BIOINFORMATICS

Vol. 14 no. 1 1998
Pages 68–73

SplitsTree: analyzing and visualizing evolutionary data

Daniel H. Huson

FSPM, University of Bielefeld, 33501-Bielefeld, Germany

Received on July 1, 1997; revised on September 19, 1997; accepted on September 23, 1997

Abstract

Motivation: *Real evolutionary data often contain a number of different and sometimes conflicting phylogenetic signals, and thus do not always clearly support a unique tree. To address this problem, Bandelt and Dress (Adv. Math., **92**, 47-05, 1992) developed the method of split decomposition. For ideal data, this method gives rise to a tree, whereas less ideal data are represented by a tree-like network that may indicate evidence for different and conflicting phylogenies.*

Results: *SplitsTree is an interactive program, for analyzing and visualizing evolutionary data, that implements this approach. It also supports a number of distances transformations, the computation of parsimony splits, spectral analysis and bootstrapping.*

Availability: *There are two versions of SplitsTree: an interactive Macintosh version (shareware) and a command-line Unix version (public domain). Both are available from: ftp://ftp.uni-bielefeld.de/pub/math/splits/splitstree2. There is a WWW version running at: http://www.bibiserv.techfak.uni-bielefeld.de/splits.*

Contact: *huson@mathematik.uni-bielefeld.de*

Introduction

Evolutionary relationships between taxa are most often represented as phylogenetic trees, and many different algorithms for tree construction have been developed (Swofford *et al.*, 1996). This is, of course, justified by the assumption that evolution is a branching or tree-like process. However, a set of real data often contains a number of different and sometimes conflicting signals, and thus does not always clearly support a unique tree.

To address this problem, Bandelt and Dress (1992a) developed the method of split decomposition. 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 are 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 are.

There exist efficient algorithms for performing split decomposition (Bandelt and Dress, 1992a) and for computing splits graphs (Wetzel, 1995; D.H.Huson, in preparation). Dress and Wetzel produced a simple implementation of split decomposition (Wetzel, 1995) as an investigative tool to help develop the general theory. Based on their work, a first public version was developed by Wetzel and Huson (SplitsTree version 1). The program described in this paper (SplitsTree version 2) is a completely new implementation.

In this paper, we first review the concepts of splits, splits graphs and the method of split decomposition, and then discuss the SplitsTree program in detail. For a number of biological applications of the split decomposition method, see, for example, Bandelt and Dress (1992b), Dopazo *et al.* (1993), Dress and Wetzel (1993), Lockhart *et al.* (1995), Wetzel (1995), Dress *et al.* (1996), McLenachan *et al.* (1996) or P.J.Lockhart *et al.* (in preparation).

Splits and splits graphs

Evolutionary relationships are generally represented by a phylogenetic tree, T, i.e. a tree whose leaves are labeled by a set X of taxa and whose remaining vertices are unlabeled and of degree at least three. (We only consider unrooted trees in this paper.) Any edge e of T defines a split $S = \{A,\bar{A}\}$ of X, i.e. a partition of X into two non-empty sets A and $\bar{A}$, consisting of all taxa on the one side, or the other, of the edge e. Such a system Σ of splits is called compatible if, for any two splits $S_1 = \{A_1,\bar{A}_1\}$ and $S_2 = \{A_2,\bar{A}_2\}$ in Σ, one of the four intersections

$$A_1 \cap A_2,\ A_1 \cap \bar{A}_2,\ \bar{A}_1 \cap A_2,\ \text{or}\ \bar{A}_1 \cap \bar{A}_2$$

is empty. Any phylogenetic tree T gives rise to a compatible split system Σ. In 1971, Buneman established that, vice versa, any compatible split system Σ corresponds to a unique phylogenetic tree T. So, tree reconstruction for a given set of taxa X is equivalent to computing a compatible system of

splits Σ for X and determining a weight for each split S that corresponds to the length of the associated edge.

