Untangling our past: Languages, Trees, Splits and Networks

David Bryant[1], Flavia Filimon[2] & Russell D. Gray[2]

[1]McGill Centre for Bioinformatics

Montreal, Quebec, H3A 2B4

Canada.

[2]Department of Psychology

University of Auckland

Private Bag

Auckland 92019

New Zealand.

In Act 1 of *The Importance of Being Earnest* Algernon quips to Jack that, "The truth is rarely pure and never simple". The idea that much of recent human history might reflect pure trees of phylogenetic descent is appealing simple. It has stimulated numerous researchers to investigate the extent to which genes, languages and cultures are bound together in codiverging trees of evolutionary history (e.g. Cavalli Sforza et al, 1994, see Cavalli Sforza & Feldman (2003) for a recent review). Increasingly studies have used explicit phylogenetic methods to make inferences about linguistic history and the evolution of cultural traits (Warnow & Ringe, 1997; Gray & Jordan, 2000; Pagel, 2000; Holden, 2002; O'Brien et al, 2002; Jordan & Shennan, 2003; Holden & Mace, 2003; Rexova et al. 2003). However, a persistent criticism of this approach is that human population history is far from tree-like (Moore 1994, Terrell 1988, Terrell et al 2001). Not only might patterns of genetic, linguistic, and cultural diversity reflect different histories (Bateman et al, 1990), each of these histories might be strikingly reticulate. As one participant at a recent symposium on phylogenetic methods in archaeology growled, "This is not history. This is history put in nested boxes!"

This debate about the shape of recent human population history has unfortunately become rather polarised between defenders of a phylogenetic approach and their rhizomatic critics. Each group cites examples that appear to support their view. Each group pokes holes in the logic and practice of the other. Rhizomophiles correctly point out that standard phylogeny programmes always produce a tree, even when a tree model is not appropriate. Phylophiles rightly argue that they are only using a tree as an initial starting point in their investigations of the complexities of human history (Jordan & Gray, 2001). While phylogenetic programmes do indeed always output a tree (or trees), it is possible to evaluate how well the data fits a tree model using these programmes. For example, consistency and retention indices can be used to measure the fit of a data set on a tree. However, what is not possible using these programmes is to evaluate explicit alternatives to a pure tree model. What is

needed to get beyond the impasse of these polarised *a priori* views is an analytic approach that enables us to assess where on the continuum between a pure tree and a totally tangled network any particular case may lie. More specifically, this approach should be able both to identify the particular populations where admixture has occurred and detail the exact characters that were borrowed. In this chapter we will outline two such methods - Split Decomposition and NeighbourNet - using Indo-European lexical data to demonstrate the potential and possible pitfalls of these approaches.

Trees and splits

What information is contained in an evolutionary tree? Ask a mathematician and you will probably get details on which node is connected to which edge. Ask a linguist or biologist and you are much more likely to get details on the *groupings* in the data. When the tree is rooted, each node in the tree gives a different group: the cluster of taxa that are descendants of that node. (The *taxa* are the objects under study, singular *taxon*). If the tree does not have a root (that is, it is *unrooted*) then the groupings are really *splits* or *bipartitions* made up of a group and its complement. A split is nothing more than a division of the set of taxa into two non-empty parts. We use splits for unrooted trees since we don't know which node is a descendent of which, so can't tell which side of the split corresponds to a cluster.

Each branch (edge) in an unrooted tree corresponds to a different split of the taxa that label the leaves of this tree, since removing that edge divides the taxa into two non-empty parts. The collection of all splits corresponding to edges in the tree (one for each branch) contains all the information we need to reconstruct the tree. The collection of splits for a tree *T* is denoted *splits*(*T*). We say that a given collection of splits is *compatible* if it is contained within *splits(T)* for some tree *T*.

Most collections of splits are not compatible, but it is easy to characterise those

which are (Buneman 1971; Semple and Steel 2003). Our main interest in compatible splits is that in order to generalise trees, we construct collections of splits that are *not* compatible. Indeed, thinking of trees as collections of splits allows us to go beyond trees in a natural manner, allowing us to represent groupings that cannot be represented with a single tree.

Our first step towards constructing phylogenetic networks is to generate an appropriate collection of splits that may not necessarily be compatible. But this is just the first step. After this, we also assign weights (the equivalent of branch lengths) to these splits and then represent the weighted splits graphically. The representation is a snap-shot of the data, but it is a snap-shot with a particular lens. Often the network is easy to interpret, as is the case when the network closely resembles a tree with no, or a small number, of admixture events. In other cases, the complexity of the signal (or presence of noise and error) leads to complicated networks where it is sometimes difficult to separate noise and the underlying signal.

We will focus on just two methods for building phylogenetic networks: split decomposition (Bandelt and Dress 1992; Huson 1998) and Neighbor-Net (Bryant and Moulton 2002; Bryant and Moulton 2003). Both follow the same general scheme: (a) we determine a distance matrix from the data (although we note that discrete data can also be used for split decomposition); (b) use the distance matrix to generate a collection of splits; (c) compute weights (branch lengths) for these splits; and (d) represent the weighted splits using a special kind of network called a *splits graph*. We discuss each of these steps in turn.

(a) Constructing distance matrices from lexical data

Languages diverge with time. One obvious way of measuring this divergence is by calculating the extent to which their vocabularies have diverged. Since at least the 1950s historical linguistics have used divergence in lists of basic vocabulary to

estimate historical relationships in a quantitative manner (e.g. Swadesh 1952, see Embleton 2000 for a recent review). This approach, termed “lexicostatistics”, subgrouped related languages on the basis of the percentage of cognates they shared from a list of basic vocabulary. An extension of this method, known as glottochronology, attempted to estimate divergence dates between languages using the formula,

$$t = \log c / 2\log r$$

where *t* is time, *c* is the percentage of shared cognates, and *r* is the retention rate per thousand years. The retention rate was assumed to be roughly constant at around 81% per thousand years for the Swadesh 200 word list.

