ONLINE SUPPLEMENTARY FILE – A SHORT DESCRIPTION OF THE ALGORITHMS FOR MINIMUM SPANNING NETWORKS, STATISTICAL PARSIMONY, AND MEDIAN NETWORKS

Minimum Spanning Networks (MSN)

Several algorithms exist for constructing MSN, but a useful one is given by Bandelt *et al.* (1999):

1. Calculate the distance between any two sequences and arrange these in increasing order.
2. Let δ_i be the smallest distance observed in the sample (which may be observed more than once). At the beginning, no sequences are connected and each can be thought of as a sub-network.
3. Connect all haplotypes of *different* sub-networks that are at distance δ_i.
4. Repeat step (3) with distance δ_{i+1} (the second smallest) and so on until a network of all haplotypes is formed.

This algorithm is analogous to the algorithm described by Excoffier and Smouse (1994) and implemented in the widely used program ARLEQUIN (Excoffier et al., 2005) (see Bandelt et al., 1999).

Statistical Parsimony (SP)

Statistical parsimony consists of two steps: first the limit of parsimony is evaluated and then the network summarising all most parsimonious solutions is constructed within this parsimony limit, which means that haplotypes distant by more mutations than the parsimony limit will not be connected. Described below is the algorithm implemented in the program TCS (ver. 1.21; Clement et al., 2000).

Evaluating the limit of parsimony.—A limit of parsimony of x steps means that only haplotypes that differ by x nucleotides or less have a probability greater than 95% (default value) to be parsimonious (i.e. of not containing homoplasious mutations). This limit is calculated using a Bayesian estimator and formulas for calculating it are given in Templeton *et al.* (1992).

Constructing the network.—The following description for reconstructing statistical parsimony networks is taken from Clement et al. (2002). First, the algorithm calculates the distance between each taxa and every other taxa. This calculation is performed by comparing the characters for each sequence and recording the raw number of changes between the sequences. Once the distance matrix has been computed, the TCS algorithm proceeds to connect the taxa into a cladogram using the following algorithm:

1. A cluster is created for each of the N taxa in the sequence file.
2. The distance matrix is examined to determine which two clusters have the minimum distance M. This distance is computed by taking each taxa in one cluster and finding the taxa in each of the other clusters that has the smallest number of changes.
3. All of the taxa in the two minimum distance clusters that have distance M are then joined. For this discussion we assume that taxa S in the source cluster and taxa D in the destination cluster are two of these taxa that have distance M. Connections that

have a distance greater than 1 will be made by adding intermediates in the following way:

3.1. The minimum number of intermediates should be added to make the distance between the two taxa correct while preserving other distances in the matrix. Intermediates from another connection can be used in joining a pair of taxa as long as the connection does not form a connection that is shorter than the minimum distance between any two taxa in the source and destination clusters. This step is implemented using the following algorithm

3.1.1. All of the possible connections between *S*, an intermediate in the source cluster, an intermediate in the destination cluster and *D* are evaluated.

3.1.2. These connections are evaluated to determine which connection has the maximum metric. The metric is computed in the following way:

3.1.2.1. The distance between every pair of taxa in the source and destination cluster is examined to determine the global quality of a possible connection. A distance metric is created by comparing each of these possible distances with the real distance computed from the sequence file.

3.1.2.2. If a possible connection creates a distance that is correct, the metric is incremented by 20 points.

3.1.2.3. If the distance is shorter than the correct distance, but longer then the minimum distance for the taxa, then the metric is decremented by 10 points.

3.1.2.4. If the distance is less than the minimum for the taxa, then the metric is set to negative infinity (to indicate that this connection is undesirable).

3.1.2.5. If the distance is longer than the correct value, then the metric is decremented by 5 points.

3.1.3. The connection with the best metric is made in the tree data structure.

3.2. Combine the two clusters into one, reducing the number of clusters by one.

4. If there is more than one cluster, go to step 2.

Full Median Networks (MN)

Full Median networks, hereafter simply referred to as Median Networks (MN), were first described by Bandelt *et al*. (1995), but Huber *et al.* (2001) also explain the method with a useful example. Briefly,

1. The matrix is first recoded in binary (0-1) characters and all constant sites are excluded. The way this could be done for DNA characters with more than two states at a site is described elsewhere (Bandelt et al., 1995).
2. Then, the matrix is reduced by pooling identical sequences together (to obtain haplotypes) and by representing identical site patterns with only one column, although the number of times each pattern is found in the original matrix constitutes the weight of the pattern
3. Then, for each triplet of haplotypes, the median haplotype is computed by attributing, at each site, the majority state among the three haplotypes. For example, the median haplotype of [0011], [0100] and [1001] is [0001]. If the median haplotype obtained is identical to one of the haplotypes already present in the matrix, it is simply ignored. This is done for every triplet of haplotypes, including the newly formed median haplotypes

4. When no more median haplotypes can be obtained, the median network can be constructed: each haplotype (including median sequences) represents a node on the network and the edges are simply constructed by connecting nodes that differ by a single character in the new matrix. The length of each edge (and their parallel edges when boxes are formed) corresponds to the weight that was associated with the site in the recoded matrix.

Literature Cited

Bandelt, H.-J., P. Forster, B. C. Sykes, and M. B. Richards. 1995. Mitochondrial portraits of human populations using median networks. Genetics 141:743-753.

Bandelt, H.-J., P. G. Forster, and A. Röhl. 1999. Median-joinning networks for inferring intraspecific phylogenies. Mol. Biol. Evol. 16:37-48.

Clement, M., D. Posada, and K. A. Crandall. 2000. TCS: a computer program to estimate gene genealogies. Mol. Ecol. 9:1657-1659.

Clement, M., Q. Snell, P. Walker, D. Posada, and K. A. Crandall. 2002. TCS: estimating gene genealogies *in* First IEEE International Workshop on High Performance Computational Biology (HiCOMB).

Excoffier, L., L. G. Laval, and S. Schneider. 2005. Arlequin ver. 3: An integrated software package for population genetics data analysis. Evol. Bioinform. Online 1:47-50.

Excoffier, L., and P. E. Smouse. 1994. Using allele frequencies and geographic subdivision to reconstruct gene trees within a species: molecular variance parsimony. Genetics 136:343-359.

Huber, K. T., V. Moulton, P. J. Lockhart, and A. Dress. 2001. Pruned median networks: a technique for reducing the complexity of median networks. Mol. Phylogenet. Evol. 19:302-310.

Templeton, A. R., K. A. Crandall, and C. F. Sing. 1992. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. III. Cladogram estimation. Genetics 132:619-633.