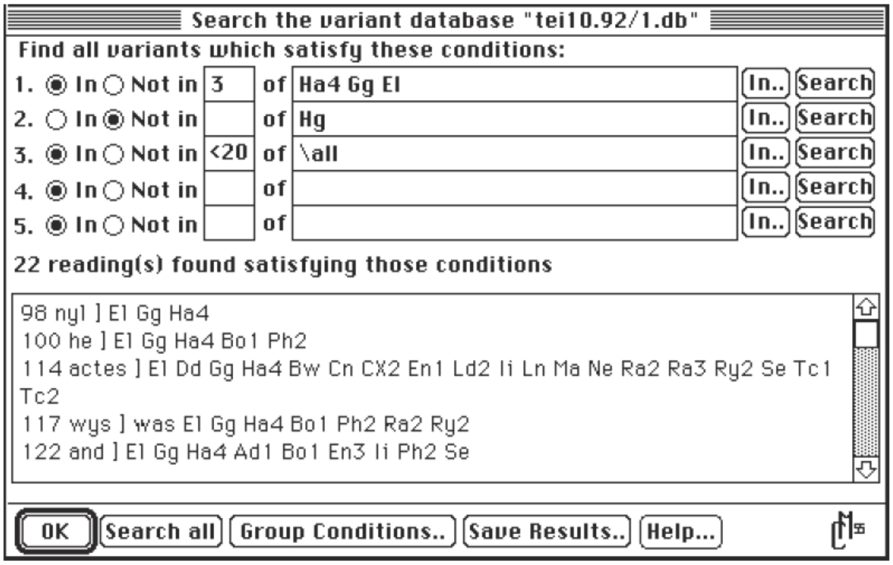
Database analysis: refining the tree

To address the details of the history of the *Canterbury Tales* manuscripts we need answers to questions like 'what variants are found in El and in Ha^4 and in no more than three other manuscripts but not in Hg'. Or, 'what readings are shared by all of Gg, Ha^4, and El against Hg', or 'by any two of these three manuscripts against Hg'. A database facility within *Collate* can be used to give answers to such questions. Figure VI shows us the readings in all three of Ha^4, Gg, and El which are not in Hg and which occur in less than twenty manuscripts for the first half of the Wife of Bath's Prologue. There are twenty-two such readings in these 428 lines. The search was repeated for the second half of the poem: there are just two such readings in this half. The readings themselves are also shown. This can be used to find out just where and how particular manuscripts, and groups or subgroups of manuscripts, agree or disagree with one another. For example: Manly and Rickert (1940, 2: 196) assert that after line 387 El shifts exemplars, leaving the archetype of Gg and joining the archetype of their 'constant group a'. The database confirms (as does preliminary cladistic analysis) that this shift does not occur. It takes about two seconds for the database to find this information; compare this with the labour of sorting necessary to find out this manually, as Manly and Rickert had to do.

Database analysis: contamination and coincident variation

Contamination, as described above, includes all cases of 'horizontal transmission' of readings, such as the deliberate importation of readings from one manuscript into another not its direct copy, rather than usual practice of 'vertical transmission' from exemplar to copy. Coincident variation, by contrast, refers to the independent introduction of the same reading into otherwise unrelated manuscripts. Cladistic analysis effectively assumes that instances of vertical transmission will outnumber instances of both coincident variation and contamination. As noted above this is broadly true of the mass of variants in most manuscript traditions, but there may be subgroups of variants in subgroups of manuscripts which have been much influenced by these factors.

That there is massive coincident variation and contamination in the *Canterbury Tales* tradition is certain: witness the number of times Manly and Rickert (1940) must appeal to 'acco' (accidental coincidence) and 'ctm' (contamination) in order to sustain their groupings. Initial cladograms of any part of the *Canterbury Tales* may be distorted by coincident variation and contamination, and so place particular manuscripts and manuscript groups too close together or too far apart. The *Canterbury Tales* Project will address these problems by means of database analysis of the corpus of variants. The theory behind the use of a database to distinguish between variants as present in a manuscript through descent, contamination, or coincidence is based on Robinson's work with the *Svipdagsmál* tradition (1989; 1991). The process is as follows:



Search the variant database "tei10.92/1.db"

