My Attempt to Understand the Buneman-Steiner-Tree theorem in Phylogenetics
Dan Gusfield

November 15, 2022
Why do it?

The theorem that I call the Buneman-Steiner Theorem is widely stated and (I think) very important, but very few people know its proof because the main published proof (by Bandelt) is very mathematical and relies on another very difficult theorem. I have not mastered that proof.

So, I want to find a more direct and less-mathematical proof.
An alternative proof

There is a less cited proof of the theorem in a book by M. van de Vel (Theory of Convex Structures), which is actually more general, but also very abstract and mathematical. However, it seems more constructive and direct, and so I tried to translate it from mathematize to computer-scientist-friendly, and to specialize it to the case that we care about in Phylogeny.

I have not (yet) succeeded. But talking about my efforts may help me, and may be educational for you.
Hypercubes

The $m$-dimensional hypercube $H_m$ has $2^m$ nodes, each labeled by a distinct binary string of length $m$. Two nodes in $H$ are adjacent if and only if their labels differ at exactly one site (i.e., position).
Hypercubes $H_2$, $H_3$, $H_4$
Hypercube $H_{100}$
Why hypercubes in Phylogenetics?

Because
Buneman Graphs

Given an \( n \)-by-\( m \) binary matrix \( M \), the Buneman graph, \( B \), for \( M \) is a subgraph of the \( m \)-dimensional hypercube, whose nodes and edges will be defined below. An example of matrix \( M \) is:

\[
\begin{array}{cccc}
  s: & 1 & 2 & 3 & 4 \\
  \hline
  v1: & 1 & 1 & 0 & 0 \\
  v2: & 0 & 1 & 0 & 1 \\
  v3: & 1 & 1 & 0 & 1 \\
  v4: & 0 & 0 & 0 & 0 \\
  v5: & 0 & 0 & 1 & 1 \\
\end{array}
\]

We define a *gamete* as an ordered pair of binary values. For example, the two sites \( s1 \) and \( s2 \) in matrix \( M \) shown above have gametes 11, 01 and 00; but not 10.
Buneman Graphs

There are several equivalent definitions of a Buneman graph for \( M \). The following is a convenient definition for this note.

A node \( v \) in a hypercube \( H \), with label \( s \), is defined to be in the Buneman graph, \( B \), for \( M \), if and only if for each pair of sites \( i, j \), the gamete in \( s \) at sites \( i, j \) is the gamete at sites \( i, j \) in some string (row) of \( M \).

For example, if \( v \) is labeled 0110, then it will be a node in \( B \) if and only if there is a string 01xx in \( M \), and a string 0x1x, and a string 0xx0; and string x11x; and a string x1x0; and a string xx10 in \( M \), where \( x \) can be either 0 or 1.

There need not be a single string in \( M \) that satisfies all of these conditions.
Buneman Graphs - Equivalently

Equivalently, a node \( v \) in \( H \) is \textit{not} in \( B \) if and only if its label has a pair of sites \( i, j \) whose gamete is \textit{not} the gamete at sites \( i, j \) in any string in \( M \).

For example, since there is no string \( x11x \) in the matrix \( M \), the node \( v \) labeled 0110 is not in the Buneman graph for \( M \).
Buneman Graphs - Edges

We have defined the nodes of $B$ but not the edges. There is an edge between two nodes in $B$ if and only if the labels for those two nodes differ at exactly one site.

Note that the Buneman graph for a matrix $M$ is unique.
Example

The Buneman graph for the matrix $M$ above is shown is:

![Buneman graph diagram](image)

Figure: The Buneman graph for matrix $M$ shown above. Since $m = 4$, the hypercube for $M$ would have 16 nodes, and so would have 11 more nodes than the five nodes required for the five strings in $M$. The Buneman graph has only two more nodes than those five nodes. The optimal Steiner tree for $M$ has five edges.
Steiner Trees for $M$

Given an $n$-by-$m$ binary matrix $M$, with no duplicate rows or columns, we define the Steiner Tree for $M$ as a subtree of $H_m$ connecting the $n$ nodes of $H$ that are labeled by the strings (rows) of $M$.

An optimal Steiner Tree of $M$ is a Steiner Tree for $M$ with the minimum number of edges.

What does an optimal Steiner tree for $M$ mean for phylogeny?
Assumptions

We have already assumed that all the rows and columns of $M$ are distinct. As is well known, we can also assume that each column contains at least one 0 and one 1. Also

Lemma

*We can assume that for any pair of sites, $i, j$, in $M$, there are at least three distinct gametes in columns $i, j.*
The Buneman-Steiner Theorem

**Theorem**

*Given M, there is an optimal Steiner Tree T for M that is contained in the Buneman graph for M.*

Why do we care? Because for biological data in an \( n \)-by-\( m \) matrix \( M \), the Buneman graph for \( M \) is much smaller than the hypercube \( H_m \), and the Steiner-tree problem (even in hypercubes) is NPC. So, having a smaller graph is very helpful for practical computation.
How small?

Exclusion-Inclusion argument for a rough, genomic-relevant upper-bound on the number of nodes in a Buneman graph.