[insert table 1 about here]

However, both glottochronology and lexicostatistics have been shown to give inaccurate results in cases of languages evolving at markedly different rates (Trask, 1996, Blust 2000). Both approaches are also likely to give inaccurate answers in cases where there has been substantial borrowing. Today glottochronology is rejected by most historical linguists and lexicostatistics is generally regarded as only useful as a crude initial approximation of possible subgroups (Trask, 1996, Campbell 1988). However, the idea that there is useful historical information in cognate distances need not be abandoned provided the quantitative methods allow languages to evolve at different rates and do not assume that this evolution is always strictly tree-like. For the sake of simplicity in the examples that follow we will calculate distances between Indo-European languages based on the mean percent difference in cognacy. We note, however, that just as biologists correct estimates of genetic divergence using models of sequence change (e.g. Nei, 1991), we could correct these lexical distances to allow for factors like a distribution of word rate changes (Sankoff, 1970, 1973).

[Insert Table 2 about here].

(b) Generating splits

Split decomposition and Neighbor-Net differ considerably in how they generate splits from the distance matrix. Split decomposition starts with an explicit formula for scoring splits, and returns exactly those splits with positive score. Neighbor-Net, on the other hand, more closely resembles agglomerative clustering algorithms like the single and average linkage methods. It constructs splits by progressively combining clusters in a way that allows overlap.

Split decomposition

Consider the network in figure 1. We can define distances in the network just as for any graph (or tree): the distance between two nodes equals the length of the shortest path between them. For example, the distance between taxa *a* and taxa *c* is

$$D[a,c] = 0.1 + 0.3+ 0.5 + 0.2 = 1.1$$

In this case, we can go in two directions. Given the edge lengths we can compute taxon to taxon distances. However if we are given the taxon to taxon distances we can, in fact, recover all of the edge lengths.

For any four taxa *a,b,c,d* define the value *w(ab|cd)* by

$$w(ab|cd) = 1/2 \left(\max\{D[a,c]+D[b,d],D[a,d]+D[b,c]\} - D[a,b] - D[c,d] \right)$$

Looking at the network, you can quickly convince yourself that w(ab|cd) equals the length of the internal (horizontal) edges. In this case, we have

$$w(ab|cd) = 1/2 (\max\{1.1 + 1.0, 0.7 + 0.8\} - 0.5 - 0.6) = 0.5$$

The length of the vertical internal length equals w(ac|bd). In fact, if we can stretch the formula a little, we see that the edge adjacent to *a* is equal to w(aa|bc).

The function w only gives a score for (at most) four taxa at a time. However we can easily extend the formula to a scoring function on entire splits. For any split A|B (that is, the split dividing the taxa into two sets A and B), define the weight of A|B by

$$w(A|B) = \min\{w(aa'|bb') : a,a' \text{ in } A,\ b,b' \text{ in } B\}$$

That is, w(A|B) is the *minimum* score w(aa'|bb') over all choices of taxa a,a' from one side and taxa b,b' from the other side of the split. The score w(A|B) indicates how separated the taxa in A are from the taxa in B. The larger the score, the more support for this split.

We have now a way of scoring splits given a distance matrix. The idea behind split decomposition is to construct the collection *all* splits A|B with a positive score w(A|B). There are a lot of ways of splitting a set of *N* taxa ($2^{N-1} - 1$, to be exact), but it turns out that there is always only a small number (at most N(N-1)/2) of splits A|B with a positive score w(A|B). What's more, there exist fast algorithms for generating this collection of splits (Bandelt and Dress 1992, Berry and Bryant 1999).

As an example, consider the distance matrix computed for the 9 Germanic languages in Table 2. The weight of the quartet separating English and Sranan from German and Swedish equals 0.1495. In fact, all of the quartets separating English and Sranan from any two other languages have positive weight. Hence the split separating English and Sranan from the other taxa appears in the set of splits generated by split decomposition.

While Sranan and English are separated from the other taxa, there are other data that group Sranan with other languages apart from English. The quartet separating Sranan and German from English and Swedish (the same four languages as before) also has a positive weight (0.0525), though this weight is smaller than for the first quartet. In fact if we examine all sets of two taxa from English, Swedish, Riksmal, Faroese and two taxa from Sranan, German, P. Dutch, Dutch and Flemish we see that all of these quartets have positive weight. Therefore this split is also generated by split decomposition.

There are a total of 16 splits generated by split decomposition. The splits separating more than one taxon from the remainder are:

[Insert TABLE 3 about here]

In the table, we should consider the 1's and 0's interchangeable: we use 1 and 0 only to indicate the separation into two groups.

One attractive feature of split decomposition is that after computing the weights w(ab|cd) for each set of four taxa, we no longer make use of the distance matrix D. In fact, we can go straight to the weights for the quartets and bypass the need to construct a distance matrix at all. The only constraint is that, for every four taxa, at most two out of the three weights w(ab|cd), w(ac|bd), w(ad|bc) are positive.

An example of this approach is "parsimony splits". For each set of four taxa a,b,c,d, we count up how many characters separate a,b from c,d; how many separate a,c, from b,d and how many separate a,d from b,c. We assign weight one to the two highest supported groupings and zero to the remaining grouping (e.g. w(ab|cd) = w(ac|bd) = 1, w(ad|bc) = 0). The algorithm then constructs a set of splits representing these groupings. One note of caution, however. Just like parsimony, this approach

can be shown to be inconsistent when the distances between the taxa become too large.

Split decomposition is effective when the number of taxa is small, or when the signal is not too complicated. The criterion w(A|B)>0 that a split appear in the split decomposition is very strict: it is the minimum of *w(ab|cd)* over *all* a,b on one side and *c,d* on the other. When there is variance in the distances, this value will be bias negative. In practice, as the number of taxa increases the values w(A|B) become smaller and smaller, and, subsequently, the number of splits in the split decomposition decreases. Thus extremely `un-treelike' data can appear as completely `tree-like', simply because all of the w(A|B) values have become small or negative.

Neighbor-Net

Many clustering and tree construction algorithms are variants of the same general method. We start with each taxon in associated with a separate node. At each iteration we use some optimality criterion to select two nodes. These are fused together to make a composite node, and the process repeats until only two or three nodes remain. At this point, the history of fusions defines a tree or collection of clusters. Neighbor-Net follows the same general scheme, with one important difference. When we select a pair of nodes we do not combine and replace them immediately. Instead we wait until a node has been paired up a second time. We then replace the three linked nodes with two linked nodes and reduce the distance matrix. If there is still a node linked to two others we perform a second agglomeration and reduction. The process continues until only three nodes remain. Because each node is combined with two others, the resulting collection of clusters is overlapping: so the splits we generate are not compatible. The selection criteria, and formulae for computing distances between composite nodes, are directly analogous to those used for Neighbor-Joining.