Find all variants which satisfy these conditions:

1. ● In ○ Not in 3 of Ha4 Gg El
2. ○ In ● Not in of Hg
3. ● In ○ Not in <20 of \all
4. ● In ○ Not in of
5. ● In ○ Not in of

22 reading(s) found satisfying those conditions

```
98 nyl ] El Gg Ha4
100 he ] El Gg Ha4 Bo1 Ph2
114 actes ] El Dd Gg Ha4 Bw Cn CX2 En1 Ld2 Ii Ln Ma Ne Ra2 Ra3 Ry2 Se Tc1
Tc2
117 wys ] was El Gg Ha4 Bo1 Ph2 Ra2 Ry2
122 and ] El Gg Ha4 Ad1 Bo1 En3 Ii Ph2 Se
```

Figure VI. Sample database output from Robinson's *Collate* program.

- The major 'genetic groups' of manuscripts—those manuscripts having a common subarchetype—are identified. Rapid identification of these groups will be the greatest benefit of cladistic analysis, permitting quick progress to the next stage.
- The variants characteristic of these manuscript groups are identified from the database. The theory here is that a distinct group of manuscripts will be characterised by a set of readings unlikely to be archetypal and found together in significant numbers only in the manuscripts of that group (Robinson 1991, 158). Thus, for a hypothetical group of eight manuscripts, one might ask the database to extract all variants found in any five of the eight manuscripts and not found in two manuscripts thought to be close to the ultimate ancestor of the group. With the *Svipdagsmál* material, this method was found to be successful in locating the variants characteristic of the group. Specifying that variants present in manuscripts close to the common ancestor should be eliminated means that 'ancestral' variants should be disregarded and only 'introduced' variants found.
- Once these sets of characteristic readings are identified, the database is used to count, for each manuscript, how many variants the manuscript contains from each of the sets of characteristic readings. This gives a group profile of each manuscript, a snap-shot of the relations of that manuscript to each of the groups of manuscripts.

In Robinson's *Svipdagsmál* work, interpretation of these group profiles proved to be the key to coping with contamination and coincident variation. For the

Svipdagsmál tradition, grouping of the manuscripts was guided by the relative percentages of characteristic readings found in each manuscript:

- If less than 7% of the readings characteristic of a given group appeared in a manuscript, then this was likely to be the result of coincident variation and of no significance.
- If more than 7% but less than a third of the readings characteristic of a given group appeared in a manuscript, then that manuscript was likely to have been contaminated by readings from that group.
- If over half the readings characteristic of a given group appeared in a manuscript, then that manuscript was likely to belong to that group.

The imprecision of these numbers reflects that this method is a set of working hypotheses, not of ironclad rules. For every individual manuscript, and for all the manuscripts taken together, the classification this method offers must be tested by every available means. One may use these guidelines to interpret the group profile for the *Svipdagsmál* manuscript NkS 1109 fol. (Royal Library, Copenhagen) as follows:

	Readings	Text	Margin
Total readings in manuscript	1041	954	87
B text	248	15	4
St	28	0	3
Stockholm	53	4	13
Ra	34	24	1
Luxdorph	60	2	15
Langebek	97	5	0
Egilsson	29	20	7

The top line of the table gives the total number of readings in the manuscript of potential significance (i.e., not found in all the manuscripts) in the text and the number in the margin. The two left-hand columns list the seven groups of characteristic readings, and the total number of readings in each group. The two right-hand columns give the numbers of readings from each group which this manuscript has in the text or as marginalia. Then, one may read off how that manuscript stands in relation to all the characteristic groups of variants:

- 1109 fol. is not a B manuscript (it contains only 19 of the 248 B readings).
- It is not descended from Stockholm papp 8: 15, containing only 3 of the 28 St readings characteristic of that manuscript.
- There is some evidence of contamination from the Stockholm group in the 13 of 53 readings characteristics of the so-called 'Stockholm manuscripts' and found in the marginalia of this manuscript.
- It appears to be descended from Rask 21a: hence the presence of 25 of the 34 Ra readings characteristic of that manuscript.
- There is evidence of contamination from the Luxdorph group in the marginalia, with 15 of the 97 Luxdorph readings appearing there.

