

CHAPTER IV: ANALYSIS OF THE TREES

PRODUCED USING PHYLOGENETIC SOFTWARE

This chapter discusses the validity of the results produced using phylogenetic programs and attempts to explain these results based on a manual analysis of particular aspects of the different tale-orders. First, I briefly survey some of the different phylogenetic approaches and point out the basic differences between them.¹ Later I explain why the programs represent the data as they do and compare the results obtained with two different encoding systems (breakpoint distance and IEBD or Inverse of Expected BreakPoint Distance).² I particularly focus on the cases in which the traditional Manly and Rickert tale-order and word-variant groups do not overlap and attempt to explain the reasons for this divergence.

1. METHODS OF PHYLOGENETIC ANALYSIS

Evolutionary biology, like any other discipline, has various ways to approach and solve a given problem, some of which are more successful than others. In this way, we find that there are many different approaches that could be used for the reconstruction of phylogenies. There are systems that produce tree-like representations of a given phylogeny while others represent the same relationships by using cycles. The latter are referred to as networks.³ My choice is to concentrate on tree-building methods. This choice is related to two main issues: in the first place, tree-building methods have similar

goals to those of the stemmatic analysis of manuscript traditions; moreover, they display their results in a similar way to the one employed by stemmatologists. In the second place, some tree-building methods, as I have explained in previous chapters, have been tried before with word-variant manuscript data with successful results, which makes them suitable to be tested with a different aspect of a textual tradition.⁴

Phylogenetic tree-building methods can be divided according to the kind of data they use (distance, discrete) and by the actual method employed by the programs for the construction of trees (clustering algorithm, search). Phylogenetic software was designed to work with deoxyribonucleic acid (DNA), but different methods approach the same problems differently. In order to understand how these programs work with manuscripts, it is necessary to understand how they work with DNA.⁵

1.1 How does Phylogenetic Software Work?

DNA is built by a sequence of nucleotides, and each nucleotide contains one of the four bases: adenine, cytosine, guanine and thymine.⁶ A DNA sequence contains the equivalent of words, but each word is written using three letters (See example below). Each of these three-letter sequences of nucleotides forms a specific amino acid. When a mutation occurs, one or more nucleotides are substituted by others, so changing the amino acid referred to by the sequence of three-letter words. If the change is successful it will be copied, individuals will inherit it, and it will become a feature of those individuals; otherwise it is just a random mutation which does not survive into further generations of copying. Similar processes can be observed in the scribal copying of

manuscripts: sometimes a change persists and is inherited by other witnesses; other times it is just random variation that gets corrected, and the variant is not inherited. Because of the similarity in these processes, there are many points of contact between the problems presented in evolutionary biology and those embedded in the study of manuscript traditions. For example, a single nucleotide could have been replaced on two successive occasions and the final result would be a single difference. Page offers the following example:

AAA

CAA

TAA (Page and Holmes 145)

The first nucleotide (A) is substituted by a second one (C), which is, in turn, replaced by a third (T). In this case, if we only knew about the first and last states, it would be impossible to infer the second state (unless the second step was necessary to reach the third one⁷). This problem is central for evolutionary biologists; and it is also relevant for the study of manuscript traditions. In evolutionary biology, most approaches deal with these data in either of two ways: distance methods and discrete methods.

1.2 Different Tree-Building Methods

1.2.1 Distance vs. Discrete Methods

Distance methods have their basis on the assumption that if the number of changes between the elements being compared were known to us, we would be able to calculate

the total number of these and reconstruct the complete sequence of changes (Page and Holmes 1979). In the proposed case, where two substitutions result in a single difference, we would need to know how many changes occurred between the first state and the third state. Although there are many difficulties involved in calculating this, several good methods have been developed.⁸ Another problem which is usually mentioned while referring to distance methods is that they have as their first step the conversion of the data into a distance matrix and as their second step they build trees based on such matrices. Of course, there are occasions in which distance methods are the most appropriate ones, especially when the data one is working with have not been produced by an automated system. Minimum evolution and neighbour joining are methods which employ distance measurement before processing the data.

Discrete methods consider the data directly in each nucleotide site instead of attempting to calculate the number of changes in a sequence as distance methods do. The main difference between these two approaches is that the discrete methods “endeavour to avoid the loss of information that occurs when sequences are converted into distances” (Page and Holmes 1979). Indeed, it would seem that using directly the data without further processing should be a more straightforward approach and should yield undistorted results. Maximum likelihood and maximum parsimony are methods which consider the data directly, without any further processing. About these methods Page and Holmes observe:

The two major discrete methods are maximum parsimony (MP) and maximum likelihood (ML). Maximum parsimony chooses the tree (or trees)

that require the fewest evolutionary changes. Maximum likelihood chooses the tree (or trees) that of all trees is the one that is most likely to have produced the observed data. (187)