Exclusion-inclusion formula for the fraction of nodes of the hypercube knocked out as a function of the number of disjoint compatible pairs of columns.

For one pair (two cols.) the fraction removed is 1/4. For two pairs the number is 7/16. For three is is 37/64. For four disjoint pairs, it is

\[
\frac{4}{4} - \frac{6}{16} + \frac{4}{64} - \frac{1}{256} - \frac{96}{256} + \frac{16}{256} - \frac{1}{256} = \frac{175}{256} = 0.6835
\]

How fast does this approach 1? How do we handle non-disjoint compatible pairs?
There are empirical data where \( B \) is small, and other arguments hinting that \( B \) is expected to be small for biological data.
Proof attempt for the Steiner-tree Theorem

First, it will be convenient to use an equivalent definition of a Steiner tree for $M$ as a tree, $T$, connecting the nodes labeled by the strings in $M$; and possibly containing additional nodes, each labeled by a distinct $m$-length binary string, such that any two adjacent nodes in $T$ must be labeled by strings that differ in exactly one site.

This definition of a Steiner tree allows us to talk about changing the label of a node in $T$, or merging nodes, or sub-dividing an edge in $T$, without explicitly referring to $H$. 
Proof attempt

Let $T$ be an optimal Steiner tree for $M$ that (over all optimal Steiner trees) minimizing the number of nodes not in the Buneman graph $B$ for $M$.

If every node in $T$ has the label of a node in the Buneman graph $B$ for $M$, then $T$ is already contained in $B$, and there is nothing to prove.
Proof attempt

Otherwise, let \( v \) be a node in \( T \), but not in \( B \).

By the definition of the nodes in \( B \), there must be a pair of sites \( i, j \) such that the \( i, j \) gamete at node \( v \) is not the \( i, j \) gamete for any string in \( M \).

We call any such node an \( i, j \)-violating node; and call such a pair of sites \( i, j \) a violating pair.

For ease of exposition, and without loss of generality, suppose that the \( i, j \) gamete at node \( v \) is 01. Then, no string in \( M \) has the gamete 01 at sites \( i, j \). (But since every pair of sites has at least three gametes, the gametes 11, 00, and 10 do appear at sites \( i, j \).)
Example

\begin{tabular}{ccc}
\hline
s: & 1 & 2 & 3 & 4 \\
\hline
\end{tabular}

\begin{tabular}{c}
\hline
i & j \\
\hline
\end{tabular}

v1: 1 1 0 0 \\
v2: 0 1 0 1 \\
v3: 1 1 0 1 \\
v4: 0 0 0 0 \\
v5: 0 1 1 1 \\

There is no string in $M$ with gamete 01 at sites $i = 2, j = 4$. So, if Steiner-tree $T$ has a node $v$ labeled 1001, that is a 2,4-violating node, and 2,4 is a violating pair of sites, and node $v$ is not in $B$. 
Creating Steiner tree $T'$

Now, we will make some changes to $T$. First, change the value of site $i$ from 0 to 1, for every $i,j$—violating node $v$ in $T$. This changes the $i,j$ gamete at each such node from 01 to 11. Note that 11 is a non-violating $i,j$-gamete, since every pair of sites has at least three gametes.

After those $i,j$ changes, we will need to make additional changes so that the modified tree is a Steiner-tree. But we will first explain those with an example.
Figure: Tree $T$ is an optimal Steiner-tree for an input matrix $M$. Only the $i,j$ gametes are shown at each node. Tree $T'$ is $T$ after changing the $i,j$ gametes at nodes $b$ and $c$ from 01 to 11. Tree $T''$ is the modified $T$ after changing the $i,j$ gametes at $b$ and $c$ from 01 to 00.
A word is worth one-thousandth of a picture: The additional changes in writing

Second, if any $i, j$–violating node $v$ in $T$ has a neighbor $w$ in $T$ whose $i, j$ gamete is 11, merge nodes $v$ and $w$, i.e., contract the $(v, w)$ edge. Since $v$ and $w$ are neighbors in $T$, their labels differ at exactly one site, which must be site $i$. So, after the $i, j$ gamete at $v$ is changed from 01 to 11, the labels of $v$ and $w$ become identical, and hence $v$ and $w$ can be merged.

Third, if any $i, j$–violating node $v$ in $T$ has a neighbor $w$ in $T$ whose $i, j$ gamete is 00, insert a new node $w'$ in the $(v, w)$ edge, and give it a label that is identical to that of nodes $v$ and $w$ at every site other than $i$ and $j$. Then, set the $i, j$ gamete for node $w'$ to 10. Note that the label of node $w'$ differs from the label of $v$ in exactly one site, and similarly, differs from the label of $w$ in exactly one site. We call the resulting graph $T'$. 
It’s a tree

Note that no \( i,j-\text{violating} \) node \( v \) (with assumed \( i,j \) gamete 01) can have a neighbor \( w \) in \( T \) whose \( i,j \) gamete is 10, since then its label would differ from the label at \( v \) at more than one site. Hence, we have covered all of the cases of neighbors of \( v \).