- The 5 Langebek variants are consistent with coincident variation.
- It is a member of the Egilsson group of manuscripts (20 of 29 variants; a further 7 in the margin).

On this evidence, the method appears capable of making fine distinctions about manuscript affiliation. It can show what group a manuscript might belong to (here, the Egilsson group); what group it might be descended from (here, the Ra group and hence, perhaps, from Rask 21a itself); what group its text has been contaminated by (here by the B text, rather lightly) and even what groups its marginalia have been contaminated by (here by the Luxdorph and Stockholm groups).

It is observed above that the arrangement of the manuscripts offered by this method must be tested by every available means. One such means is scrutiny of the distribution of sets of variants which appear particularly likely to be the result of contamination or of coincident variation. As the result of deliberate import of certain variants from one manuscript to another, the spread of the oilslick of contamination should be definable. One should be able to distinguish certain manuscripts and certain groups of manuscripts which are contaminated from those which are not. Coincident variation should not be so definable: particular variants should be found appearing quite at random in manuscripts otherwise unrelated.

These methods proved successful in Robinson's *Svipdagsmál* analysis. The *Canterbury Tales* analysis will present additional special problems. The peculiar circumstances of the early copies of *Canterbury Tales*, with just two scribes responsible for four important early manuscripts (Hg and El; Cp and Ha[4]), and the evidence of extensive 'editing' in some parts of these early copies imply a high probability of deliberate 'improvement' of the text by scribes working together with various exemplars. Attempting to recover the flow of readings through these manuscripts will be a challenging and important task. The length of *Canterbury Tales* will also create difficulties, as manuscript affiliations may shift over different regions of the text. Again, use of the database facility will help materially in meeting this challenge.

Preliminary use of these tools on the Wife of Bath's Prologue manuscripts

At the time of writing, we have transcribed and collated only forty-six of the fifty-nine manuscripts and pre-1500 printed editions of the Wife of Bath's Prologue. Our use of these analytic tools on the results of this collation has thus far been exploratory, aimed only at showing their likely utility before we committed ourselves to the immense labour of transcription of all the manuscripts of all the *Canterbury Tales*. Preliminary cladograms based on partial data suggest that there are several well-defined genetic groups among the collated manuscripts, corresponding approximately to Manly and Rickert's a, b, c, and d 'constant groups'. Two large groups of manuscripts containing Cp and Dd are similar to Manly and Rickert's c/d and a/b groups, and the constitution of these two groups is remarkably uniform whether stemmata are constructed on the basis of readings taken from the beginning or the end of the poem. But our preliminary results also indicate considerable uncertainty in the relationships among the four vital early manuscripts Ha^4, Gg, El, and Hg. Data from the beginning of the poem suggests that three of these—Gg, Ha^4, and El—are very close to one another, and distinctly separate from Hg. But analysis of data from the second half of the poem places El adjacent to Hg, away from Gg and Ha^4. Indeed, setting aside four passages present in El but not Hg, El and Hg are so similar to one another in this second half that one could be copied directly from the other. Simultaneously, Ra^2 has moved away from Ha^4. The database analysis of the relationship among the manuscripts El, Gg, Ha^4, and Hg, reported above (p. 66), supports this picture of shifting affiliation, with El apparently close to Gg and Ha^4 in the first half but apart from them and with Hg in the second. What we see in these manuscripts, and in two other important early manuscripts Cp and Dd, may be the traces of intense editorial activity. Recall that four of these manuscripts are almost certainly the work of just two scribes— Ha^4 and Cp; El and Hg—and recall too that we know these two scribes worked together on the Trinity College Gower manuscript as hands 'd' and 'b' (Doyle and Parkes 1978; cf. Ramsey 1982 and 1986). What we may have here, in the shifting patterns of affiliation in these manuscripts, is the record of an early cooperative editorial effort to recover Chaucer's text. Somewhere about here we might find, if anywhere, the register of Chaucer's own revisions. It is striking how often readings shared by Gg and El against all the other four appear superior: they may have had a joint ancestor of peculiar authority. Manly and Rickert appear to have had something of this in mind in their assertion that Gg 'represents, in the main, the El tradition without the El editing' (1940, 1: 176).

