Manuscript evolution*

Christopher J. Howe, Adrian C. Barbrook, Matthew Spencer, Peter Robinson, Barbara Bordalejo and Linne R. Mooney

Frequently, letters, words and sentences are used in undergraduate textbooks and the popular press as an analogy for the coding, transfer and corruption of information in DNA. We discuss here how the converse can be exploited, by using programs designed for biological analysis of sequence evolution to uncover the relationships between different manuscript versions of a text. We point out similarities between the evolution of DNA and the evolution of texts.

Gutenberg is widely credited with the invention of printing with movable type in the 15th century. Before this development, which dwarfs the importance of the appearance of the Internet in the 20th century, scribes copied texts by hand. The popular image of these people is of monks working in a scriptorium, but that is a very limited picture. Some scribes were educated amateurs, like John Shirley (*ca*. 1366–1466), secretary to Sir Richard Beauchamp, Earl of Warwick (d. 1439), who copied books of literary works, including Geoffrey Chaucer's minor works, after Warwick's death, apparently for his own and his friends' reading entertainment[1]. Many were professionals earning a living by copying books, often to order for wealthy individuals, and copying was a significant part of the overall cost of the book, depending on how lavish it was. For example, a collection of 15th century manuscripts from Peterhouse in Cambridge includes a breakdown of the cost of producing them. The parchment was 3 or 6 pence a gathering (a folded section of sheets) depending on the size of the book, whereas the copying was 16 or 20 pence and the binding 24 to 30 pence[2]. For comparison, a day's wage for an agricultural labourer was about 3 to 6 pence (M.J. Hatcher, pers. commun.). Some of the professional scribes might have specialized in certain texts, copying them several times.

Scribes frequently made mistakes while copying a text, and corrections could be made by erasing or crossing out words or inserting corrections in the margin. However, not all the errors would be noticed, and indeed scribes would sometimes deliberately alter a text as they were copying it – perhaps in an attempt to enhance the rhythm of a poem or to 'improve' the meaning. The altered text, whether modified deliberately or accidentally, might in turn serve as a template (or 'exemplar') for other copyists and the changes would thereby be propagated.

C.J. Howe*, A.C. Barbrook and M. Spencer

Are at the Dept of Biochemistry, University of Cambridge, Tennis Court Road, Cambridge, UK CB2 1QW.
*e-mail: c.j.howe@bioc.cam.ac.uk

P. Robinson and B. Bordalejo

Are at the Centre for Technology and the Arts, De Montfort University, The Gateway, Leicester, UK LE1 5XY.

L.R. Mooney

Is at the Dept of English, University of Maine, Orono, ME 04469-5752, USA.

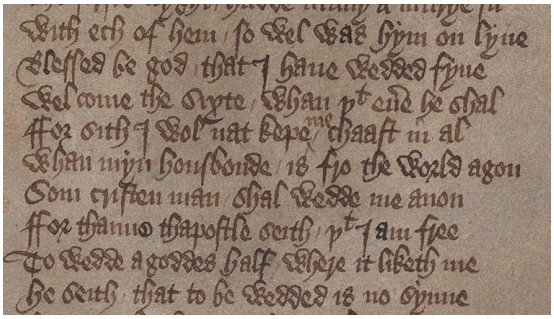

Figure 1 Folio 58 verso of the Hengwrt Chaucer, Peniarth 392D, showing lines 43 to 51 of the Wife of Bath's Prologue. Line 46 (reading 'for sith'…) appears as 'for sothe' in other manuscripts, an example of the type of variant reading exploited by phylogenetic analysis. Reproduced with permission of the National Library of Wales.

Manuscript scholars have long studied the differences among a set (or 'tradition') of extant versions of a text to try to understand how the individual versions are related. This approach, termed stemmatic analysis, or stemmatics, is often attributed to Karl Lachmann in the 19th century. It aims to construct, for a set of copies of the 'same' text, a diagram (or 'stemma', plural 'stemmata') showing how individual texts are related in terms of shared differences from the other manuscripts, and therefore which groups of manuscripts were likely to have been copied from the same template. The amounts of data manuscript scholars need to handle can be very large. For example, the Greek New Testament is represented by several thousand manuscripts. Although some progress has been made by manuscript scholars in developing computer methods for handling these datasets[3–8], they have not been widely applied.