Of course, there are some disadvantages in both methods. For example, because maximum parsimony gives preference to the tree with the least number of changes, it is liable to leave out more complex alternatives which might have occurred in the real evolutionary path. In theory at least, because maximum likelihood produces trees that explain the extant data in the most satisfactory way, stemmata built using this method should be very useful to map the relationships between the witnesses which have reached us.⁹

1.2.2 Clustering vs. Search Methods

The second way in which phylogenetic methods can be classified is in regard to the procedure employed to build the trees. As stated above, one of the tree-building methods is clustering analysis. This uses algorithms to construct the trees, that is, it uses a series of steps represented in mathematical formulae to reach its results. Page synthesises some of the positive aspects of clustering analysis by pointing out that:

Clustering methods have the advantage of being easy to implement, resulting in very fast computer programs. Furthermore, they almost always produce a single tree. This combination of speed and an apparently unambiguous answer is naturally very appealing, and accounts for much of the sustained popularity of clustering methods. However, they have some severe limitation

as analytical tools. The result obtained from simple clustering algorithms often depends on the order in which we add the sequences to the growing tree. (Page and Holmes 174)

It seems clear from the above quotation that there are also serious concerns about the use of clustering methods. In fact, when one has used search methods and has seen the software evaluate more than 400,000 possible branch rearrangements before arriving at the best solutions, it is difficult to accept that any method could produce a single tree (although it is conceivable that one could arrive at a best tree). On the other hand, search methods (minimum evolution, maximum likelihood and maximum parsimony) may yield many equally good trees from a single data set. However, because so many trees can be produced, each of these methods has to determine which of the trees are the best ones. Minimum evolution searches for the tree with the shortest overall length, that is the one that requires the minimum number of changes to be produced. In a way, minimum evolution is related to maximum parsimony, but the former builds its trees by using pairwise distances, while the later is a discrete method.

1.2.2.1 Testing the Tree-building Methods

Scientists agree in saying that the true tree is very difficult to identify, so they have used models which attempt to show the accuracy of the diverse methods by reconstructing trees which are very similar to the proposed tree or to a known tree (Nei and Kumar 109). This means that a known phylogeny could be used to test the validity of

a given tree-building method. However, known phylogenies are as rare as known manuscript traditions and this poses a problem for the testing of software. One possible way to overcome this problem is to use phylogenies that have been produced in laboratories, but even with these the software has been shown not to be able to reconstruct the relationships perfectly and completely. It has been pointed out, especially, that neighbour joining and minimum evolution present problems while handling branches of length zero:

[Z]ero length interior branches in realized trees are the source of topological errors in reconstructed trees, particularly when there are many such branches. Unfortunately, we usually do not know such interior branches in real data, and therefore it is difficult to evaluate the effect of this factor, though parsimony methods are capable of identifying such branches under certain conditions. (Nei and Kumar 111)

Nei and Kumar suggest that factors such as the mentioned branches with length zero are likely to generate problems even in a known phylogeny.¹⁰ Known phylogenies, however, could be useful to test the degree of accuracy of the different methods. This represents a problem when dealing with manuscript traditions, since we could potentially have an internal branch of length 0 (a manuscript from which we know others were copied). This difficulty is not insurmountable, especially if a stemma is taken as a representation of reality. In fact, currently Matthew Spencer and Elizabeth A. Davidson are working on a pilot experiment of an artificial manuscript tradition. In their first attempts, Spencer and Davidson have included all the witnesses of the artificial tradition. The tree built using

these data is extremely accurate and reflected all the divisions and subdivisions where the researchers expected these, that is, where it was known that breaks and subdivision of the tradition had occurred. Further study of artificial textual traditions, in which parts of the tradition are suppressed, might be helpful to give a clearer idea of how accurate the programs are depending on how much data has been lost, since these factors could be monitored and controlled.¹¹

1.2.3 Split Decomposition

Another method which should also be considered is split decomposition (Dress, Huson and Moulton). Splits represent the relationships of a data set by arranging the different elements in two groups, therefore creating a partition. If the splits are compatible or suggest a derivation there would be no problem. However, when the data is conflicting, as it would be the case with highly contaminated manuscript traditions, then the scholar (or the software) has to decide which one of the splits is to be followed. Split decomposition expresses evolutionary relationships differently than other methods: it does not assume that evolution is a “*branching* or *tree-like* process” (Hudson). For this reason, split decomposition offers the advantage that the data does not have to be drawn as a tree at all. Hudson describes it as follows:

In contrast to methods such as maximum parsimony and maximum likelihood that reconstruct phylogenetic trees by optimizing certain parameters, split decomposition is a transformation-based approach. Essentially, evolutionary

data is transformed or, more precisely, "canonically decomposed", into a sum of "weakly compatible splits" and then represented by a so-called splits graph. For ideal data, this is a tree, whereas less ideal data will give rise to a tree-like network that can be interpreted as possible evidence for different and conflicting phylogenies. Further, as split decomposition does not attempt to force data onto a tree, it can provide a good indication of how *tree*-like given data is. ("Analyzing")

The above quotation explains how this method can be of help when the data is conflicting (such as in the cases of contamination or convergent variation). Given conflicting data, split decomposition can offer a different alternative for its representation. Because other methods assume that the evolutionary data is tree-like they try to impose this format upon them *SplitsTree* (Hudson), which implements split decomposition, when faced with ambiguous or contaminated data represents them as networks.