Database analysis on the results of collation of the unregularized transcripts—giving access to all the variant spellings as well as to the substantive variations in readings—has also given interesting results. It is well known that the spelling practices of Hg and El differ in some respects, and Vance Ramsay (1982) has used this variation to argue that the two manuscripts were written by different scribes. We have been able to collate the original spelling transcripts of El and Hg against four other early manuscripts: Gg, Dd, Ha^4, and Cp. Over the 856 lines of the Wife of Bath's Prologue we found some 958 places in which Hg and El agreed against the other four manuscripts—a far higher proportion of agreement than that achieved by all but one other pair of these six manuscripts.[9] These agreements are virtually all in spelling or punctuation: of the first fifty agreements of Hg/El against the other four, forty-seven are spelling, three punctuation. This remarkable and persistent agreement in these spellings of the two manuscripts against that of all other manuscripts, and especially against those manuscripts closest to them in date, must weigh in favour of the argument that the two manuscripts were written by just one scribe. It also adds to the evidence that the two manuscripts either have a single common exemplar or that their exemplars are very closely related.

One should not expect that the reconstructions of the manuscript tradition of the *Canterbury Tales* developed with these tools will always provide conclusive evidence of the originality of every reading. The nature of the evidence is such that we can deal only with probabilities, not certainties. But in the same way as it is highly probable that the Hengwrt and Ellesmere manuscripts were written in the first decades of the fifteenth century, though it is not certain, we might assert that particular relationships bearing on the history of the text are highly probable even though not certain. The information we will provide will be an additional tool in the hands of scholars which they can use to determine the weight behind this reading or that.

Notes

[1] This paper draws together elements from several papers published elsewhere. Cladistic analysis, along with an historical account of the parallels between systematics and stemmatics, is discussed further by Robinson and O'Hara (1992; forthcoming). Database stemmatic analysis is discussed further by Robinson (1989), which summarizes the treatment in his dissertation (1991). Robinson (1993) gives a fuller account of the preliminary use of these methods on the Wife of Bath's Prologue.

[2] On the failure of Manly and Rickert (1940), see Kane (1984). Kane regards their enterprise as doomed because of inherent weaknesses in the theory of recension. In a forthcoming article in *Poetica* Robinson argues that the failure was not primarily due to their theory (though this was certainly flawed) but to their practice, specifically the inability of the manual techniques they used to cope with the volume of information generated by their collation. Compare Partridge's comments, p. 87 below.

[3] The first scholar to have tried such a computer-assisted approach appears to have been John Griffith, who applied cluster analysis to some variant readings in the manuscripts of Juvenal (Griffith 1968; 1984). Most subsequent studies of this type have also used either cluster analysis or multivariate analysis (see the reviews in Hockey 1980 and Pierce 1988). Only one of these studies—that of Xhardez (forthcoming) on some fifty manuscripts of a twelfth-century text—attempted to apply statistical techniques to all the data from a complete manuscript tradition. Xhardez found that this gave a 'general but fairly accurate idea of the broad relationships between the manuscripts'. For reasons that will become apparent, we have found these statistical clustering techniques less satisfactory than cladistics.

[4] The number of possible trees for a given number of endpoints has been calculated by Felsenstein (1977) in systematics, and in parallel by Flight (1990) in stemmatics.

[5] The collaboration on this project began in July 1991 when Robinson posted a challenge to the HUMANIST electronic discussion group to discover whether anyone could duplicate his *Svipdagsmál* stemma using only the raw

matrix of agreements and disagreements among the manuscripts. O'Hara, an evolutionary biologist, responded. Prior to our collaboration, Platnick and Cameron had outlined some of these parallels in the journal *Systematic Zoology* in 1977, and Cameron had detailed them further in an excellent review published in 1987. Arthur Lee, in a paper presented to the 1987 Patristics conference in Oxford, seems to have been the first actually to apply cladistic techniques to a particular problem in manuscript studies: the relationships of some twenty-five manuscripts of Augustine's *Quaestiones in Heptateuchum* (Lee 1989).

[6] This evidence is collected in Robinson's doctoral dissertation, 1991.