To illustrate, consider the distance matrix for the nine Germanic taxa. The first two languages to be paired up are Faroese and Riksmal. However we don't agglomerate the two nodes immediately. The next two languages to be paired up are Faroese and Swedish. Now Faroese has been paired up twice: we might envisage this incomplete network as in figure 2(i). The three nodes are agglomerated to make two `composite' nodes (figure 2(ii)). The method then pairs up Sranan and English, then Dutch and Flemish. Once again, we don't immediately combine these nodes (figure 2(iii)). The next step pairs English with Swedish/Faroese. Since both of these are already paired up, we get two agglomerations immediately after each other (figure 2(iv)).

[Insert figure 2 about here]

When all the agglomerations have taken place we end up with three nodes and a list of which ordered pairs of nodes replaced which ordered triples of nodes. To construct the set of splits we first use this list of agglomerations to order the taxa around a circle. We start with the three remaining nodes in any order. We work backward through the list. Each agglomeration will replace two adjacent nodes with three adjacent nodes in a given order. When we have worked backwards all the way up to the front of the list. The ordering produced from the 9 Germanic languages is

Flemish - Dutch - German - Penn. Dutch - Sranan - English - Swedish - Faroese - Riksmal - Flemish

The splits generated by Neighbor-Net are then the splits formed from consecutive countries. In practice, however, when we estimate split lengths a lot of these splits are assigned weight zero and removed from the network.

Neighbor-Net can be seen as a heuristic method for obtaining splits in the data, just like Neighbor-joining (Saitou and Nei, 1987) is a heuristic method for constructing trees. There is no explicit definition for splits in the Neighbor-Net like there is for splits generated by split decomposition. The upside is that Neighbor-Net will generate many splits, and informative networks, even for large collections of taxa. This advantage, combined with the speed of the algorithm, have enabled network analysis of data sets with hundreds of taxa, data for which split decomposition generally returned no informative splits (Bryant and Moulton, in press).

(c) Computing split weights

When constructing trees from distances we distinguish between two different distance matrices: the observed or corrected distances (i.e. the data) and the tree-like or *phenetic* distances (i.e. the hypothesis). The phenetic distance between two taxa in a tree equals the sum of the branch lengths along the unique path that connects them. The general aim is to find a tree with branch lengths so that the corresponding phenetic distances best fit (or best explain) the observed distances.

The phenetic distances can be also re-expressed in terms of splits. The branches along the path between two taxa correspond exactly to the splits in the tree that separate those two taxa (that is, the taxa are on different sides of the split). Assigning branch lengths to a tree is the same as weighting the splits of that tree. The phenetic distance can then be defined only with reference to the weighted splits, without using to involve the actual tree. Of course, this formulation generalises directly to the case of incompatible splits.

Suppose we have some collection S of splits, generated by split decomposition or Neighbor-Net or whatever method we have decided on. If we also have weights (or lengths) for the splits, we can define a phenetic distance. The phenetic distance between two taxa is defined to be the sum of the weights of the splits separating

them. We can now work backwards. Given observed distances, and a collection of splits, we can choose the split weights so that the corresponding *phenetic* distance best fits the *observed* distances. With Neighbor-Net and split decomposition we use a least squares measure of fit: if **D** represents the observed distances, and **p** represents the phenetic distances, we want to minimise

$$\sum_{ij} w_{ij} (D_{ij} - p_{ij})^2$$

The weights w_{ij} equal one for ordinary least squares and the reciprocals of the variances for weighted least squares. In practice we optimise the fit with a non-negativity constraint on the weights.

With split decomposition we can alternatively weight the splits using the scores $w(A|B)$ that we computed. Since they are defined using minima, these weights underestimate the distances between taxa. Even with as few as nine taxa, the differences can be significant (cf TABLE 3).

(d) The splits graph

At this point, we have inferred two types of information: the splits, which represent the groupings in the data; and the branch lengths, which indicate the degree of separation for each split. When the splits are compatible, we can represent both kinds of information using a tree. The branches represent the splits and the branch lengths indicate the split weights. When the splits are incompatible, we use a *splits graph*. A splits graph is a graphical representation of a collection of weighted splits.

In a tree, each split corresponds to a single edge. Removing that edge partitions the taxa set into the two parts making up the split. In a splits graph, each split corresponds to a *collection* of parallel edges, all with length equal to the weight of

the split. Removing those edges partitions the graph, and therefore taxa set, into the two parts making up the split.

The simplest splits graph that is not a tree is depicted in Figure 1. The graph represents six splits: the four splits separating one taxa from the rest, one split separating a,b from c,d, and another separating a,c from b,d. Split weights are marked on the figure. The two darker internal edges correspond to the split {a,b}|{c,d}, while the dotted edges correspond to {a,c}|{b,d}.

[Insert figure three about here]

In Figure 3 we have splits graphs for the splits generated by Split decomposition and Neighbor-Net from the example distance matrix. In both cases we use ordinary least squares to determine branch lengths. The conflicting splits have generated boxes. The two splits graphs are very similar, as tends to be the case when the number of taxa is small. Both show strong conflicting signal in the placement of Sranan. This is exactly what the methods should do given that Sranan is a creole (developed by African slaves in Surinam on the northern coast of South America) with words derived from both English and Dutch. The English established Surinam in 1651 as a slave colony but Dutch has been the official language since 1667 (McWhorter, 2001). Table 4 shows some cognate sets that reveal evidence of this dual influence. Note, however, that Neighbor-Net picks up a significant split separating German, Dutch and Flemish from the other languages. This split is missed by split decomposition, but supported by several cognates. For example, one of the two Dutch words meaning “to blow” (“waaien”) is cognate with the German “wehen’ and the Flemish “waeijen” but not with the words for “to blow” found in the other Germanic languages (see Table 5). The conflicting signal probably reflects a combination of German borrowings into Dutch and Flemish, and English borrowings into Pennsylvania Dutch. "Pennsylvania Dutch" is a dialect spoken in Pennsylvania,

USA, and is actually a variant of German, not of Dutch. Its name arose from the fact that German settlers in Pennsylvania referred to their language as "Deutsch". The influence of the English-speaking environment on Pennsylvania Dutch can be seen in the fact that it shares a cognate for “smoke” (“schmoeck”) with English and Sranan (“smoko”) that is not found in German or Dutch (where the terms are “rauch” and “rook” respectively).