2. PHYLOGENETIC SOFTWARE AND THE ORDER OF THE *CANTERBURY TALES*

The fact that there are so many choices and different approaches to the reconstruction of phylogenies indicates that evolutionary biology is far from reaching a consensus about which method or software is the best.¹² The choice among all of these is difficult, but the main reason for the choices of software used in this research was the fact that these had already been tested by Robinson with the *Svigdagsmál* and the *Canterbury Tales* traditions. Both *SplitsTree* (split decomposition) and *PAUP* (originally designed as

maximum parsimony, but now also implementing maximum likelihood and distance methods) have been shown to be reliable with textual data, but in fact the latest version of *PAUP* allows the use of methods other than parsimony.

When using word-variant data, *Collate* (Robinson), one of the computer programs used at the Canterbury Tales Project, can encode them to be used with either *PAUP*, *SplitsTree* or any other phylogenetic software. In the case of the tale-order data, as was mentioned in chapter 2, the modified Manly and Rickert table had to be made computer readable. A special program was written by Spencer (Spencer et al.) to interpret the table and calculate the distance between the different tale orders (sequences); that is, the program can carry out calculations of how many changes have to occur in the tale-order of a particular manuscript so it can become that of another. The process originally used to encode the data is described as follows:

We calculated pairwise distances between manuscripts based on differences in item order. The natural choice for a distance measure is edit distance, the minimum number of editing operations (here insertions/deletions, adding/removing one or more items; and transpositions, moving one or more items to a new location) needed to convert one order into another. . . . We therefore used scaled breakpoint distance between the items common to each pair of manuscripts. . . . Breakpoint distance is the number of items common to both manuscripts and having different right-hand neighbours. (Spencer et al., “Gene Order”)

When there is missing data, it is not possible to know how many possible different right-hand neighbours might be in the data. For example, Gg has the sequence CL-L13-L14...ME...SQ...FK, due to loss of leaves. If someone tried to measure the difference between this manuscript and El, based in this data, he or she would have to speculate how many changes occurred between the El order to reach the Gg order (or vice versa). The method devised by Spencer had to cover two kinds of situations: when manuscripts have many common items which are now missing, or when they have fewer common items missing. This is what generated the lower and upper bound data. These two data sets differ in the fact that: “[t]he lower limit occurs when no common items were lost and the upper limit is approached if there are many lost common items” (Spencer et al., “Gene Order”). However, breakpoint distance “is only reliable when the number of transpositions is small” (Spencer et al., "Analizing" 102). In order to obtain more reliable data IEBP (Inverse of Expected BreakPoint Distance) was used.¹³ Both methods are described in “Analyzing the Order of Items in Manuscripts of *The Canterbury Tales*” (Spencer et al. 98-102).

2.1 *SplitsTree*'s Tale-Order Stemmata

In the first instance the encoded tale-order data was fed into *SplitsTree*.¹⁴ The first results produced by the program can be seen in plates1 and 2.¹⁵ The graphs are not only uninformative, but also very different from the trees that one would expect. As stated above, when faced with data deemed non tree-like, *SplitsTree* would not force a tree shape on them; however, the nature of these trees indicates that there might be some other

reason for this. In fact, *SplitsTree* gives a warning ‘Non metric: Triangle inequalities are not satisfied.’ In order for distances to be metric, they have to fulfil four mathematical requirements: non-negativity, symmetry, triangle inequalities and distinctness.

The third property is the *triangle inequality*, which states that the dissimilarity between any two sequences cannot exceed the sum of the dissimilarities between each sequence and a third. (Page and Holmes 25)

What this means is that the data can be represented only if it can be built into triangles. See (Page and Holmes 24 and ff.). The command ‘Force triangle inequalities’ alters the distance matrix in order to make it satisfy the triangle inequality. This command allows the tree to be re-drawn by completing any distances which were lacking in the construction of the previous graph. The result of forcing the triangle inequalities can be seen in plates 3 and 4. There is a slight difference in the fit for these two plates, but if they are superimposed the differences are extremely difficult to detect. These differences are reflected by small changes in the length of some of the branches. The problems of *SplitsTree* attempting to handle the tale order data can be summed up as follows: on one hand, even when forcing the triangle inequalities, the trees are not very helpful in interpreting the relationships among the data. On the other hand, because forcing the triangle inequalities alters the data, trees built using this command might not be reliable in reflecting the true nature of the relationship between the manuscripts.