[7] This tree was estimated from a matrix of 43 manuscripts by 3138 readings, of which 2063 were informative. It is the single shortest tree found by *PAUP*'s heuristic search procedure under 500 random permutations of the addition sequence, and has a length of 8249, a consistency index of 0.38, and a retention index of 0.625 (excluding uninformative characters l = 7181 and c.i. = 0.287).

[8] Researchers in systematics who have produced superficial numerical analyses of cladistic data have themselves been vigorously criticized. See, for example, the recent commentaries on the work of Cann *et al.* 1987 and Vigilant *et al.* 1991 by Maddison 1991 and Templeton 1992.

[9] The only other pair to approach such a high incidence of agreement against the other manuscripts is the pair Cp/Ha^4: we found 777 unique agreements between this pair. Cp and Ha^4 are also thought to have been written by the one scribe: see Parkes 1978, 212–15, but cf. Ramsay 1986, 126–34. The two pairs Hg/El and Cp/Ha^4 apart, the figures for unique agreements between pairs are far lower: thus 203 for Gg/Dd, 154 for Gg/Cp, 49 for Hg/Gg, 47 for Cp/Hg, 46 for Cp/El, etc.

Bibliography

Bédier, J. 1928. 'La Tradition Manuscrite du "Lai de L'Ombre": Réflexions sur l'Art d'Éditer les Anciens Textes.' *Romania* 54: 161–96, 321–56.

Brooks, D. R., and McLennan, D. A. 1991. *Phylogeny, Ecology, and Behavior: A Research Program in Comparative Biology.* Chicago: University of Chicago Press.

Cameron, H. D. 1987. 'The Upside-down Cladogram: Problems in Manuscript Affiliation.' In Hoenigswald and Wiener 1987, 227–42.

Cann, R. L., Stoneking, M., and Wilson, A. C. 1987. 'Mitochondrial DNA and Human Evolution.' *Nature* 325: 31–36.

Darwin, C. 1859. *On the Origin of Species.* London: John Murray.

Donaldson, E. T. 1970. *Speaking of Chaucer.* London: Athlone.

Doyle, A. I., and Parkes, M. B. 1978. 'The Production of Copies of the *Canterbury Tales*.' In *Medieval Scribes, Manuscripts and Libraries: Essays Presented to N. R. Ker*, ed. M. B. Parkes and A. G. Watson, 163–210. London: Scolar Press.

Felsenstein, J. 1977. 'The Number of Evolutionary Trees.' *Systematic Zoology* 27: 27–33.

Flight, C. 1990. 'How Many Stemmata?' *Manuscripta* 34: 122–28.

Griffith, J. G. 1968. 'A Taxonomic Study of the Manuscript Tradition of Juvenal.' *Museum Helveticum* 25: 101–38.

—. 1984. 'A Three-Dimensional Model for Classifying Arrays of Manuscripts by Cluster Analysis.' *Studia Patristica* XV, Part I, 79–83.

Hockey, S. 1980. *A Guide to Computer Applications in the Humanities*. London: Duckworth.

Hoenigswald, H. M., and Wiener, L. F. (Eds.) 1987. *Biological Metaphor and Cladistic Classification: An Interdisciplinary Perspective*. London: Frances Pinter.

Kane, G. 1984. 'John M. Manly (1865–1940) and Edith Rickert (1871–1938).' In *Editing Chaucer: The Great Tradition*, ed. P. G. Ruggiers, 207–29. Norman, Ok.: Pilgrim Books.

Kane, G., and Donaldson, E. T. 1975. *Piers Plowman: The B Version*. London: Athlone.

Kenney, E. J. 1974. *The Classical Text*. Berkeley: University of California Press.

Lee, A. 1989. 'Numerical Taxonomy Revisited: John Griffith, Cladistic Analysis and St. Augustine's *Quaestiones in Heptateuchum*.' *Studia Patristica* XX: 24–32.

Maas, P. 1958. *Textual Criticism*. Tr. B. Flower. Oxford: Clarendon Press.

Maddison, D. R. 1991. 'African Origin of Human Mitochondrial DNA Reexamined.' *Systematic Zoology* 40: 355–63.

Maddison, W. P., and Maddison, D. R. 1992. *MacClade: Analysis of Phylogeny and Character Evolution*. Version 3. Sunderland, Mass.: Sinauer Associates.