Clearly, the model of changes being introduced during copying and then propagated in subsequent rounds closely resembles the introduction of mutations into DNA and their subsequent propagation. Similarly, the process of using comparisons between texts to infer a tree of relationships has a close parallel in the use of nucleotide or amino acid sequence data from a range of different organisms to construct a phylogenetic tree showing how they are related[9]. There is a wide range of powerful

*Commissioned by *Trends in Genetics* and published in the March issue of *TiG*.

BOX 1. ENCODING JOHN LYDGATE'S *KINGS OF ENGLAND*

This is a historical poem with stanzas describing the reigns of each of the Kings of England from William the Conqueror (1066–1087) to Henry VI (1422–1471), and beyond in some cases, and exists in over 30 different manuscript forms.

As an example of how texts can be encoded for phylogenetic analysis, we pick line 13, referring to William II (William Rufus; Figure 5), from six manuscripts and a printed version. (We have modernized letter forms where necessary, but retained the original spellings.) The line is as follows:

ffourtene yeere he bare his crowne I reede	Ashmole 59
xiiij^{e} yere he bare his crowne in dede	Bodley 48
xiiij^{e} yere bare his corone in dede	Bodley 686
ffourtene yere he bare his croune I rede	CUL Ad6686
ffourtene yer^{e} bare he his crowne in dede	Harley 2261
fortene bare hys crown in dede	Lansdowne 210
Bare th^{e} crowne xij yere xi monthes & xvi dayes in dede	de Worde (printed)

The texts are aligned for this line as follows:

ffourtene	yeere	he	bare	his	crowne	I	reede
xiiij^{e}	yere	he	bare	his	crowne	in	dede
xiiij^{e}	yere		bare	his	corone	in	dede
ffourtene	yere	he	bare	his	croune	I	rede
ffourtene	yer^{e}	bare	he	his	crowne	in	dede
fortene			bare	hys	crown	in	dede
Bare th^{e} crowne	xij yere xi monthes & xvi dayes					in	dede

The last line can be rearranged to follow the structure of the other lines as follows:

xij	yere xi monthes & xvi dayes	bare	th^{e}	crowne	in	dede

And the lines are encoded as follows (see also Table 1):

AAAAAAAHH	Ashmole 59
AAAAAAAAA	Bodley 48
AAASAAAAA	Bodley 686
AAAAAAAHH	CUL Ad6686
AAARAAAAA	Harley 2261
ALASAAAAA	Lansdowne 210
HAESRMAAA	de Worde

The coding of the last of these texts is derived as follows:
'H' indicates the change from fourteen to twelve
'A' indicates the unchanged word 'year'
'E' indicates a portion of a line that is changed (insertion of months and days)
'S' indicates omission of 'he'
'R' indicates the rearrangement of the line, shifting 'bare'
'M' indicates the substitution of 'the' for 'his', without major change in meaning
'AAA' indicates the unchanged 'crowne in dede'

methods and computer programs available to handle the sequence data used for phylogenetic inference[10], and these can be used more or less unchanged to handle manuscript data to generate credible stemmata[11]. We will describe how this can be done, and then show how several other well-documented features of the evolution of manuscript traditions have close parallels to genetic processes (Figure 1).

Phylogenetic analysis

Transcription

This is the first and most time-consuming stage. It requires access to the manuscript, ideally in its original form. Although texts are starting to become available in digitized form and thus over the web, this applies only to a tiny fraction at present. The process of digitization is slow and expensive, requiring sophisticated equipment if

good resolution is to be preserved. Furthermore, making images of manuscripts available over the web also poses copyright problems, which are not yet fully resolved.