Even if *SplitsTree* were successful in the analysis of word-variant data, the way in which this program handles the tale-order data is not very efficient.¹⁶ For this reason, the trees built using this software are not helpful for the interpretation of the relationships

between the different tale orders, and an alternative method to reconstruct trees using the tale-order data was required.

2.2 PAUP and Tale-Order Stemmata

Since 1992, the time when *PAUP* was first used by Robinson to construct computer generated trees for the Norse text *Svigdagsmál*, the program has been altered to deal with several alternative phylogenetic approaches. In fact, *PAUP* 4.0b10 allows the construction of trees using maximum parsimony, maximum likelihood and distance methods as stated above. The modified Manly and Rickert table was made computer readable by calculating distances (breakpoint and IEBP) and, for this reason, this is the method used to build trees with *PAUP*. The criterion for the choice of trees is minimum evolution, which, as I have pointed out before, has some links in principle to maximum parsimony but is a distance method.

2.2.1 Stemmata Based on Data Encoded Using Breakpoint Distance

PAUP gives a single tree for each of the nexus files,¹⁷ one for upper bound and one for the lower bound data. The overall architecture of these trees is very similar, although we can find some alterations in some of the relationships in the branches of the trees. Manuscripts belonging to Manly and Rickert's tale-order groups with the exact same order, such as En1 and Ds or Cn Ma En3 and Ad1, are placed with E1 in both the upper

and lower bounds trees (See plates 5 and 6). This group of manuscripts is in the same position, relative to the rest of the analysed witnesses, in both trees (it appears at the top of the stemmata). The only exception in constant position is Gg, a manuscript that has many missing leaves, when the breakpoint method has been designed to take into account the loss of common items or their preservation. If one assumes that no common items were lost (lower bound tree, see plate 6) Ps appears nearer to the majority of the **a** manuscripts and apart from the group formed by To Gg Hk Se and Ch. Ps was classified as anomalous by Manly and Rickert, but its only differences from the **a** group are the position of CL, which is separated from ME and placed between ML and WPB, the lack of TM, the lack of L14 (as Ad3 and Ha5) and the loss of L37 PA and RT because of loss of leaves in the manuscript. This close inspection of the order of this manuscript shows that its order could easily be related to that of the **a** manuscripts. Moreover, on the assumption that common elements were lost in the witnesses, Ps still groups near **a**, but it appears closer to Gg and Hk. In the upper bound tree (plate 5) these three manuscripts separate from Se To and Ch which now form a small unit farther from the **a** group.

Another major difference between plates 5 and 6 is the position of S11 Py and Ld2. In the lower bound tree (plate 6), these manuscripts appear grouped with Pw Mm and Ph3, towards the middle left side of the stemma. However, in the upper bound, S11 Py and Ld2 move (as a group) form part of the **b d** cluster. Once more, this is due to the fact that some elements have been lost. Both groups of manuscripts still appear in the indistinct cluster formed by the **b** and **d** groups (a lack of distinction which had been

suspected by Dempster when she suggested that both of these orders were derived from a common exemplar [see Dempster 1123-42]).

A very small difference between the two stemmata is the relative position of Hg and Ra1, since both manuscripts appear closer in plate 5, separated by a single node, while plate 6 shows two nodes between them. Ra1 is an incomplete manuscript which has lost many leaves, and its main characteristic is the separation of NP from the rest of the fragment VII.

In both trees, Ch is separated from Ha4 by a single witness: Wy. The significance of this is that it is possible that the source for the order in Wy might have been the ancestor of Ha4.¹⁸

This summary of differences and similarities between the two trees suggests that they are equivalent in many features. However, I have not yet offered any explanations about the possible reasons for the relative positions of the manuscripts. If we take as an example the upper bound unrooted cladogram (plate 7), we can clearly see Manly and Rickert's **a** group at the top of the tree. Dd, one of the **a** group witnesses with the most missing items appears between the **a** group and the rest of the witnesses. The reason for this is the encoding of the data, which was done without presuming the text that might have existed in Dd's missing leaves. For this data set, even though one might think that the Dd order when the manuscript was intact was very similar to that of E1, missing leaves have been encoded as items not present.

Gg, on the other hand, wants L10, L11, L15, L17, L20, L21, L24, L28, L30 and RT. One cannot be completely certain that these are indeed the only missing items in the

manuscript, but a combination of tale-order and codicological analysis might indeed suggest that the missing leaves might have contained these links.¹⁹ Gg appears between Hk and Ps (classified as anomalous by Manly and Rickert).²⁰ However, as explained above, the Ps order seems to be closely related to the **a** order. Agreeing with Manly and Rickert's classification, Ad3 Bo2 and Ha5 are nearer to Dd than to the manuscripts at the top of the tree (Ds En1 En3 Ad1 Ma El and Cn). The modified tale-order table shows that Ad3 and Bo2 have quite similar orders, even with some elements in different places.