Hence, to obtain more general graphs, one must consider less restricted systems of splits. Let X be a set of taxa. A system of splits Σ of X is called weakly compatible if, for any three splits S_1, S_2, S_3 and all $A_i \in S_i$ ($i = 1, 2, 3$), at least one of the four intersections

$$A_1 \cap A_2 \cap A_3,\ A_1 \cap A_2 \cap A_3,\ A_1 \cap A_2 \cap A_3, \text{ or } A_1 \cap A_2 \cap A_3$$

is empty (Bandelt and Dress, 1992a). So, in particular, any two splits are permitted to be incompatible. Intermediately, Σ is called circular if there exists an ordering $x_1, x_2, \ldots, x_m$ of the taxa such that for every split $S \in \Sigma$ there exists $A \in S$ with $A = \{x_{p(S)}, x_{p(S)+1}, \ldots, x_{q(S)}\}$ and $1 \leq p(S) \leq q(S) \leq m$. One can prove that a circular split system is always weakly compatible and a compatible split system is always circular (Bandelt and Dress, 1992a; Wetzel, 1995).

A splits graph representing a weakly compatible split system Σ is a graph $G(\Sigma) = (V,E)$ whose vertices $v \in V$ are labeled by the set of taxa X and whose edges $e \in E$ are straight-line segments that represent the splits in Σ (see Figure 1). More precisely, each split $S = \{A,A\} \in \Sigma$ is represented by a band of parallel edges of equal length in such a way that deleting all edges in such a band partitions the graph into precisely two components: one containing all vertices labeled by taxa in A and the other containing all vertices labeled by taxa in A. The length of the edges representing a given split S indicates its weight or support and is calculated as the isolation index of S. For algorithms that compute splits graphs, see Wetzel (1995) and D.H.Huson (in preparation).

Consider a weakly compatible system of splits Σ of a set X of taxa. If Σ is compatible, then $G(\Sigma)$ is a phylogenetic tree. What is the situation if Σ is merely circular? Then $G(\Sigma)$ can be realized as a planar graph (Wetzel, 1995; D.H.Huson, in preparation). Finally, if Σ is not circular, then in general $G(\Sigma)$ will not be planar. In biological applications, the arising split systems are often either circular or mildly non-circular.

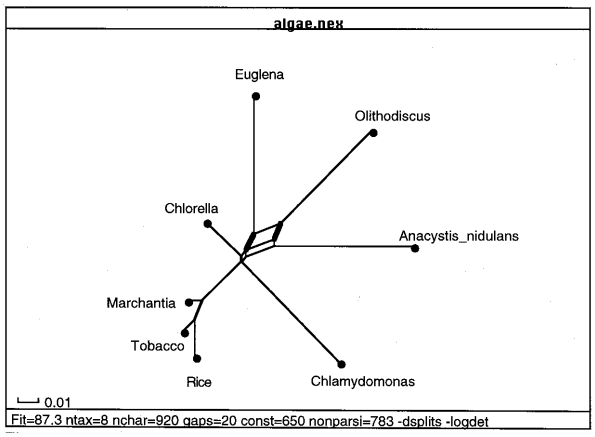

Fig. 1. The splits graph for the distances listed in Figure 3. Each band of parallel edges indicates a split. For example, the two bold lines represent the split {*Euglena*, *Olithodiscus*} versus the other taxa. The distance between any two taxa x and y corresponds to the sum of weights of all splits that separate x and y, i.e. the sum of edge lengths of any shortest path from x to y.