Transcription also requires a great deal of experience in reading scribal hands, which are often hard to decipher. Indeed some scribes are recognized by the peculiarities of their handwriting, such as the distinctive form of the letter 'g' by which the so-called 'Hooked-g' scribe is identified[12]. The aim is to transcribe the text directly into an electronic file, still recognizable as text, although this might be done through a paper copy first. As well as expertise in reading scribal hands, this also requires some judgement as to what characters it is feasible to record. For example, some letters might be decorated, or be drawn in an unusual way or a different colour. In general, these differences are not used in the phylogenetic analyses that follow, but it is important to record as much data as possible in case it might be useful later[13].

Encoding

This stage aims to turn the transcribed text into a form that can be used directly as an input file by standard phylogenetic programs. This file will be a matrix of single-letter symbols in which each row is a separate text and each column a position within the text. Where texts agree at a given position, the symbol is the same, and where they differ, a different symbol is used. An example is given in Box 1. In general, each word of the manuscript corresponds to a different column in the output file, and different symbols are used to denote different kinds of change, as summarized in Table 1. Some changes, such as spelling alterations, are excluded in this protocol. In general, spelling was not systematic and a scribe might spell a word differently in different places, sometimes for trivial reasons such as making a line fit a page. Similarly, punctuation was flexible, and also is usually omitted in the phylogenetic analysis. Where a change in a word might simply reflect a local dialect, this too is omitted, as scribes working in the same geographical area might independently make the same change. So, for example, substitution of 'kirk' for 'church' would not usually be included. The process of encoding the transcribed texts in this way can be done manually, although a computer program, COLLATE, has been developed for this purpose[14].

Inferring a tree

The final stage is to use the datasets as inputs to standard phylogenetic programs. In principle, any program can be used. Early studies used parsimony as a tree recovery method[15,16], where the aim is to produce a tree requiring the smallest possible number of changes. We now most commonly use split decomposition as implemented in SplitsTree[17]. Split decomposition attempts to represent the differences between manuscripts as distances measured along a graph, while also retaining information on the amount of support for conflicting evolutionary pathways. One of the advantages of split decomposition is that it does not presuppose that the data can be fitted to a bifurcating tree. That is, it does not attempt to fit the data to a model in which one branch splits into two, and each of those can split into two more and so on. Such a model would not necessarily be appropriate to manuscripts, because a single text could be copied many times. In fact the output from SplitsTree need not be a conventional tree at all, but can be a network which allows the analysis to show signals within the data which conflict with a simple tree (whether bifurcating or not).

An example of a SplitsTree analysis of texts is shown in Figure 2, which shows a tree obtained with 43 different texts of the Prologue to the Wife of Bath's Tale in Chaucer's *Canterbury Tales* (Figure 3). Reassuringly, the manuscript groups suggested by the analysis are broadly

TABLE 1. CODING SCHEME USED FOR DATA FROM *KINGS OF ENGLAND*

Change	Symbols
Base text[a]	A
Line changed completely[b]	B, C, J, O, W[c]
Word affecting rhyme	D
Variant portion of line, changes meaning[b]	E, Y, Z[c]
Portion of line omitted[b]	F
Word variant, changes meaning	H, I, T, V[c]
Proper noun variant, changes meaning	K
Major word added/omitted, changes meaning	G, L
Word variant without change in meaning	M, N, P, Q[c]
Two (or more) words in reverse order	R
Minor word added/omitted, without change in meaning	S
Missing data	–

[a]The consensus at a given location is selected as the 'base' text [which text(s) this is does not affect the subsequent analysis], and changes in the other manuscripts are indicated.
[b]These changes are applied once at the start of the changed section and followed by X until the end of the section. For example, if the base text has 'the quick brown fox jumped over the lazy dog' and manuscript 1 has 'the lazy dog', we could code manuscript 1 as 'AFXXXXXAA' or 'FXXXXXAAA', with zero weighting given to the X characters.
[c]Some variants need several symbols, because there are some locations at which several distinct variants of the same kind occurred.