In plate 5, Bo1 and Ph2, the two last manuscripts in Manly and Rickert's **a** group, are the only ones which appear far from the rest of **a** in the tree (they also appear separated from the **a** groups in plates 6, 7 and 8). In fact, according to this tree, they are more closely related to the **c** group and to the **b d** cluster. Both Bo1 and Ph2 have FK followed by NU-L33-CY, and all three **c** manuscripts have FK followed by NU. The number of changes required to move NU-L33-CY to the position they have in Bo1 and Ph2 is very significant for the overall placement of these manuscripts on the tree, but there is yet another element that displaces them from a position near the **a** group. In Bo1 and Ph2, CL is directly followed by FK, instead of the El sequence: CL-L13-L14-L15-ME-L17-SQ-L20-FK. Cp and La only have L13 and ME between CL and FK. Bo1 and Ph2 appear together in the tree, and this is not surprising since these two manuscripts present a very similar order. What seems strange is that Manly and Rickert decided to include these two manuscripts in the **a** group. The main difference in order between Bo1 and Ph2 and manuscripts of the **c** group is that the latter have L8-SQ and L13 ME which

are not present in Bo1 and Ph2. These two manuscripts have also included L34 between CY and PH and L22 between PD and SH (where La has L23).

The **b** and **d** group manuscripts appear mixed at the bottom of tree (plate 7). Dempster's ideas about the possible common origin for the **b** and **d** groups seem to be supported by the fact that the trees generated by *PAUP* represent them in this way. Ne and Cx1 have the same tale order, but it is not clear why Ra2 appears next to these. The tale order in Ra2 is clearly different, since it includes TG, L12, L22 and L34 and lacks L2, L3 and CO.

In plate 7, Ra1, classified as **b**, is the nearest manuscript to Hg. The Hg order is truly anomalous, in the sense that its links have been altered to cover for mistakes made by its scribe. It is likely that the scribe copied the SQ ME and FK tales before he copied the links. The scribe had left the space for the links to add them between the tales, but he had copied the tales in an order that was not supported by the links:

[T]he link between Squire and Merchant copied onto the verso of fol. 137 is copied in the hand of the Hg scribe but in the ink used for the last half of section III, suggesting a late addition. The link used may have been altered to fit this position though it could be an early version. The same applies to the link on the inserted fol. 153 which, along with decorative gaps, also contains the first twelve lines of the Franklin's Tale. It has been argued. . . . that the texts of the Nun's Priest and Manciple and the two linking passages in Section IV were probably the last work of the Hg scribe as he endeavoured to 'complete' the manuscript. It has also been noted that at this final stage the

supervisor in charge of the supply of texts and the ordering and organising of the material took no further part. None of the folios in yellow ink show any features of his work. It would seem then that the Hg scribe knew that he could not adequately link the tales in the order in which they had been copied. Nevertheless he used or adapted available material to make the manuscript appear more complete. If the Hg scribe acted on his own initiative, the supervisor may not have overseen the copying of material in the last half of Section III or advised on the placement of the two linking passages in Section IV (Inks).

In fact, Hg has two lines of SQ in f. 137v; the rest of the page probably was, as Stubbs suggests, written at a later time in a yellowish ink. The text following these two lines in f. 137v is L20, which goes to the end of the page. The textual variants between Hg and El at this point witness the scribe's alteration of the text to fit the order in which he had already copied the tales. The initial rubric in Hg reads: '☛ The prologe / of the Marchaūtes tale ~.' El, on the other hand has: '☛ Heere folwen the wordes of the Frankeleyn to the Squier ~ and the wordes of the hoost7 to the Frankeleyn ~ ~ ~.' Each manuscript follows the pattern established in the rubrics.

Quod the Marchant7 considerynge thy youthe

So feelyngly thow spekest7 sire I allow the (Hg ll. 3-4)

☛ Straw for youre gentillesse / quod oure hoost

What Marchaūt / pardee sire wel thow woost (Hg ll. 23-24)

☞ That knowe I wel sire / quod the Marchant c^oteyn
I prey yow / haueth me nat in desdeyn (Hg ll. 27-28)

Quod the Frankeleyn / considerynge thy yowthe
So feelyngly thou spekest^r sire I allowethe (El ll. 3-4)

☞ Straw for your^e gentillesse / quod oure hoost^r
What Frankeleyn / p^rdee sire wel thou woost^r (El ll. 23-24)

☞ That knowe I wel sire / quod the Frankeleyn
I prey yow / haueth me nat in desdeyn (El ll. 27-28)