Split decomposition

Split decomposition is a method for obtaining a system of weakly compatible splits with weights from a given set of evolutionary data. So, assume we are given a set of taxa X and a distance map $d: X*X \rightarrow R^{\geq 0}$ on X, i.e. a matrix representing the evolutionary distances between pairs of taxa. Bandelt and Dress (1992a) showed that such a distance map d has the following canonical decomposition:

$$d = \Sigma\, \alpha_S \cdot \delta_S + d_0$$

Here, we sum over all possible splits S of X; the map $\delta_S: X*X \rightarrow R^{\geq 0}$ is the split metric on S that equals 1 if x and y lie on different sides of S, and 0 otherwise; the number $\alpha_S \geq 0$ is the weight or isolation index of the split S; and the map $d_0: X*X \rightarrow R^{\geq 0}$ is the so-called split-prime residue and cannot be decomposed further. A split S with $\alpha_S > 0$ is called a d-split, and the system Σ of all d-splits is weakly compatible and can be computed efficiently [see Bandelt and Dress (1992a) for details].

If there is no split-prime residue, then the distance between any two taxa x and y is precisely equal to the sum of weights of all d-splits that 'separate' x and y, and thus proportional to the sum of all edge lengths along a shortest path from x to y in the splits graph. However, in general, the split-prime residue will be positive and so the sum of weights will only give an approximation (from below) of the original distances. The fit of the approximation is measured by the sum of all approximated distances divided by the sum of all original distances. In biological applications, the fit is often quite high and a small split-prime residue can be considered as 'noise'.

If we are given a set of aligned sequences, then to apply split decomposition we must first compute a distance matrix d using an appropriate distance transformation. Alternatively, one can compute the so-called parsimony splits, or p-splits, directly from the sequences, as described in Bandelt and Dress (1993). Yet another possibility is to use spectral analysis (Hendy and Penny, 1992; M.D.Hendy and P.J.Waddell, in preparation) to assign a weight (the so-called γ-value) to each possible split of X. One can then greedily extract a weakly compatible (or compatible) system of splits, i.e. by considering all such splits S in decreasing order of weight

and inserting the split S into Σ if it is weakly compatible (or compatible) with all splits already in Σ.

Description of SplitsTree

SplitsTree is an easy-to-use Macintosh application that takes as input a file containing sequences, distances, or a system of splits, and produces as output a weakly compatible system of splits and a splits graph representing the given data. It contains a number of transformations to obtain distances from sequences and methods for obtaining compatible or weakly compatible split systems from distances or sequences.

Menus

SplitsTree offers the following menus: File, Edit, Layout, Options, Method and Window. The File menu contains the usual items for opening, closing, saving and printing documents. The Edit menu contains items for copying and pasting, etc.

The first group of items in the Layout menu can be used to change the position, orientation and size of the displayed splits graph. The Cycle item allows the user to specify the circular order in which the taxa appear around the outside of the splits graph. This feature can be used to produce the same layout for different splits graphs produced from the same data set by different methods. The Vertex Labels and Edge Labels submenus can be used to decide whether the vertices are to be labeled by the names or numbers of the taxa and whether the edges are to be labeled by weight, number or bootstrap support. The Equal Edges and To Scale items determine whether the edges of the displayed splits graph are drawn all with the same length, or in proportion to the isolation index of the corresponding splits.

The Options menu determines how the given data are preprocessed. The Taxa item enables the user to exclude certain taxa from the analysis. Similarly, the Sites item can be used to exclude certain sites and also codon positions. Moreover, items are available for excluding whole groups of sites: Exclude Gaps, Exclude Missing, Exclude Non Parsimony and Exclude Constant…. In the latter case, one can choose to exclude only a proportion of the constant sites, which can be useful in connection, for example, with the LogDet transformation (see Figure 1), as it provides a way of approximating a more continuous distribution for rates across sites (Adachi and Hasegawa, 1995; Waddell, 1996).

The Options menu also offers a number of distance transformations such as Hamming distances, Kimura 3ST (Kimura, 1981), Jukes Cantor (Jukes and Cantor, 1969) and LogDet (Steel, 1994). The Nei Miller item is for computing distances for restriction site data (Nei and Miller, 1990), and the PAM 250 item applies to protein data (Dayhoff *et al.*, 1983). A user-defined weight matrix can be supplied using the User Matrix item.

