Explanation of HIV data table

The first seven columns of the "Data Table" *Excel* sheet present general information regarding each clinic visit. The data are organized by **Subject** and **Visit**. The numbering scheme for visits is complicated by the fact that some visits lack sequence data. Since the Markham *et al.* study focused on sequence analysis, they skipped over some of these no-sequence data points when numbering subjects' visits. For example, subject 12's first visit with sequence data (on 4/25/1989) was labeled V1, and that subject's next visit *with sequence data* (on 12/19/89) was labeled V2. To keep our notation consistent with theirs, therefore, we have labeled the intermediate no-sequence visit as V1.5.

Continuing with the visit summary columns, the two columns after **Visit** contain data on when each clinic visit occurred. Note that although the experimental design called for six-month intervals between visits, the actual schedule is significantly more variable. The **CD4s** column lists that patient's CD4 count at each visit. **# Sequences** lists the total number of viral sequences obtained at each visit, while **# Distinct Sequences** lists the number of distinct sequences (*clones*) within that sample. For example, if the viruses sampled at a particular visit included 7 copies of sequence A, 2 copies of sequence B, and 1 copy of sequence C, we would record that sample as containing 10 sequences but only 3 distinct sequences. Visits for which we don't have HIV sequence data are listed as having zero sequences.

The remaining columns contain statistics summarizing the sequence data from each visit. The first such column, ***S***, represents the number of segregating (polymorphic) sites within a sample, computed by counting the number of sites that are not fixed for a single nucleotide. This gives us one measure of the sample's genetic diversity, although not a good one since *S* depends on the sample size. Correcting this bias gives us the next column, ***θ***, which is calculated by the formula

$$\theta = S / \left(\sum_{i=1}^{n-1} \frac{1}{i} \right),$$

where *n* is the sample size. The next column, ***Π***, contains another diversity statistic: the average pairwise genetic distance. This value is computed by adding up all entries in a sample's pairwise distance matrix and then dividing by the total number of entries.

Here's a worked example demonstrating how to calculate these statistics:

```
Sequence 1:  A G T C G A T C G A T C G A T C G
Sequence 2:  A G C C G A C C G A T C G A T C G
Sequence 3:  A G A T G A T G A A T A G A C C A
Sequence 4:  G G T T G A T C A A T C G G C C A
Sequence 5:  A G T T G A T C A A T C G A C C A
             -----------------------------------
Segregates?  +   + +     + + +     +   + +   +
```

First, we count the number of segregating sites in our sample; this gives us S=10. We then construct the distance matrix:

	1	**2**	**3**	**4**	**5**
1	0	2	7	6	4
2		0	8	8	6
3			0	5	3
4				0	2

The sum of all entries is 51 and the total number of (off-diagonal) entries is 10; so Π = 51/10 = 5.1 . The sample size is 5, so the denominator of θ is 1 + 1/2 + 1/3 + 1/4 = 25/12, meaning that $\theta = 10 \times 12 / 25 = 4.8$.

Note that the two statistics are not equal. This is because, while both statistics measure the amount of polymorphism in the sample, Π is influenced by the frequency of those polymorphisms while θ is not. The final column of the combined data, ***D***, contains a new statistic *D* that describes the frequency distribution of polymorphisms:

$$D = \frac{\Pi - \theta}{\sqrt{\hat{V}(\Pi - \theta)}}.$$

Here, $\hat{V}$ represents the expected variance, which is calculated from *S* and *n* by a straightforward but cumbersome formula (see Tajima 1996 for details). The case $D = 0$ corresponds to the expected distribution of polymorphism frequencies if mutations occur at random. If $D < 0$, most polymorphisms occur at lower frequencies than expected, suggesting a recent decrease in population diversity (possibly due to a selective sweep or population bottleneck). $D > 0$, on the other hand, suggests that some force (such as diversifying selection or population subdivision) is acting to maintain genetic diversity.

One way to tease apart the effects of selection vs. population structure is to study synonymous and nonsynonymous mutations separately (Leigh Brown 1997, Hahn *et al.* 2002). Assuming synonymous substitutions to be selectively neutral, we can first use them to test for population equilibrium. We can then subtract these population effects from the pattern observed by studying nonsynonymous substitutions, thereby isolating the effects of selection:

	D_{SYN}	**D_{NON}-D_{SYN}**
>0	Population subdivision	Diversifying selection
<0	Population bottleneck	Selective sweep

Interpretation of results from a synonymous/nonsynonymous Tajima test.

The next three sections of the data table present the results of this partitioned analysis. The first section lists values of θ, Π, and *D* for synonymous substitutions; the second section calculates the same statistics for nonsynonymous substitutions. Finally, the discrepancy $D_{SYN} - D_{NON}$ is computed. <Significance tests?>

References:

Hahn et al. 2002
Leigh Brown 1997
Tajima, F. 1996. Statistical method for testing the neutral mutation hypothesis by DNA