In these lines, the changes made in Hg by the scribe are quite evident. Sometimes, they have dramatic results on the meter of the line (l. 27). In L17 we also find that the meter of the line has been altered to such a degree that it would be difficult to accept that it is not the result of scribal intervention:

☞ Sire Frankeleyn / com neer / if it your^e wille be
And sey vs a tale / for certes ye (Hg ll. 23-24)

☞ Squier com neer / if it your^e wille be
And sey somewhat of loue / for c^otes ye ~ (El ll. 23-24)

It seems evident that after copying the tales, the scribe realized that the links suggested a different order; Stubbs' suggestion of the ink-color indicating a later addition seems correct, and the idea that the links arrived at a later date than the tales is acceptable.

Instead of correcting the mistake, the scribe attempted to cover it by altering the names in the links to make them look as if they should be linking the tales in the order in which he had copied them.²¹

The section from SQ to L14 has been completely modified in Hg and it includes NU. The Hg sequence FK NU CL is only found in two other manuscripts Ad3 and Ha5. There are not many points of coincidence between Hg and Ad3, but this link is important because Hg does not have (and never had) L33 and CY, while Ad3 has these two immediately before L37 and PA. There is only one other manuscript with the same sequence, L33-CY L37-PA, Ch. Another remarkable fact about Hg is that an alteration was made to the name of the pilgrim in L37 (the current reading in Hg in the Parson's Prologue is 'manciple'). This might indicate another change in the order in Hg, perhaps due to the intervention of the scribe or maybe due to the fact that part of the text never became available to him to be copied into Hg.

The previous discussion shows that the overall shape of the tree can be explained if we observe the closeness of certain witnesses to one another and compare it to the tale-order table. To do this by hand is a much more complex task, so much so that Manly and Rickert did not see the possible relationship between Bo1 and Ph2 and the c group. In other cases, such as the situation of Ch near To, we discover that the connection implied by the consecutive positions which Manly and Rickert assigned to these manuscripts is confirmed by the results yielded by the phylogenetic software.

It is important to stress that, not only the positions, but also the distances between the different items are informative. For example, in the case of Gg Ps and Hk, it is

possible to see that even if their positions remain the same the length of the branches of the tree indicates that there are not as many similarities between these as the unrooted cladogram seems to suggest. However, even with all the differences between Hk and Gg, *PAUP* still proposes that these have a closer relationship with each other than with any other witness.

2.2.2 Stemmata Based on Data Encoded Using IEBP

After the analysis of the trees built based on Spencer's method was carried out, STEMMA came across new ongoing research at the department of computer sciences at the University of Texas. There, Tandy Warnow and Li-San Wang have been working on new distance coding methods (Wang and Warnow 636-46). The exchange between STEMMA and Warnow and Wang resulted in a new coding of the tale-order data. This new method is the Inverse of Expected BreakPoint distance method (IEBP), which calculates the possible number of movements of each item in a sequence:

IEBP estimates the true evolutionary distance using an approximation to the relationship between evolutionary distance and expected breakpoint distance, under the assumption that all transpositions are equally likely. Simulations show that phylogenetic reconstructions based on IEBP distance are more accurate than those based on breakpoint distance. (Spencer, "Gene Order")

To produce the new coding, the tale-order table was revised and sent to Wang and Warnow. They coded it according to the principles of their method (IEBP). The nexus

file produced at Texas can be seen in the appendix. Although the Wang-Warnow method clearly offered better results than breakpoint distance when used with *SplitsTree*,²² the trees produced with these program were, once more, uninformative (See plates 11 and 12).

The new nexus file was fed to *PAUP* and yielded the results shown in plates 13 and 14. If we look at the unrooted cladogram (plate 13), we immediately discover that the groups appear in different positions on this tree. In order to establish any differences between the IEBP and the breakpoint stemmata, we have to rotate one of the trees until both of them are in the same relative position. By rotating the stemma in plate 13, 90° to the right, we have the **a** group at the top of the tree and **b** and **d** groups at the bottom left of the stemma.²³ If we now compare plates 7 (upper bound unrooted cladogram) and 13 (IEBP unrooted cladogram) we can see that the **a** group remains in the same relative position from the cluster where we find Gg. However, in this group, El has moved and it appears between Gg and the majority of the **a** group, while Ps and Hk pair with other witnesses. The cluster formed by Bo2 Dd Ad3 and Ha5, in plate 13, appears closer to the rest of **a**, instead of being separated by Ps Hk and Gg (as it is in plate 7). In fact, this is partly due to the new placement of Gg near El. When using IEBP encoding, we find that Hk pairs with To and that both of these spring from the same node as the Bo2 Dd Ad3 and Ha5 cluster. Analyzing this, the only difference at this point is that plate 7 shows an extra node between the Ad3 cluster and the Hk one. Ps, which in plate 7 shows in the Gg group, pairs with Se in plate 13. If we keep going down the tree, we find that Ha4 and Wy appear next, but in inverse positions to those of plate 7.