Manly, J. M., and Rickert, E. 1940. *The Text of the Canterbury Tales*. 8 vols. Chicago: University of Chicago Press.

Mayr, E., and Ashlock, P. D. 1991. *Principles of Systematic Zoology*. Second edition. New York: McGraw-Hill.

Moorman, C. 1982. 'Computing Housman's Fleas: A Statistical Analysis of Manly's Landmark Manuscripts in the General Prologue to the *Canterbury Tales*.' *Association for Literary and Linguistic Computing Journal* 3: 15–35.

O'Hara, R. J. 1988. 'Homage to Clio, or, Toward an Historical Philosophy for Evolutionary Biology.' *Systematic Zoology* 37: 142–55.

Pierce, R. H. 1988. 'Multivariate Numerical Techniques Applied to the Study of Manuscript Traditions.' In *Tekstkritisk Teori og Praksis*, ed. B. Fidjestol *et al.*, 24–45. Oslo: Novus Forlag.

Platnick, N. I., and Cameron, H. D. 1977. 'Cladistic Methods in Textual, Linguistic, and Phylogenetic Analysis.' *Systematic Zoology* 26: 380–85.

Quentin, H. 1926. *Essais de Critique Textuelle*. Paris.

Ramsey, R. V. 1982. 'The Hengwrt and Ellesmere Manuscripts of the *Canterbury Tales*: Different Scribes.' *Studies in Bibliography* 35: 133–54.

—. 1986. 'Palaeography and Scribes of Shared Training.' *Studies in the Age of Chaucer* 8: 107–44.

Reeve, M. D. 1986. 'Stemmatic Method: "qualcosa che non funziona."' In *The Role of the Book in Medieval Culture*, ed. P. F. Ganz, Bibliologia; 3, 57–70. Turnhout: Brepols.

Robinson, P. M. W. 1989. 'The Collation and Textual Criticism of Icelandic Manuscripts (2): Textual Criticism.' *Literary and Linguistic Computing* 4: 174–81.

—. 1991. 'An Edition of *Svipdagsmál*.' Unpublished doctoral dissertation, University of Oxford.

—. 1993. 'An Approach to the Manuscripts of The Wife of Bath's Prologue.' In *Computer-based Chaucer Studies*, ed. I. Lancashire, CCH Working Papers 3, 17–47. Toronto: University of Toronto Press.

—. Forthcoming. 'Collate: A Program for Interactive Collation of Large Textual Traditions.' In *Research in Humanities Computing* 3, ed. S. Hockey and N. Ide. Oxford: Oxford University Press.

Robinson, P. M. W., and O'Hara, R. J. 1992. 'Report on the Textual Criticism Challenge 1991.' *Bryn Mawr Classical Review* 3: 331–37.

—. Forthcoming. 'Cladistic Analysis of an Old Norse Manuscript Tradition.' In *Research in Humanities Computing* 4, ed. S. Hockey and N. Ide. Oxford: Oxford University Press.

Sober, E. 1988. *Reconstructing the Past: Parsimony, Evolution, and Inference*. Cambridge, Mass.: MIT Press.

Swofford, D. L. 1991. *PAUP: Phylogenetic Analysis Using Parsimony*. Macintosh Version 3.0r. Computer program distributed by the developer, Smithsonian Institution, Washington, DC 20560.

Swofford, D. L., and Olsen, J. 1990. 'Phylogenetic Reconstruction.' In *Molecular Systematics*, ed. D. M. Hillis and C. Moritz, 411–501. Sunderland, Mass.: Sinauer Associates.

Templeton, A. R. 1992. 'Human Origins and Analysis of Mitochondrial DNA Sequences.' *Science* 255: 737.

Vigilant, L., Stoneking, M., Harpending, H., Hawkes, K., and Wilson, A. C. 1991. 'African Populations and the Evolution of Human Mitochondrial DNA.' *Science* 253: 1503–507.

West, M. L. 1973. *Textual Criticism and Editorial Technique Applicable to Greek and Latin Texts.* Stuttgart: Teubner.

Xhardez, D. Forthcoming. 'Computer-Assisted Study of a Textual Tradition.' In *Research in Humanities Computing* 3, ed. S. Hockey and N. Ide. Oxford: Oxford University Press.