[Table 5 about here]

Figure 3 provides only one split graph representation of these data. We can rearrange the graph by rotating edges and moving nodes. This kind of modification is best illustrated by downloading SplitsTree from http://www-ab.informatik.uni-tuebingen.de/software/jsplits/welcome_en.html, and playing with the display options. When the graphs become more complicated it is possible that the same splits might have a representation using a completely different splits graph. This ambiguity necessitates care when interpreting splits graphs. The nodes in a splits graph do not necessarily correspond to the ancestor of recombinant species (Strimmer and Moulton, 2001). The graphs represent conflicting signal rather than a reconstruction of evolutionary history. That said they provide a valuable exploratory data analysis tool in the reconstruction of evolutionary history and representation of data.

Split support

Splits graphs not only represent different groupings, they can indicate the degree of support for that grouping. The significance of the weights (or lengths) assigned to different splits depends, of course, on how the distance matrix was constructed and which method was used to estimate branch lengths. Mathematically, the weights are chosen so that the distances between taxa in the network best fits the distances in the input matrix. When it comes to interpretation, we can generally interpret the length of a split to indicate the degree to which the two sides of the split

are separated.

To illustrate, consider the simple case when there is absolutely no conflict in the data: the characters correspond exactly to the splits of a tree. The weights of the splits will then correspond exactly to the lengths of the branches in the tree. If we started with discrete binary character data, the length of the edge would (generally) correspond to the number of characters that differed along that edge.

The situation becomes more complicated if the data is actually a mixture between different evolutionary histories. This scenario can arise when there is borrowing or hybrids, although conflicting signals can also arise because of errors in the data. In this case, the weights will reflect a combination of branch lengths of the splits in different trees. Suppose, for example, that 1/3 of the characters evolved along one tree and 2/3 on another. If a branch had length 8 in the first tree and the corresponding branch had length 10 in the second tree then we would expect the split to appear with weight 8 * 1/3 + 10 *2/3 in the splits graph. On the other hand, if a split only appears in the first tree, and there it corresponding to a branch of length 5, then we would expect the split to have weight 5*1/3 + 0*2/3 in the final splits graph. In complicated situations, with several trees, it may not be possible to represent all of the splits for all of the trees for all the different implicit evolutionary histories. When this happens, both split decomposition and Neighbor-Net can be seen as ‘approximations' of the splits.

In any case, the strength of both methods lies in their ability to represent data without assuming the data is tree-like. Thus any conclusions drawn from the networks can generally be validated easily by returning to the data and identifying the exact characters giving rise to a particular conflict. *We strongly encourage this.*

Case study: How tree-like are Indo-European languages?

Casual consideration of the history of the English language would lead one to believe that language evolution is anything but tree-like. English is a veritable fruit salad of a language with chunks of vocabulary from the Celts, Romans, Angles, Saxons, Jukes, Vikings, Normans, and slices of Latin, French, Greek, and Italian tossed with some more recent garnishes from Arabic, Persian, Turkish and Hindi. There is even the odd Polynesian borrowing like “tattoo”. Ninety-nine percent of words in the Oxford English Dictionary are in fact borrowings from other languages (McWhorter, 2001). So how can linguists routinely construct language family trees and happily classify English as a Germanic language? The answer to this question is first that the incredibly hybrid history of the English language is not typical of most languages. Second, even for English its Germanic roots are plainly evident in its grammar, morphology and basic vocabulary. Although over fifty percent of the total English lexicon comes from Romance languages post the Norman conquest, this figure falls to around 6% for basic vocabulary such as the Swadesh 200 word list (Embleton, 1986). This explains why recent phylogenetic analyses of Indo-European linguistic data have found that the language relationships are strikingly tree-like (Warnow and Ringe, 1997, Rexova et al. 2000, Ringe et al, 2003).

However, one problem with these analyses was that while these methods can determine whether or not the data are tree-like, they lack any systematic way of investigating any non-tree-like signals in the data. This is particularly true if there is a lot of admixture present. NeighbourNet enables us to perform such an investigation. The Dyen et al (1992, 1997) database of cognates sets for the Swadesh 200 wordlist provides an appropriate Indo-European test data set. As is standard in lexicostatistical studies, Dyen et al excluded obvious borrowings from their cognate sets (they were placed in the 001 set of non-cognates). This means that the cognate sets in Dyen et al do not enable us to obtain an unbiased assessment tree-likeness in Indo-European basic vocabulary. To remedy this problem Filimon (2000) returned

all the borrowings found in the 001 set to the appropriate cognate sets.

The result of the NeighbourNet analysis of this data is shown in figure 4. Not only does this analysis confirm that Indo-European language relationships are strikingly tree-like (for basic vocabulary), it also uncovers some interesting cases of conflicting signal. As might be expected given English's history of French borrowings post the Norman Conquest, and Latin borrowings with the Renaissance, it shows a small conflicting signal grouping it with the Italic family of languages (split 1 in Figure 4). More significantly, Vlach (also called Macedo Romanian and Aromanian) and Romanian show strong conflicting signals in their relationships (splits 2 and 3 in figure 4). While Vlach is most closely related to Romanian, they separated at least a thousand years ago (Grimes, 2000). Since this time the mobile nature of the Vlach-speaking people, and the distribution of Vlach dialects in Greece, Albania, Bosnia-Herzegovinia, Bulgaria, Macedonia, and Yugoslavia, has resulted in considerable borrowing from the non-Romance languages spoken in these areas. These borrowings "pull" Vlach away from Romanian and its Italic relatives towards the base of the splits graph, producing the box-like sections of the graph.

The small number of box-like sections of the splits graph supports the claim of Warnow and Ringe (1997) and Rexova et al. (2000) that Indo-European linguistic relationships are strikingly tree-like (for basic vocabulary at least). It is reassuring that Neighbor-Net produces a tree-like network when the data are tree-like. It means that we could use Neighbor-Net for the first assessment of the data, and then turn to more specialised tree-based methods if the data turned out to be tree-like. (For an example of this approach see Gray & Atkinson's (2003) application of Bayesian phylogenetic methods to the estimation of the age of proto-Indo-European).