IEBP seems to suggest that Ha4 is closer to the **a** group than Wy, and it also places Cx2 one node removed from the latter and another from Ch. This could be further confirmation of a possible common source for the variants in both of these printed editions and Ha4 and Ch (Bordalejo 359 and ff.). Other similarities between IEBP and the breakpoint trees are the separation of Bo1 and Ph2 from **a** and their close relationship to one another, as well as the lack of a clear distinction between groups **b** and **d**, and the evident separation of **c** as a distinct and individual group.

However, even among so many common features in the stemmata produced using these two methods for the treatment of the data, in the IEBP unrooted cladogram it seems, once more, that Hg could be related to the **c** group. A peculiarity of this tree is that Hg appears closer to Fi and Ii than to any other witness. Once more, this can be explained by the sequence SQ ME FK which is unique to Fi Ii, and the **d** group. However, Fi and Ii, like Hg, separate CL with other items from this sequence.²⁴ Robinson's hypothesis that hand b (the Hg scribe) made changes to the Hg exemplar which were later transmitted to the **b** group appears supported by the association of Fi and Ii and Hg.

There is a striking similarity between the lower bound (plate 8) and the IEBP unrooted cladograms in the cluster which in the latter includes Ld2 S11 Tc2 Ph3 Mm and Pw. This cluster is comparable with the one in which in the lower bound unrooted cladogram includes Ld2 S11 Py Ph3 Mm and Pw. Py was classified as anomalous by Manly and Rickert, but the overall structure of the order in this manuscript could have fitted the **d** group pattern (although Ps does not have TG, a tale characteristically found in

the **d** group. In the unrooted phylogram, Py appears very close to Ra2 and Ln). In fact, it is closer to these than the other witnesses in this branch.

In plate 14, the IEBP unrooted phylogram, we can see the distances between the different witnesses. For example, in the case of Hk and To (two manuscripts that appear very close to the **a** group) the distance shown in plate 14 suggests that, although they are related, they are not as close to each other as some other witnesses. Once more the length of the branch indicates that it is likely that there were many permutations between one order and the other.

The relationship between Hg Fi and Ii appears to become more evident, since these manuscripts seem to derive from a same common origin in the branch. Another remarkable feature of this stemma is the closeness of Wy Ha4 and Cx2. Cx2 only differs from Cx1 in the addition of L31 and the movement of L8-SQ, but just these characteristics are enough to make Cx2 appear next to Wy and Ha4 instead of in the position of Cx1, which shows up together with the **b** and **d** groups. Cx2 is in the same branch with Ch and Ld1 (both of which were classified as anomalous by Manly and Rickert). The closeness of these five witnesses, all of them anomalous, suggests a genetic relationship between them. Not only do these share the sequence L15 ME...SQ L20 FK, but Cx2 and Wy also have L8 between ME and SQ (Ha4 has L8 after ML).

The changing position of Hg and Ch, for example, could suggest that the order in these manuscripts is not closely related to any other order or that the changes on them, rather than being genetic, are the result of scribal intervention. If a scribe or his supervisor, actively and radically changed the order of an exemplar, it becomes difficult

to sustain the notion that there should be a clear genetic relationship between the resulting manuscript and the one used as its exemplar (of course, this might depend on the number of introduced changes and their originality). However, even when changes have been purposely introduced (a fact which might appear as the unstable position of a manuscript on the trees) the consistency of certain groups in the stemmata points towards a genetic origin and subsequent transmission of many of the tale-orders found in manuscripts of the *Canterbury Tales*.

¹ This is by no means an attempt to survey all the different methods currently available. Instead, it is an attempt to clarify some differences and explain why some methods work or not.

² For further information about this method see Wang and Warnow (637-646).

³ STEMMA is currently exploring the applicability of networking programs to manuscript traditions. However, this research is just starting and it might be some time before it has yielded practical results. Networking might be useful to display the relationships of highly contaminated textual traditions, but this remains to be proven.

⁴ See also Griffith (101-38); Cameron (227-42); Platnick and Cameron (380-5); Robinson and O'Hara, ("Report" 331-37; "Computer-Assisted" 53-74; "Cladistic" 115-137); Robinson ("Best-Text" 71-103; "Stemmatic" 69-132); Robinson et al. ("Phylogeny" 839); Howe et al. ("Manuscript" 147-52).

⁵ See also Howe et al., ("Parallels").

⁶ In ribonucleic acid (RNA) thymine is replaced by another base called uracil (represented as U).