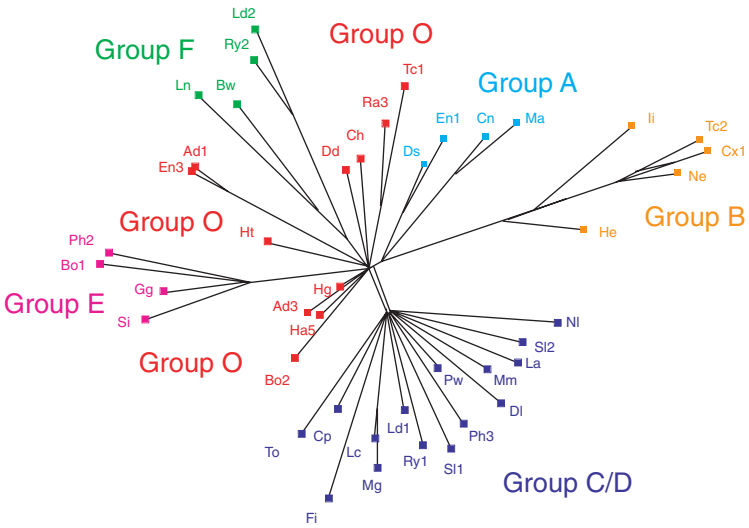

Figure 2 Analysis by SplitsTree of 43 manuscripts of the Prologue to the Wife of Bath's Tale. Individual manuscripts are indicated by two or three letter codes[15] (e.g. Bo1 is Bodley 414, Cx1 is Caxton's first printed edition, El is Ellesmere, Hg is Hengwrt and Ln is Lincoln 110). Groups of manuscripts identified in the phylogenetic analysis are marked in the same colour. Distances are a measure of the amount of difference between manuscripts. Reproduced by permission from *Nature* 394, 839, copyright 1998 Macmillan Magazines Ltd.

Figure 3 Illustration of the Wife of Bath from Chaucer's *Canterbury Tales*, GG.4.27(1) University Library Cambridge. By permission of the Syndics of Cambridge University Library.

consistent with those suggested by earlier manuscript scholars, but the phylogenetic tree is generated in a fraction of the time[11]. It seems that each group except 'O' is descended mainly from a single scribal copy. The tree shown in Figure 2 is unrooted. That is to say that it does not tell us *a priori* which the oldest manuscript of the set (and therefore probably closest to Chaucer's original) is, and it is important to note that the central point of the tree need not necessarily represent the root. However, it is interesting that the manuscript that scholars have traditionally favoured as closest to the original, Hengwrt (Hg), is close to the centre of the tree, and the analysis certainly allows us to identify others that might be close to Hengwrt (Figure 1).

Genetic parallels

The analysis described above highlights the similarity between point mutations in sequences and textual evolution. However, other genetic processes also have parallels in manuscripts.

Recombination

Some manuscripts vary in their position in a phylogenetic tree depending on which part of the manuscript is used. Figure 4 shows an example from the Prologue to the Wife of Bath's Tale. Figure 4a shows a phylogenetic tree constructed from the first half, whereas Figure 4b shows the tree based on the second half. Note that manuscript El (Ellesmere, which has been used widely in preparing modern editions of the *Canterbury Tales*) varies in its position. Analyses using parsimony and electronic databases of variant readings show that this shift is not simply owing to lack of resolution in the data[15]. The reason for the shift in the position is a phenomenon recognized by manuscript scholars for a long time – change of exemplar. The scribe used a manuscript close to the E/F group as the exemplar for the first part of the Wife of Bath's Prologue, then switched to a manuscript close to the O group. There are many possible reasons for such a change; perhaps the scribe felt that a different manuscript was more reliable for the second part. This process has obvious parallels in genetic recombination. Identifying the point where the exemplar changes by constructing sequential trees on sections of the text is clearly laborious, and work is in progress to develop computer programs to locate exemplar shifts[7]. Programs developed to detect recombination in viral evolution could be useful in this context[18]. In some cases, a manuscript can resemble a patchwork of two or more exemplars, with short sections from each interspersed. This is likely to prove very difficult to deal with, and has analogies in some cases of recombination where a large region of heteroduplex is produced between the recombining molecules and mismatches between the heteroduplex strands are resolved by the host repair machinery in different directions in different places[19].