Moreover, there are two items for determining distances between groups of taxa, both suggested by Mike Steel: the Fitch Sidow… item computes the distances between given groups using a combination of methods from Fitch (1971) and Sidow *et al.* (1992), whereas the Covarion… item is based on Moulton *et al.* (1997).

Finally, SplitsTree checks for given distance data whether the triangle inequalities hold. If they do not, then the Force Triangle Inequalities item can be used to force them to, i.e. by adding an appropriate offset to all distances.

The Method menu is the most important menu, as it determines which method is applied to produce a split system from the given data. The first group of items all produce weakly compatible split systems. The choices are: Split Decomposition (as described above), Parsimony Splits (Bandelt and Dress, 1993) and Spectral Analysis… (Hendy and Penny, 1992; M.D.Hendy and P.J.Waddell, in preparation, followed by a greedy selection of a weakly compatible split system). The second group of items all produce compatible split systems: Buneman Tree (Buneman, 1971; Bandelt and Dress, 1992a), P-Tree (Bandelt and Dress, 1993) and Spectral Tree (spectral analysis followed by a greedy selection of a compatible split system).

For larger data sets, methods such as split decomposition or computing the 'Buneman tree' tend to produce unresolved split systems. This is because they involve computing the minimum of a certain index over all quartets of taxa that are separated by a given split to determine whether that split should be included in the split system (Bandelt and Dress, 1992a). In an attempt to solve this problem, one can replace the minimum by the average over a given number of quartets with smallest indices to obtain a refined system of splits, as suggested in Moulton *et al.* (1997). The Refine menu item implements this idea.

The Bootstrap item runs bootstrap sampling from given sequence data (Felsenstein, 1985). This is a way to test the statistical robustness of the computed splits graph. To be precise, the program repeatedly generates new artificial data sets by randomly choosing k (not necessarily distinct) sites in the original data set. The user is prompted to supply the number of times this is done, whereas k usually equals the length of the original sequences. For each such data set, the splits graph is then computed. At the end of this procedure, each split in the original splits graph is labeled by the percentage of computed splits graphs that it occurred in, thus indicating the statistical robustness of each split. The st_bootstrap block contains a full listing of all splits that occurred.

Finally, the Window menu contains a Syntax and Show submenu that can be used to obtain a listing of the syntax or current contents of a selected 'nexus block'. The Get Info item gives general information on the current document. Moreover, the menu contains a list of the currently open windows.

Windows

SplitsTree displays two windows. The SplitsTree Console is used to print messages when reading or computing data. It also accepts typed commands and nexus blocks. Moreover, it is used to present information requested using the menu items described in the preceding paragraph. The second window, called the document window, displays the splits graph computed for the given data set. The bottom of this window contains a line of information on the current data and how they were computed (see Figure 1).

The splits graph displayed in the document window can be manipulated using the mouse. Clicking on an edge will highlight that edge and all other edges representing the same split. Then, grabbing and dragging any other part of the graph will rotate the selected edges and thus reshape the graph, without changing any of the edge lengths. Moreover, the vertex labels can also be grabbed and dragged.

File format

SplitsTree is based on the new nexus format (Maddison *et al.*, 1995), which was originally developed for the programs PAUP (Swofford, 1997) and MacClade (Maddison and Maddison, 1989). Input data are described in the three standard block types: taxa, characters and distances. More precisely, an input file will typically consist of a taxa block listing the names of the given taxa and either a characters block containing a set of, for example, DNA, RNA, protein or RFLP sequences, or a distances block containing a distance or dissimilarity matrix. In Figure 2, we describe the syntax of these blocks and in Figure 3 an example input file is given.

An output file typically contains a number of additional blocks that are computed by SplitsTree and are specific to the program. The names of such blocks all have the prefix 'st_'. The st_splits, st_graph and st_assumptions blocks contain the split system, the splits graph and the assumptions made, respectively. More precisely, the latter block describes how the data were processed, e.g. whether sites were excluded, which distance transformation was applied, and which method was used to compute the splits, in other words, which items from the Options and Method menus were in effect.