To see that \( T' \) is a tree, note that \( T \) is changed to \( T' \) only through edge contractions and subdivisions, so the resulting graph remains connected. Further, edge contractions and subdivisions of a tree cannot create cycles, so \( T' \) is a connected graph with no cycles, i.e., a tree.
Actually, $T'$ is a Steiner-tree

Also, since no $i,j$-violating node can have a label that is a string in $M$, and only the labels of $i,j$-violating nodes are changed, every string in $M$ continues to label some node in $T'$. And, the labels of any adjacent nodes in $T'$ differ at exactly one site. Hence, $T'$ is a Steiner tree for $M$. But is it an optimal Steiner-tree?
What does $T'$ cost?

We use $E(T)$ and $E(T')$ respectively to denote the number of edges of $T$ and $T'$. Let $n_1$ be the number of edges of $T$ that were contracted, and $n_2$ be the number of edges of $T$ that were sub-divided to convert $T$ to $T'$. Then, since $T$ is an optimal Steiner tree,

$$E(T') = E(T) + n_2 - n_1 \geq E(T),$$

so $n_2 \geq n_1$. 
Creating Steiner tree $T''$

Now we will create a second tree, $T''$, from $T$ by using operations that are symmetric to those used to create $T'$:

First, change the value of site $j$ from 1 to 0, for every $i,j$—violating node $v$ in $T$. This changes the $i,j$ gamete at each such node from 01 to 00.

Second, if any $i,j$—violating node $v$ has a neighbor $w$ in $T$ whose $i,j$ gamete is 00, contract the $(v, w)$ edge.

Third, if any $i,j$—violating node $v$ has a neighbor $w$ in $T$ whose $i,j$ gamete is 11, insert a new node $w'$ in the $(v, w)$ edge, and give it a label that is identical to that of nodes $v$ and $w$ at every site other than $i$ and $j$. Then, set the $i,j$ sites at node $w'$ to 10.
What does $T''$ cost?

By the same reasoning we used for $T'$, the resulting graph $T''$ is a Steiner tree for $M$. To determine the number of edges, $E(T'')$ in $T''$, note that an edge $e$ in $T$ is contracted (or sub-divided) to create $T'$ if and only if edge $e$ is sub-divided (or contracted) to created $T''$. 

So, $n_2$ is the number of edges of $T$ that were contracted, and $n_1$ is the number of edges of $T$ that were sub-divided, to create $T''$. This is exactly symmetric to the counting we did for $T'$. Therefore,

$$E(T'') = E(T) + n_1 - n_2 \geq E(T),$$

so $n_1 \geq n_2$. 
But, but, but

But in the analysis of \( T' \), we concluded that \( n_2 \geq n_1 \), so now we conclude that \( n_1 = n_2 \), and \( E(T) = E(T') = E(T'') \). This means that \( T' \) and \( T'' \) are also optimal Steiner trees for \( M \).

All the \( i,j \) gametes in \( T' \) and \( T'' \) are now either 11, 10, or 00, which are all \( i,j \) gametes for some string in \( M \). So, \( i,j \) is no longer a violating pair. Further, no gametes other than \( i,j \) gametes were changed, so \( T' \) and \( T'' \) are both Steiner trees for \( M \) with one fewer violating pairs of sites than \( T \) has. That's Progress.
Repeating these modifications for any remaining violating pair of sites, we ultimately create an optimal Steiner tree for $M$ with no violating pair of sites, so every node in the tree satisfies the definition for a node to be in the Buneman graph for $M$.

And this ends the proof of the Buneman-Steiner Theorem.
QED
Right?

No! Fooled you – and myself.

The problem is that by changing every 01-gamete to 11 in site pair \( i, j \), we converted \( i, j \) from a violating pair to a non-violating pair, but we may have created an \( i, k \)-violating pair for some site \( k \neq j \).

Suppose, for example that node \( v \) in \( T - B \) has \( i, j \)-gamete 01, and that the bit a site \( k \) is 0. If there is a row in \( M \) with \( i, k \)-gamete 00, then after the changes at site \( i \), that \( i, k \)-gamete becomes 10, which might create an \( i, k \)-violating pair.

A similar problem arises when we change \( i, j \)-gamete 01 to 00 at an \( i, j \)-violating pair.
So, what progress?

Given the problem that removing an $i,j$-violation might introduce an $i,k$-violation, what is the measure of progress in this “proof”?

One idea: Maybe after removing the $i,j$-violation, even if additional violations are introduced, the $i,j$-violation (i.e., the 01-gamete in $i,j$) can never come back. That is progress we can use to finish the proof.

But, is it true that the 01-gamete can never come back? Not without some additional rules on how violating pairs of sites are chosen.
What other kind of progress might be possible?

The general proof in van de Vel uses a different kind of progress based on medians and medians graphs. It is more abstract, and it is a bit more involved to describe. So, we can’t cover it in this talk.

But their argument seems to have the same problem as mine. Did they overlook the issue, or is it so obvious that it isn’t a problem, that they didn’t even mention it?
Thanks

Thanks for listening and for the therapy. Send me a bill in the morning.