⁷ A good example of word variation in which the final state implies at least one previous state is SQ 491, where the Hg line reads Hg "ƒ Ther I was bred / allas that ilke day" while El reads "ƒ That I was bred / allas that harde day." Manly and Rickert's explanation states that there must have been an intermediate state in which a scribe miscopied 'ilke' as 'ille' and that this reading, in turn, was converted into 'harde.'(4:482-3)

⁸ See Page and Holmes; Spencer and Howe 467-84.

⁹ The key-words here are 'extant data,' while other methods might try the least amount of changes (minimum evolution) to build their trees, maximum likelihood attempts to explain the data that has been provided. In textual criticism terms, this philosophy is very close to that of the New Stemmatics, which presupposes that because the extant manuscripts are not the totality of the textual tradition a stemma built based on them would only map the relationships of these and should not attempt to express witnesses we do not have. If other witnesses were to be discovered in the future, these could be added to the analysis and the new results would be another step in the study rather than a unique solution to the problem.

¹⁰ The version of PAUP used for this research allows us to produce distance trees using options such as "Constrain branch lengths to be nonnegative" and "Collapse branches of effectively zero length when searching." The use of these options should help with the accuracy problems foreseen by Nei and Kumar, and greatly reduces the number of resulting trees, by excluding some branch re-combinations.

¹¹ Spencer and Davidson also present an interesting list of divergences between the aims of evolutionary biology and textual studies. Among these differences, the most interesting one is about the aims of both disciplines: "Another important difference is that biological evolution is continuous. After an evolutionary divergence, both species continue to change, so neither directly represents the ancestor. Phylogeneticists

therefore assume that contemporary species should always appear on the tips (terminal nodes) of the tree. In contrast, once a manuscript is produced, the text it contains does not change (except through corrections and damage). Some extant manuscripts may be the ancestors of others, and should therefore be represented by internal nodes of the stemma” (Spencer et al. “Artificial”). This observation also applies to the tale-order stemmata (since it is likely that some of the extant orders originated the other ones), but as this work does not presuppose a direct correlation between stemma and reality, the conceptual difference of the possibility of allowing internal nodes or presenting each witness as a terminal node should not present an insurmountable problem.

¹² In fact there are other methods (and even more programs which follow them) such as least squares and spectral analysis, which are not discussed here. A discussion of least squares can be found in New and Kumar, and spectral analysis in Page.

¹³ IEBP was devised by Wang and Warnow. For more details refer to their article “Estimating true evolutionary distances between genomes.”

¹⁴ See the nexus files for the distance matrix of the tale-order data in the appendix.

¹⁵ There are two sets of trees here the upper and lower bound because of the way in which the data was processed. BreakPoint distance is defined as “the number of items common to both manuscripts and having different right-hand neighbours.” The difference between upper and lower bound is that: “The lower limit occurs when no common items were lost and the upper limit is approached if there are many lost common items” (Spencer, “‘Gene Order’ Analysis”). When there is missing data, it is not possible to know the number of items which are missing, if the presupposition is that two witnesses have not lost corresponding parts of the text, this is best represented by the lower bound. When witnesses are so damaged that one can assume that many corresponding items are lost then the data is best expressed by the upper bound. Because it is impossible to decide *a priori* which one of these might offer better results, I have analyzed both the upper and the lower bounds.

¹⁶ The word-variant data is generated automatically using *Collate* while the tale-order data has been coded using different methods (Spencer et al., “Analyzing;” Wang and Warnow 637-46). It is possible that this difference might be the result of the difference in coding generated by the use of distance methods, while *Collate* offers a format which might be better for *SplitsTree*.

¹⁷ The nexus files contain both the data and a protocol for the software. These can be seen in the appendix.

¹⁸ See Feinstein, 45-60 and Bordalejo, “The Manuscript Source of Caxton's Second Edition of the *Canterbury Tales* and its Place in the Textual Tradition of the *Tales*.”

¹⁹ At this point of the research, I want to present the data as it is in the manuscripts. After my codicological analysis of the chosen witnesses, the data will be revised to conform with the making of the data provided by a more detailed analysis of the manuscripts.

²⁰ But see plate 5 for an idea of the distance between the Gg and Hk orders, for example.

²¹ This and other aspects of the codicological analysis of the manuscripts are discussed in Chapter 6.

²² The fit of the tree is tree times as high when the encoding method used is IEBP than the one obtained with breakpoint distance. The fit of the tree is an indicator of how adequately *SplitsTree* has been at handling the data.

²³ The rotation does not affect the relationships between the branches of the trees. Because the tree is unrooted what really matters are the relative positions of the items in the tree (Robinson, “Analysis”).

²⁴ Hg has NU between SQ L20 ME L17 FK and CL, while Fi and Ii have WB L10 FR L11 SU in that position. Ht has WB L10 FR L11 SU NU L33 CY L34 PH L21 and PD between SQ L20 ME L17 FK and CL.