Conclusion

Recovering the truth about population history might not be simple, but it is starting to look tractable. There are a growing number of tools for reconstructing evolutionary history. Tree based methods are a powerful tool for recovering language relationships, provided that these relationships are close to tree-like. When this is not the case (as it may often be) we require tools that make fewer assumptions about the data and have the capability to recover conflicting signals. Neighbor-Net and split decomposition are two such methods, and there will no doubt be many more in the near future. We saw that Neighbor-Net and split decomposition could detect known reticulations in the data examined. On the other hand, both methods return tree-like networks when the data is close to tree-like.

No method is perfect, and these network approaches have several shortcomings. Most importantly, it is not readily apparent how we might validate the inferences made from the networks. Are the splits we see actually significant? Or are they the outcome of a method that is too eager to detect conflict, even if when it is not present? We would also like to construct networks directly from the data, instead of boiling everything down to a distance matrix (or, in the case of parsimony splits, quartet scores).

Both of these shortcomings demand a greater incorporation of the underlying evolutionary model into network construction. The development, and improvement, of these models is a pre-requisite for significant advances in analysis techniques. And of course, improved analysis techniques are exactly what we need to refine the evolutionary models. Recovering the truth about population history might be tractable, but not every tractable problem is simple.

References

Barbrook, A. C., Howe, C. J., Blake, N. & Robinson, P. (1998). The phylogeny of *The Canterbury Tales*. *Nature,* **394**, 839.

Bandelt, H.-J. and A. W. M. Dress (1992). A canonical decomposition theory for metrics on a finite set. *Advances in Mathematics* **92**, 47-105.

Bateman, R., Goddard, I., O'Grady, R., Funk, V. A., Mooi, R., Kress, W. J. & Cannell, P. (1990). Speaking of forked tongues. The feasibility of reconciling human phylogeny and the history of language. *Current Anthropology, **31***, 1-13.

Bergsland, K. & Vogt, H. (1962). On the validity of glottochronology. *Current Anthropology,* **3**, 115-153.

Berry, V. and Bryant, D. (1999). Faster reliable phylogenetic analysis. *Proc. 3rd international conference on Computational Molecular Biology (RECOMB),* **3**, 59-68.

Blust, R. (2000). Why lexicostatistics doesn't work: the 'universal constant' hypothesis and the Austronesian languages, In: *Time Depth in Historical Linguistics* (Renfrew, C., McMahon, A. & Trask, L. eds). The McDonald Institute for Archaeological Research, Cambridge, pp. 311-332

Bryant, D. and Moulton, V. (2002). NeighborNet: An agglomerative algorithm for the construction of planar phylogenetic networks. In: *Workshop in Algorithms for Bioinformatics (WABI).* (Guigo, R. & Gusfield, D. eds.). pp 375-391.

Bryant, D. and. Moulton, V. (in press). NeighbourNet: An agglomerative method for the construction of phylogenetic networks. *Molecular Biology and Evolution*.

Buneman, P. (1971). The recovery of trees from measures of dissimilarity. In: *Mathematics in the Archaeological and Historical Sciences*, (Hodson, F.R.,

Kendall, D.G. and Tautu, P. eds.). Edinburgh University Press, Edinburgh. pp. 387-395.

Campbell, L. (1998). *Historical Linguistics: An Introduction*. Edinburgh University Press, Edinburgh

Cavalli-Sforza, L.L., Piazza, A., Menozzi, P., and Mountain, J. (1988). Reconstruction of human evolution: bringing together genetic, archaeological, and linguistic data. *Proceedings of the National Academy of Sciences,* **85**, 6002-6006.

Cavalli-Sforza, L. L., Minch, E. and Mountain, J. L. (1992). Coevolution of genes and languages revisited. *Proceedings of the National Academy of Sciences,* **89**, 5620-5624.

Cavalli-Sforza, L. L., & Feldman, M. W. (2003). The application of molecular genetic approaches to the study of human evolution. *Nature Genetics Supplement*, **33**, 266-275

Cavalli-Sforza, L. L., Menozzi, P. & Piazza, A. (1994). *The History and Geography of Human Genes.* Princeton University Press, Princeton.

Dyen, I., Kruskal, J. B. & Black, P. FILE IE-DATA1. Available online: http://www.ntu.edu.au/education/langs/ielex/IE-DATA1 [retrieved from the world wide web 2000, 15 March].

Embleton, S. (1986). *Statistics in Historical Linguistics*. Brockmeyer, Bochum.

Embleton, S. (2000). Lexicostatistics/glottochronology: from Swadesh to Sankoff to Starostin to future horizons. In: *Time Depth in Historical Linguistics* (Renfrew, C., McMahon, A. & Trask, L. eds.). The McDonald Institute for Archaeological Research, Cambridge, pp. 143-165

Filimon, F. (2000). *Linguistic and Biological Evolution: Trees versus Networks. An analysis of the Indo-European language family using phylogenetic methods*. Honours dissertation, Department of Psychology, University Auckland, Auckland.

Gamkrelidze, T. V. & Ivanov, V. V. (1995). *Indo-European and the Indo-Europeans. Trends in Linguistics 80*. Mouton de Gruyter, Berlin.

Gray, R.D., & Atkinson, Q.D. (2003). Language-tree divergence times support the Anatolian theory of Indo-European origins. *Nature*. (in press).

Gray, R.D., & Jordan, F.M. (2000). Language trees support the express-train sequence of Austronesian expansion. *Nature* **405**, 1052-1055

Grimes, B. F., ed. (2000). *Ethnologue*, 14th ed. edn, Summer Institute of Linguistics. [Online]. Available at: http://www.ethnologue.com/

Holden, C.J. (2002). Bantu language trees reflect the spread of farming across sub-Saharan Africa: a maximum-parsimony analysis. *Proc. Roy. Soc. London (B),* **269**, 793-799.

Holden, C.J. & Mace, R. (2003). Spread of cattle led to the loss of matrilineal descent in Africa: a coevolutionary analysis. *Proc. Roy. Soc. London (B): Biological Sciences*. DOI: 10.1098/rspb.2003.2535

Huson, D. (1998). SplitsTree - a program for analyzing and visualizing evolutionary data. Bioinformatics **14**(1), 68-73.

Jordan, F.M. & R.D. Gray. (2001). Comment on “Foregone conclusions? In search of ‘Papuans’ and ‘Austronesians’” by J.E.Terrell, K.M. Kelly, and P. Rainbird. *Current Anthropology,* **42**, 114-115.

Jordan, P. & Shennan S.J. (2003). Cultural transmission, language and basketry traditions amongst the California Indians. *J. Anth. Archaeology,* **22**: 42-74.