Additionally, the program will generate a st_spectra block if spectral analysis was used, a st_bootstrap block if bootstrapping was applied, or an st_extras block if one of the additional computations offered by the program was employed. As mentioned above, the program offers an online description of the syntax of all blocks that it understands.

Implementation

This paper describes the interactive Macintosh version of SplitsTree, which is based on a kernel program that is essentially a nexus interpreter that reads nexus blocks from a file

```
------------------------------------------------------------
#NEXUS
BEGIN TAXA;
    DIMENSIONS NTAX=number-of-taxa;
    TAXLABELS taxon_1 taxon_2 ... taxon_ntax;
END;

BEGIN CHARACTERS;
    DIMENSIONS [NTAX=number-of-taxa] NCHAR=number-of-
                                                characters;
    [FORMAT
        [DATATYPE={STANDARD|DNA|RNA|PROTEIN}]
        [MISSING=symbol]
        [GAP=symbol]
        [SYMBOLS="symbol symbol ..."]
        [[NO] LABELS]
        [[NO] TRANSPOSE]
        [[NO] INTERLEAVE]
    ;]
    [CHARWEIGHTS wgt_1 wgt_2 ... wgt_nchar;]   (*)
    MATRIX
        sequence data in the specified format
    ;
END;

BEGIN DISTANCES;
    [DIMENSIONS [NTAX=number-of-taxa] [NCHAR=number-of-
                                                characters];]
    [FORMAT
        [TRIANGLE={LOWER|UPPER|BOTH}]
        [[NO] DIAGONAL]
        [[NO] LABELS]
        [MISSING=symbol]
    ;]
    [[NO] FORCE_METRIC [:OFFSET=number];];     (*)
    MATRIX
        distance data in the specified format
    ;
END;

------------------------------------------------------------
```

Fig. 2. Syntax of the three main input blocks. In this figure, square brackets indicate optional items and curly brackets indicate a choice of items. The syntax follows the standard definition of these blocks (Maddison *et al.*, 1995), expect for the two additional commands marked by a (*). The CHARWEIGHTS item is used to enter weights when specifying RFLP data. The FORCE_METRIC item can be set by the program when the triangle inequalities do not hold and an offset must be added to force them to.

or the keyboard and outputs nexus blocks and PostScript. The kernel is written in C++ and thus can be compiled on any computer, and executables are available for a number of different Unix systems. We plan to develop an interactive Windows version in the future.