Lateral transfer

Some texts show a more extreme form of 'recombination', resembling lateral gene transfer. Box 1 gave an example of coding texts from Lydgate's *Kings of England*. The set of texts we have used includes the following stanza referring to William I of England (William the Conqueror):

> This myghti William Duk of Normandye
> As bokes old makith mencioun
> By just title and by his cheualrye
> Made kyng by conquest of brutes Albyoun
> (British Library, Harley 2251)

There exists another poem on the kings from the same period, sometimes referred to as *Kings of England II*, but written by a different author, in which the first lines in most versions of the text are as follows:

> At Westmyster William icrowned was
> The furst day of Cristemas
> A gret thyng after he dude thanne
> Made the kyng of Skottys his legeman
> (Bodleian Library, Ashmole Rolls 21)

Some versions of this second set of texts exhibit clear evidence of lateral transfer from Lydgate's text with lines 1 and 3 from *Kings of England* transferred in, to give us verses such as this:

> This myghtty William duke of Northmandy
> That by just tytill And also by chyualery
> Conquered this land And kyng bycome
> And the kyng of Scotts he made his legeman
> (Bodleian Library, Bodley 131)

… with the rest of the text following the *Kings of England II* tradition. This change, with the reference to 'duke of Northmandy', could have been made to clarify that the reference was to William the Conqueror and not his son, William Rufus (Figure 5).

Deletions

A frequent error in manuscript copying is the occurrence of small deletions between repeated words or parts of words. The scribe probably looked away from the exemplar to the text being copied and then back to the exemplar. On looking back, the scribe's eye returned not to where it had been earlier but to the later occurrence of the same word, resulting in the omission of the text in between (such eye-skips are called 'homoeoteleutons'). For example, in the stanza on Henry III, lines 53 and 54 of Lydgate's poem on the 'Kings of England' both end in 'dede'. The scribe of a manuscript (Cotton Galba E. VIII) now in the British Library omitted the second line (line 54) probably because when he looked back at his exemplar, comparing the word 'dede' which he had just written at the end of line 53 with the word 'dede' at the end of line 54 in his exemplar, he thought he was ready to copy line 55.

So the usual sequence of lines (manuscript Cotton Titus D.XX) is:

53 Gretly delyted him in almesdede
54 Lvj yere regned he in dede
55 Buryed at Westmynser by record of wrytyng

The Cotton Galba E.VIII copy has:

53 Gretly delyted in almes ded[e]
54 [omits]
55 [Buri]ed at Westmestre be record of [wrytyng]

In genetic systems the deletion of sequence between short direct repeats is well documented, either by recombination across the repeated elements or by 'slippage replication'[20].

Convergent evolution

We have already discussed a potential example of convergent evolution among manuscripts, where different scribes working in the same geographical area might make the same changes as a consequence of their dialect. Thus 'church' might be changed to 'kirk' independently by scribes working in Scotland or Northern England, and this change has probably occurred independently in two of the *Kings of England* manuscripts. The occurrence of the same change in different lineages for whatever reason is termed convergent evolution, and if extensive it will cause unrelated lineages to be grouped together. For that reason, we have omitted dialectal changes in our analyses.

Changes in gene order?

So far we have dealt with changes to the text itself and how they can provide historical information. However, other features of a manuscript tradition might provide information; for example, where a text is divided into a number of sections with a degree of independence. This is again illustrated by the *Canterbury Tales*. This comprises a number of different tales, and, though there can sometimes be a linking text, manuscripts differ in the order of the tales. Indeed almost all extant manuscripts have a different order. In the same way, members of a cluster of genes might occur in different orders in different taxa, and methods are being devised to extract phylogenetic information from the gene order[21]. The approach we are taking with manuscripts is to derive a matrix of 'breakpoint distances' that indicates for each pair of texts the number of tales where the right-hand neighbour is different between the two. That distance matrix then forms the basis of a phylogenetic tree. Although the model for re-ordering is less explicit than for changes to the text itself, preliminary analysis suggests we might be able to extract useful information.