McMahon, A. & McMahon, R. (2003). Finding families: Quantitative methods in language classification. *Transactions of the Philological Society,* **101**, 7-55.

McWhorter, J. (2001). *The Power of Babel: A Natural History of Language*. Times Books, Henry Holt and Company, New York.

Moore, J. H. (1994). Putting anthropology back together again: the ethnographic critique of cladistic theory. *American Anthropologist,* ***96***(4), 925-948.

Nei, M. (1991). Relative efficiencies of different tree-making methods for molecular data. In: *Phylogenetic analysis of DNA sequences*, (Miyamoto, M.M., & Cracraft, J., eds.), Oxford University Press, New York, pp. 90-128.

O'Brien, M.J. & R. L. Lyman, Y. Saab, E. Saab, J. Darwent, and D. S. Glover. (2002). Two issues in archaeological phylogenetics: Taxon construction and outgroup selection. *Journal of Theoretical Biology,* **215**, 133-50.

Pagel, M. (2000). Maximum-likelihood models for glottochronology and for reconstructing linguistic phylogenies. In: *Time Depth in Historical Linguistics* (Renfrew, C., McMahon, A. & Trask, L. eds.). The McDonald Institute for Archaeological Research, Cambridge, pp.189-207.

Renfrew, C. (2000). 10,000 or 5,000 years ago? Questions of time depth. In: *Time Depth in Historical Linguistics* (Renfrew, C., McMahon, A. & Trask, L. eds.). The McDonald Institute for Archaeological Research, Cambridge, pp. 413-439

Rexova, K., Frynta, D. & Zrzavy, J. (2003). Cladistic analysis of languages: Indo-European classification based on lexicostatistical data. *Cladistics* **19**, 120-127

Ringe, D., Warnow, T. & Taylor, A. (2002). IndoEuropean and computational cladistics. *Transactions of the Philological Society,* **100**, 59-129

Saitou, N. & Nei, M. (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, **4**, 406-425.

Sankoff, D. (1970). On the rate of replacement of word-meaning relationships. *Language,* **46**, 564-569.

Sankoff, D. (1973). Mathematical developments in lexicostatistical theory. In: *Current trends in linguistics 11: Diachronic, areal and typological linguistics*. (Sebeok, T. A. ed.). Mouton, The Hague, pp. 93-112.

Semple, C. & Steel, M. (2003). *Phylogenetics*. Oxford University Press, Oxford.

Strimmer, K., Wiuf, C. and Moulton, V. (2001). Recombination analysis using directed graphical models. *Molecular Biology and Evolution,* **18**, 97--99.

Swadesh, M. (1950). Salish internal relationships. *International Journal of American Linguistics,* **16**, 157-167.

Swadesh, M. (1952). Lexico-statistic dating of prehistoric ethnic contacts. *Proc. Am. Phil. Soc.* **96**, 453-463.

Swadesh, M. (1955). Toward greater accuracy in lexicostatistical dating. *International Journal of American Linguistics,* **21**, 121-137.

Terrell, J. (1988). History as a family tree, history as an entangled bank: constructing images and interpretations of prehistory in the South Pacific. *Antiquity,* **62**, 642-657.

Terrell, JE, Kelly, KM and Rainbird, P. (2001). Foregone conclusions? In search of 'Austronesians' and 'Papuans.' *Current Anthropology*, **42**, 97-124.

Trask, R. L. (1996). *Historical Linguistics*. Oxford University Press, London.

Figures

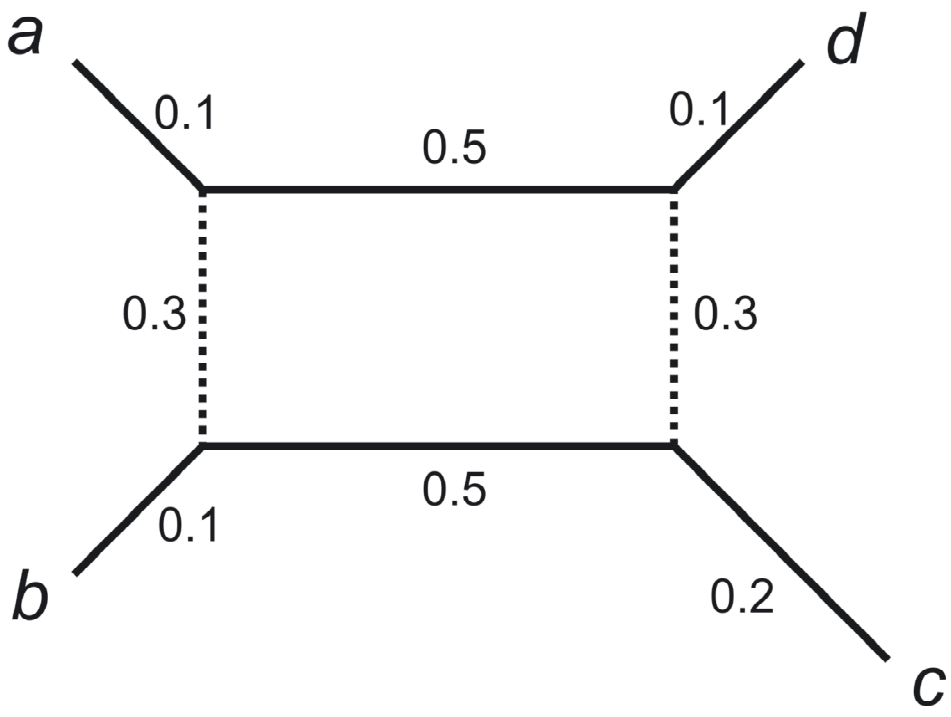

Figure 1. A splits graph network for four taxa. The distance between any two nodes equals the sum of the weights on the path between them. The two horizontal edges correspond to the split ab|cd while the dotted vertical edges correspond to the split ad|bc.