Example

The splits graph depicted in Figure 1 was obtained by applying the LogDet transformation and split decomposition to all sites in an rDNA data set (indicated in Figure 3). For these data, the splits graph in Figure 1 reveals that a conflicting relationship exists between the cyanobacterium *Anacystis* and the chloroplasts of *Euglena* and *Olithodiscus*. Previous biological studies suggest that the correct split within this unresolved part of the splits graph should actually put *Euglena* (a chlorophyll *a*/*b*-containing plastid) together with the other chlorophyll *a*/*b*-containing taxa (rice, tobacco, *Marchantia, Chlamydomonas*, *Chlorella*). That is, *Euglena* is expected to split away from the outgroup *Anacystis* and *Olithodiscus* (a chlorophyll *a*/*c*-containing plastid). The suggested reason for the conflicting signal is that the rDNA sequences in *Euglena* and *Olithodiscus* have independently and conver-

```
#NEXUS [Comments come in square brackets]
[! A comment starting with "!" is printed when read]
BEGIN taxa;
    DIMENSIONS ntax=8;
TAXLABELS Tobacco Rice Marchantia Chlamy Chlorella Euglena
          An_nidul Olithodiscus;
END;
BEGIN characters;
    DIMENSIONS nchar=920;
    FORMAT datatype=RNA gap=- labels interleave;
    MATRIX
    Tobacco       AAGAACCUGCCCUUGGGAGCUGGAAACGGCUGCUAAUACCCC
    Rice          AAGAACCUGCCCUUGGGAACUGGAAACGGUUGCUAAUACCCC
    Marchantia    AAGAACCUGCCCUUGGGAGCUGGAAACGGUUGCUAAUACCCC
    Chlamy        AAGAACCUACCUAUCGGAUUGGGAAACUGUUGCUAAUACCCC
    Chlorella     AAGAACCUACCUUUAGGAACUGGAAACGGUUGCUAAUACCCC
    Euglena       AAGAAUCUGCGCUUGGGAGAUGGAAACGUUUGCUAAUGCCUC
    An_nidul      GAGAAUCUGCCUACAGGAGUUGGAAACGACUGCUAAUACCCG
    Olithodiscus GAGAAUCUGCCUUUAGGAUUUGGAAACGAAUGCUAAUACCUU
       ...
    Tobacco       UCGCUAGUAAUCGCCGGUCAGAUACGGCGGUGAAUUCG
    Rice          UCGCUAGUAAUCGCCGGUCAGAUACGGCGGUGAAUCCG
    Marchantia    UCGCUAGUAAUCGCCGGUCAGAUACGGCGGUGAAUCCG
    Chlamy        UCGCUAGUAAUCGCCAGUCAGAUAUGGCGGUGAAUACG
    Chlorella     UCGCUAGUAAUCGCUGGUCAGAUACAGCGGUGAAUACG
    Euglena       UCGCUAGUAAUCGCCGGUCAGAUACGGCGGUGAAUACG
    An_nidul      UCGCUAGUAAUCGCAGGUCAGAUACUGCGGUGAAUACG
    Olithodiscus UCGCUAGUAAUCGCUGGUCAGACACAGCGGUGAAU-CG
    ;
END;

BEGIN distances;
    FORMAT triangle=LOWER diagonal labels;
    MATRIX
    Tobacco       0
    Rice          0.02815 0
    Marchantia    0.03241 0.04554 0
    Chlamy        0.12904 0.13856 0.11065 0
    Chlorella     0.08754 0.09909 0.06990 0.11317 0
    Euglena       0.15235 0.16207 0.13663 0.16205 0.12710 0
    An_nidul      0.14353 0.15509 0.14003 0.16569 0.13724
                  0.17907 0
    Olithodiscus 0.16007 0.16695 0.15076 0.18069 0.12851
                  0.15286 0.15208 0
    ;
END;
```

Fig. 3. Example of an input file. Typically, either a characters block or a distances block will be specified, but not both. The first token in a file must be '#NEXUS' and the first block must be the taxa block. Comments are enclosed in square brackets and all comments between the '#NEXUS' and 'BEGIN taxa' tokens are passed on to the output file by SplitsTree.

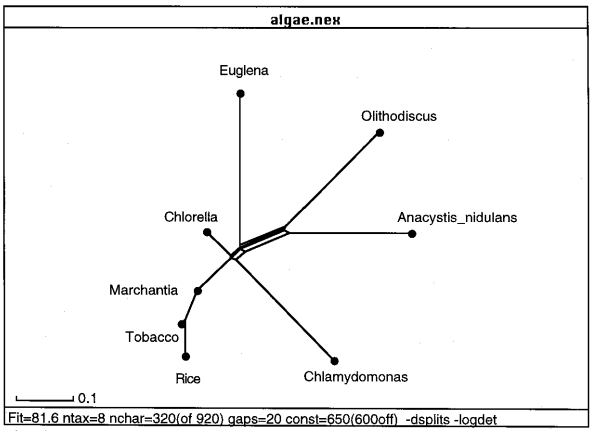

Fig. 4. The splits graph obtained from the RNA sequences indicated in Figure 3 using the LogDet transformation and split decomposition with 600 constant sites excluded. It contains a split that clearly separates *Euglena* from *Olithodiscus* and *Anacystis nidulans*, as discussed in the Example section.

gently acquired similar base compositions [see discussions in Lockhart *et al.* (1994), Delwiche and Palmer (1995) and Van der Peer *et al.* (1996)]. Hence, in this example, the splits graph indicates both the suggested true phylogenetic signal and a spurious one resulting from base composition effects.