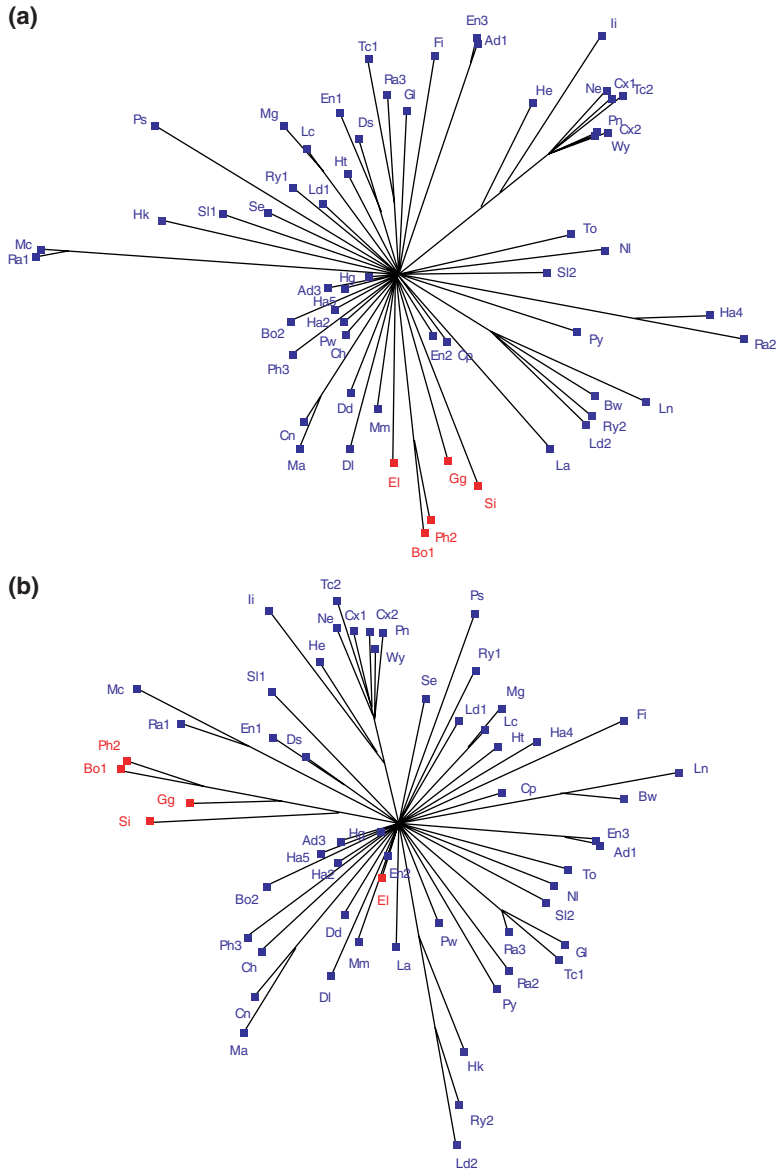

Figure 4 Shift of exemplar. (a) A SplitsTree analysis of the first half of the Prologue to the Wife of Bath's Tale; (b) the analysis of the second half. The manuscript El (Ellesmere) is placed within a group (red) in the first analysis, but its position has changed in the second.

Refinements and limitations

There is much to be done in refining the application of phylogenetic analysis to manuscript traditions. An area of particular interest is the use of weightings for different kinds of change. Are some kinds of change – such as those that significantly alter the meaning of the text – more important than others? Should more weight be given to those changes? In a simple model, all kinds of change would be consistent with the same tree topology, and the

Figure 5 British Library, Harley 4205 f.1v, stanza on and illustration of King William Rufus from the anonymous *Kings of England II*. By permission of the British Library, which owns the copyright. Further reproduction is prohibited.

topology should therefore be independent of character weightings. This prediction needs to be tested, however. If topologies are indeed independent of weightings, this will alleviate the problem of assigning rather arbitrary (and certainly potentially contentious) values. A more difficult problem, which could prove a major limitation with some manuscript traditions, will probably be that of contamination, where a single text has elements from a number of others within it. Although there is some dispute over the extent of the problem, a heavily contaminated tradition will require the application of more sophisticated phylogenetic analyses capable of dealing with, and displaying a number of, conflicting signals within a dataset. It is possible that developing better programs for stemmatic analysis will eventually prove to be useful to more conventional evolutionary biologists.