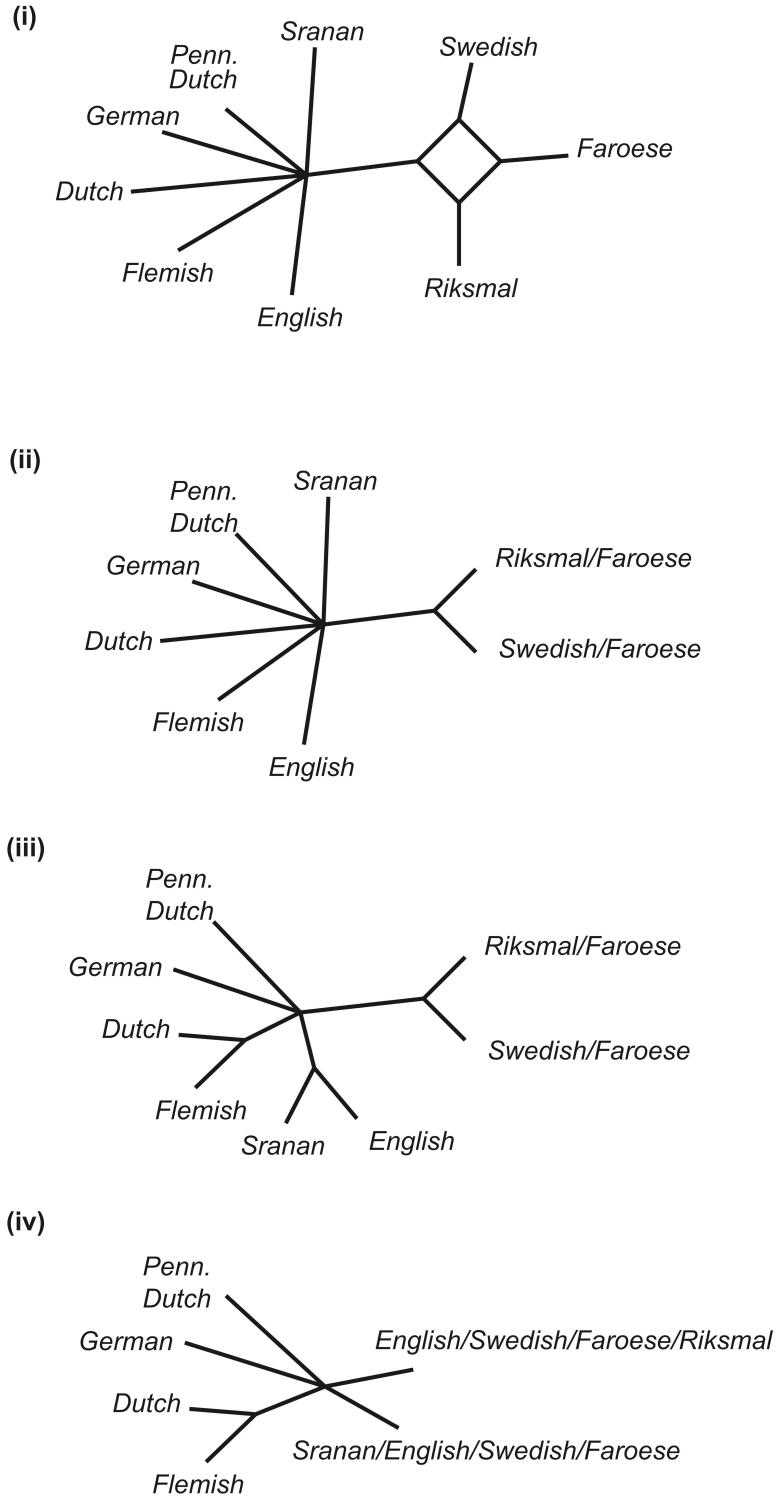

Figure 2. Neighbor-Net construction in process. The distance matrix used in this analysis was computed from Germanic lexical data (see Table 2). The steps in the NeighbourNet construction are detailed in the text.

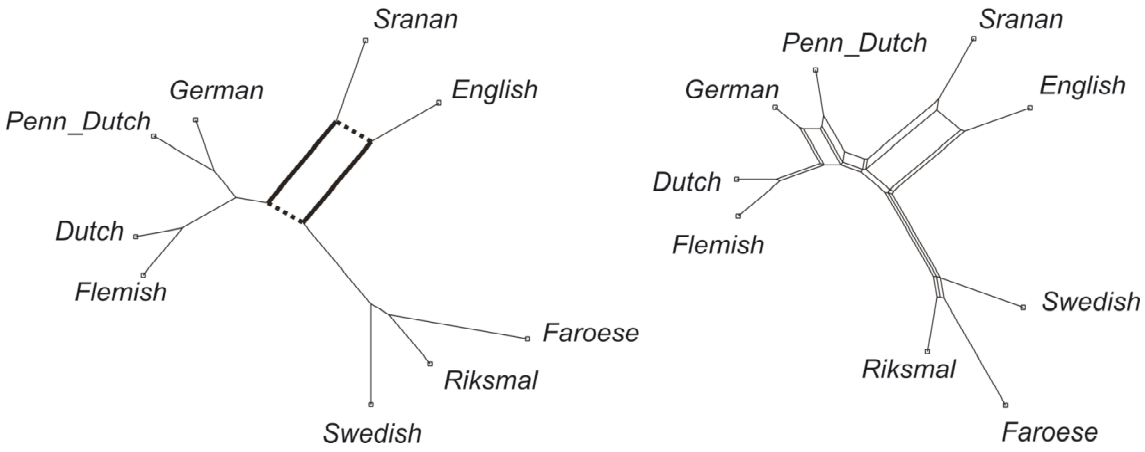

Figure 3. The networks produced by split decomposition (left) and NeighborNet (right) for the Germanic data. Both networks represent the conflicting signal introduced by the creole Sranan. NeighborNet also detects the presence of cognates separating Flemish, Dutch and German from the remaining Germanic languages. In the split decomposition graph (left), the split grouping Sranan and English is shown in bold, while the conflicting split grouping Sranan with German, Penn. Duct, Dutch, and Flemish is shown as a dotted line.

Figure 4. Network produced by the NeighbourNet analysis of a distance matrix for 84 Indo-European languages. The network is generally quite tree-like, with the exception of a few cases of strong conflicting signals. Split one shows a small conflicting signal grouping English and Sranan with Italic rather than Germanic languages. Splits two and three show substantial conflict in the relationships of Romanian and Vlach.

Table 1. Cognate sets for a selection of Indo-European languages. Non-cognate terms are shown crossed out (e.g. ~~*cos*~~). Separate cognate sets within a semantic category are denoted with a superscript (e..g *mechli*^{2} in the 'FISH' column). Anglicized orthography is used. The data were sourced from the "Comparative Indo-European data corpus" compiled by Isidore Dyen (see Dyen et al., 1992), with our addition of Latin. The Dyen et al data is available at `http://www.ntu.edu.au/education/langs/ielex/HEADPAGE.html`.