Comparison of Figure 1 with Figure 4 reiterates the point made in Lockhart *et al.* (1994) that the LogDet correction, which can overcome some such base composition problems, will not work when invariable sites are included in sequence analyses. That is, the expected split is only obtained if one removes the invariable sites from the data, i.e. an appropriate number of constant sites (using the Exclude Constant Sites… item) before applying the LogDet transformation (Figure 4 displays the result for 600 constant sites excluded).

In practice, a number of techniques can be used to estimate the proportion of constant sites that should be removed from the data when accommodating position rate heterogeneity (e.g. Lockhart *et al.*, 1996). Note that the removal of invariable positions in sequences can be important before analyses are carried out using both symmetrical (e.g. Jukes Cantor) and asymmetrical (e.g. LogDet) correction formulae (Lockhart *et al.*, 1996).

Acknowledgements

SplitsTree was developed within the framework of a joint co-operation between researchers at Bielefeld University (Germany), Massey University (Palmerston North, New Zealand) and the University of Canterbury (Christchurch, New Zealand) with support from the German Ministry of Science and Technology (BMFT), the New Zealand Marsden Fund and the University of Canterbury. Thanks to the following people for their support and co-operation: Hans-Jürgen Bandelt, Andreas Dress, Mike Hendy, Pete Lockhart, Holger Paschke, Dave Penny, Mike Steel, Udo Tsnges and Rainer Wetzel. The Example section of this paper was written with the help of Pete Lockhart, who also suggested many improvements to the program and this paper. The WWW version of the program was produced with the help of Holger Paschke.

References

Adachi,J. and Hasegawa,M. (1995) Improved dating of the human/chimpanzee separation in the mitochondrial DNA tree: heterogeneity among amino acid sites. *J. Mol. Evol.*, **40**, 622–628.

Bandelt,H.-J. and Dress,A.W.M. (1992a) A canonical decomposition theory for metrics on a finite set. *Adv. Math.*, **92**, 47–05.

Bandelt,H.-J. and Dress,A.W.M. (1992b) Split decomposition: a new and useful approach to phylogenetic analysis of distance data. *Mol. Phylogenet. Evol.*, **1**, 242–252.

Bandelt,H.-J. and Dress,A.W.M. (1993) A relational approach to split decomposition. In Opitz,O., Lausen,B. and Klar,R. (eds), *Information and Classification.* Springer, Berlin, pp. 123–131.

Buneman,P. (1971) The recovery of trees from measures of dissimilarity. In *Mathematics and the Archeological and Historical Sciences.* Edinburgh University Press, pp. 387–395.

Dayhoff,M.O., Barker,W.C. and Hunt,L.T. (1983) Establishing homologies in protein sequences. *Methods Enzymol.*, **91**, 524–545.

Delwiche,C.F., Kushel,M. and Palmer,J.D. (1995) Phylogenetic analysis of *tuf*A sequences indicates a cyanobacterial origin of all plastids. *Mol. Phylogenet. Evol.*, **4**, 110–128.

Dopazo,J., Dress,A.W.M. and von Haeseler,A. (1993) Split decomposition: a new technique to analyse viral evolution. *Proc. Natl Acad. Sci. USA*, **90**, 10320–10324.

Dress,A.W.M. and Wetzel,R. (1993) The human organism—a place to thrive for the immuno-deficiency virus. In *Proceedings of IFCS.* Paris.

Dress,A.W.M., Huson,D.H and Moulton,V. (1996) Analyzing and visualizing sequence and distance data using splitstree. *Discrete Appl. Math.*, **71**, 95–109.