Acknowledgements

We thank the Leverhulme Trust, Churchill College, Cambridge and the Broodbank Fund for their support of this work.

References

1 Connolly, M. (1998) *John Shirley: Book Production and the Noble Household in Fifteenth-Century England*, Ashgate

2 de Hamel, C. (1992) *Medieval Craftsmen – Scribes and Illustrators*, British Museum Press

3 Weitzman, M. (1985) The analysis of open traditions. *Studies in Bibliography* 38, 82–120

4 Moorman, C. (1993) *The Statistical Determination of Affiliation in the Landmark Manuscripts of the Canterbury Tales*, Edwin Mellen Press

5 Flight, C. (1994) A complete theoretical framework for stemmatic analysis. *Manuscripta* 38, 95–115

6 Wattel, E. (1996) Clustering stemmatological trees. In *Studies in Stemmatology* (van Reenen, P.Th. and van Mulken, M.J.P., eds), pp. 123–134, Benjamins

7 Wattel, E. and van Mulken, M.J.P. (1996) Shock waves in text traditions. In *Studies in Stemmatology* (van Reenen, P.Th. and van Mulken, M.J.P., eds), pp. 105–121, Benjamins

8 Wattel, E. and van Mulken, M.J.P. (1996) Weighted formal support of a pedigree. In *Studies in Stemmatology* (van Reenen, P.Th. and van Mulken, M.J.P., eds), pp. 135–167, Benjamins

9 Cameron, H.D. (1987) The upside-down cladogram: problems in manuscript affiliation. In *Biological Metaphor and Cladistic Classification: an Interdiscplinary Approach* (Hoenigswald, H.M. and Wiener, L.F., eds), pp. 227–242, University of Pennsylvania

10 Page, R.D.M. and Holmes, E.C. (1998) *Molecular Evolution: A Phylogenetic Approach*, Blackwell Science

11 Barbrook, A.C. *et al*. (1998) The phylogeny of the Canterbury Tales. *Nature* 394, 839

12 Edwards, A.S.G. and Pearsall, D. (1989) The manuscripts of the major English poetic texts. In *Book Production and Publishing in Britain 1375–1475* (Griffiths, J. and Pearsall, D., eds), pp. 257–278, Cambridge University Press

13 Robinson, P. and Solopova, E. (1993) Guidelines for transcription of the manuscripts of the Wife of Bath's Prologue. In *The Canterbury Tales Project Occasional Papers* (Vol. 1) (Blake, N. and Robinson, P., eds), pp. 19–52, Office for Humanities Communication Publications

14 Robinson, P.M.W. (1994) Collate: interactive collation of large textual traditions, version 2, Oxford University Centre for Humanities Computing

15 Robinson, P. (1997) A stemmatic analysis of the fifteenth-century witnesses to the Wife of Bath's Prologue. In *The Canterbury Tales Project Occasional Papers* (Vol. 2) (Blake, N. and Robinson, P., eds), pp. 69–132, Office for Humanities Communication Publications

16 Lee, A.R. (1989) Numerical taxonomy revisited: John Griffith, cladistic analysis and St Augustine's Quaestiones in Heptateuchum. *Studia Patristica* 20, 24–32

17 Huson, D.H. (1998) Splitstree: analyzing and visualizing evolutionary data. *Bioinformatics* 14, 68–73

18 Holmes, E.C. *et al*. (1999) Phylogenetic evidence for recombination in dengue virus. *Mol. Biol. Evol.* 16, 405–409

19 Medgyesy, P. *et al*. (1985) Interspecific chloroplast recombination in a Nicotiana somatic hybrid. *Proc. Natl. Acad. Sci. U. S. A.* 82, 6960–6964

20 Howe, C.J. *et al*. (1988) Common features of three inversions in wheat chloroplast DNA. *Curr. Genet.* 13, 343–349

21 Sankoff, D. (1992) Edit distance for genome comparison based on non-local operations. *Lecture Notes Comp. Sci.* 644, 121–135