	'FATHER'	**'FOOT'**	**'FOUR'**	**'FISH'**	**'FIVE'**
Greek	*pateras*	*podhi*	*teseris*	~~*psari*~~	*pende*
Armenian	*hayr*	*ot*	*c'ors*	~~*juk*~~	*hing*
Irish	*athair*	~~*cos*~~	*cesthair*	*iasc*	*cuig*
Afghan	*plar*	*psa*	*calor*	~~*kab*~~	*pindze*
Lahnda	*pyu*	*paer*	*car*	*mechli*^{2}	*penj*
Baluchi	*phith, pith*	*phadh*	*chiar*	*mahi*^{2}	*phanch*
Ossetic	*fyd*	*fad*	*cupppar*	~~*kaesag*~~	*fondz*
Tadzik	*padar*	*po, poj*	*cor*	*moxi*^{2}	*panc*
Persian	*pedar*	*pa*	*chahar (char)*	*mahi*^{2}	*panj*
Latin	*pater*	*pes*	*quattuor*	*piscis*	*quinque*
Romanian	~~*tata*~~	*picior*	*patru*	*peste*	*cinci*
Italian	*padre*	*pie, piede*	*quattro*	*pesce*	*cinque*
Spanish	*padre*	*pie*	*cuatro*	*pez*	*cinch*
Catalan	*pare*	*peu*	*quatre*	*peix*	*cinch*
French	*pere*	*pied*	*quatre*	*poisson*	*cinq*
Walloon	*pere*	*pi*	*cwate, qwate*	*pehon*	*cinq'*
Provencal	*paire, pai*	*ped*	*quatre*	*peis, peossoun*	*cincq*
Brazilian	*pai*	*pe*	*quatro*	*piexe*	*cinco*
Portuguese	*pai*	*pe*	*quatro*	*peixe*	*cinco*
Frisian	*faer*	*foet*	*fjouwer*	*fisk*	*fiif*
Faroese	*fadir*	*fotur*	*fyra*	*fiskur*	*fimm*
Danish	*fader*	*fod*	*fire*	*fisk*	*fem*
Swedish	*fader*	*fot*	*fyra*	*fisk*	*fem*
Riksmal	*far*	~~*ben*~~	*fire*	*fisk*	*fem*
Icelandic	*faoir*	*fotr*	*fjorir*	*fiskr*	*fimm*
English	*father*	*foot*	*four*	*fish*	*five*
German	*vater*	*fuss*	*vier*	*fische*	*funf*
Dutch	*vader*	*voet*	*vier*	*visch*	*vijf*
Penn. Dutch	*fotter*	*fuusz*	*vier*	*fisch*	*finfe*

Table 2. Mean percent difference in cognacy between Germanic languages based on the Swadesh 200 word list. Distances calculated from the the cognate sets in the “Comparative Indo-European data corpus” (Dyen et al., 1992).

German	0.000								
Penn_Dutch	0.144	0.000							
Dutch	0.203	0.269	0.000						
Flemish	0.228	0.278	0.125	0.000					
Swedish	0.406	0.467	0.425	0.456	0.000				
Riksmal	0.406	0.417	0.425	0.439	0.217	0.000			
Faroese	0.508	0.486	0.544	0.553	0.297	0.236	0.000		
English	0.419	0.408	0.411	0.436	0.453	0.442	0.506	0.000	
Sranan	0.383	0.350	0.397	0.394	0.522	0.489	0.558	0.236	0.000

Table3. The set of splits resulting the split decomposition analysis of the data in Table 2. The observed and least-squares weights are shown below each split.

German	1	1	1	1	1	1	1
Penn_Dutch	1	1	1	1	1	1	1
Dutch	1	1	1	1	1	0	0
Flemish	1	1	1	1	1	0	0
Swedish	1	1	0	0	0	1	0
Riksmal	1	0	0	0	0	1	0
Faroese	1	0	0	0	0	1	0
English	0	1	1	0	0	1	0
Sranan	0	1	1	1	0	1	0
$w(A\|B)$	0.0945	0.0055	0.1115	0.0305	0.022	0.0475	0.0445
LS weight	0.1223	0.0236	0.1219	0.0468	0.037	0.0713	0.0388

Table 4. Cognate sets showing conflicting signal for the position of Sranan. Separate cognate sets in a column are numbered. Cognates of the Sranan term are shown in bold.

	"because"	**"dirty"**	**"dog"**	**"dull knife"**	**"bad"**
Flemish	*omdat*[1]	*vies*[1]	*hond*[1]	———	***sleg***[1]
Dutch	*omdat*[1]	*vuil*[1]	*hond*[1]	***stomp***[1]	***slecht***[1]
Penn. Dutch	*weil*[2]	*dreckich*[2]	*huundt*[1]	***schtuump***[1]	***schlecht***[1]
German	*weil*[2]	*dreckig*[2], *schmutzig*[3]	*hund*[1]	***stumpf***[1]	***schlecht***[1]
Sranan	***bikasi***[3]	***doti***[4]	***dagoe***[2]	***stompoe***[1]	***slekti***[1]
English	***because***[3]	***dirty***[4]	***dog***[2]	*dull*[2]	*bad*[2]
Swedish	*emedan*[4]	*smutsig*[3]	*hund*[1]	*slog*[3]	*dalig*[3]
Faroese	*(av) ti (at)*[5]	*skitin*[5]	*hundur*[1]	*ohvassur*[4]	*illur*[4]
Riksmal	*fordi*[6]	*skidden*[5]	*hund*[1]	*slov*[3]	*darlig*[3]

Table 5. Characters that produce some of the conflicting signals for the Germanic languages shown in figure 4.

	“to blow”	**“to blow”**	**“smoke”**	**“smoke”**
Flemish	*waeijen*	———	*rook*	———
Dutch	*waaien*	*blazen*	*rook*	———
Penn. Dutch	———	*bloesz*	———	*schmoeck*
German	*wehen*	———	*rauch*	———
Sranan	———	*blo*	———	*smoko*
English	———	*blow*	———	*smoke*
Swedish	———	*blasa*	*rok*	———
Faroese	———	*blasa*	*roykur*	———
Riksmal	———	*blase*	*rok*	———