Felsenstein,J. (1985) Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, **39**, 783–791.

Fitch,W. (1971) Towards defining the course of evolution: minimum change for a specific tree topology. *Syst. Zool.*, **20**, 406–416.

Hendy,M.D. and Penny,D. (1992) Spectral analysis of phylogenetic data. *J. Classif.*, **10**, 5–24.

Jukes,T.H. and Cantor,C.R. (1969) Evolution of protein molecules. In Munro,H.N. (ed.), *Mammalian Protein Metabolism.* Academic Press, New York, pp. 21–132.

Kimura,M. (1981) Estimation of evolutionary distances between homologous nucleotide sequences. *Proc. Natl Acad. Sci. USA*, **78**, 454–458.

Lockhart,P.J., Steel,M.A., Hendy,M.D. and Penny,D.P. (1994) Recovering an evolutionary tree under a more realistic model of sequence evolution. *Mol. Biol. Evol.*, **11**, 605–612.

Lockhart,P.J., Penny,D. and Meyer,A. (1995) Testing the phylogeny of swordtail fishes using split decomposition and spectral analysis. *Mol. Evol.*, **41**, 666–674.

Lockhart,P.J., Larkum,A.W.D., Steel,M.A., Waddell,P.J. and Penny,D. (1996) Evolution of chlorophyll and bacteriochlorophyll: the problem of invariant sites in sequence analysis. *Proc. Natl Acad. Sci. USA*, **93**, 1930–1934

Maddison,W.P and Maddison,D.R. (1989) Interactive analysis of phylogeny and character evolution using the computer program MacClade. *Folia Primatol.*, **53**, 190–202.

Maddison,D.R., Swofford,D.L. and Maddison,W.P. (1995) NEXUS: An extendible file format for systematic information. *Syst. Biol.*, in press.

McLenachan,P.A., Lockhart,P.J., Faber,H.R. and Mansfield,B.C. (1996) Evolutionary analysis of the multigene pregnancy specific β1-glycoprotein family: separation of historical and non historical signals. *J. Mol. Evol.*, **42**, 273–280.

Moulton,V., Steel,M.A. and Tuffley,C. (1997) Dissimilarity maps and substitution models: some new results. *Proceedings of the DIMACS Workshop on Mathematical Hierarchies and Biology.* American Mathematical Society, in press.

Nei,M. and Miller,J.C. (1990) A simple method for estimating average number of nucleotide substitutions within and between populations from restriction data. *Genetics*, **1256**, 873–879.

Sidow,A., Nguyen,T. and Speed,T.P. (1992) Estimating the fraction of invariable codons with a capture-recapture method. *J. Mol. Evol.*, **35**, 253–260.

Swofford, D.L. (1997) *PAUP 5.0.* Sinaur Associates, Sunderland, MA.

Swofford,D.L., Olsen,G.J., Waddell,P.J. and Hillis,D.M (1996) Phylogenetic inference. In Hillis,D.M., Moritz,C. and Mable,B.K. (eds), *Molecular Systematics*, 2nd edn. Sinauer Associates, Sunderland, MA, pp. 407–514.

Steel,M.A. (1994) Recovering a tree from the leaf colorations it generates under a Markov model. *Appl. Math. Lett.*, **7**, 19–24.

Van de Peer,Y., Rensing,S.A., Maier,U.G and De Wachter,R. (1996) Substitution rate calibration of small ribosomal subunit RNA identifies chlorachniophyte endosymbionts as remnants of green algae. *Proc. Natl Acad. Sci. USA*, **93**, 7732–7736.

Waddell,P.J. (1996) Statistical methods of phylogenetic analysis, including Hadamard conjugations, LogDet transforms, and maximum likelihood. PhD Thesis, Massey University, New Zealand.

Wetzel,R. (1995) Zur Visualisierung abstrakter Ähnlichkeitsbeziehungen. PhD Thesis, University of Bielefeld.