Lecture 4 – Characters: Molecular

The application of molecular data to estimating phylogenies also has a long history. In fact, (again) as Joe describes in Chapter 10, analyzing molecular data (frequencies of blood-group alleles) was actually the motivation for Cavalli-Sforza & Edwards to begin development of computational methods for phylogeny estimation. (Figure 10.5 is the first phylogeography I know of & it’s from the early 1960’s).

I. Types of Molecular Data: Certainly the lion’s share of molecular data used in phylogeny estimation is DNA sequence data, but we should examine other types of molecular data, and some of the characteristics of those data that influence phylogeny estimation.

A. Inherently Distance-based Data

As I indicated last week, there are molecular phylogenetic approaches in which the nature of the data produced is a matrix of pairwise distances. The units for these distances vary, but the matrix can then be subjected to a number of potential phylogenetic analyses.

<table>
<thead>
<tr>
<th></th>
<th>cwk1056</th>
<th>eaa292</th>
<th>cwk1025</th>
<th>eaa448</th>
<th>dsr5032</th>
<th>eaa028</th>
<th>fac1117</th>
<th>cwk1007</th>
</tr>
</thead>
<tbody>
<tr>
<td>cwk1056</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
</tr>
<tr>
<td>eaa292</td>
<td>0.05840708</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
</tr>
<tr>
<td>cwk1025</td>
<td>0.01769911</td>
<td>0.05398230</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
</tr>
<tr>
<td>eaa448</td>
<td>0.08672567</td>
<td>0.08141593</td>
<td>0.08230089</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
</tr>
<tr>
<td>dsr5032</td>
<td>0.02566372</td>
<td>0.05929204</td>
<td>0.01946903</td>
<td>0.08495575</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
</tr>
<tr>
<td>eaa028</td>
<td>0.06725664</td>
<td>0.07433628</td>
<td>0.06371681</td>
<td>0.07522124</td>
<td>0.07168142</td>
<td>------</td>
<td>------</td>
<td>------</td>
</tr>
<tr>
<td>fac1117</td>
<td>0.02123894</td>
<td>0.05575221</td>
<td>0.00530973</td>
<td>0.08053097</td>
<td>0.02123894</td>
<td>0.0637168</td>
<td>------</td>
<td>------</td>
</tr>
<tr>
<td>cwk1007</td>
<td>0.05221239</td>
<td>0.02920354</td>
<td>0.05132743</td>
<td>0.08230089</td>
<td>0.05486726</td>
<td>0.07610620</td>
<td>0.05132743</td>
<td>------</td>
</tr>
<tr>
<td>eaa667</td>
<td>0.05840708</td>
<td>0.01238938</td>
<td>0.05221239</td>
<td>0.07787611</td>
<td>0.05752213</td>
<td>0.07433628</td>
<td>0.05398230</td>
<td>0.02743363</td>
</tr>
</tbody>
</table>

Again, historically inherently distance based methods have included things like DNA-DNA hybridization and immunological distances. We won’t focus on those because they’re never used anymore. However, as I mentioned, information on comparative genomics may be presented as inherently distance data.

An example of a simple genomic distance is provided by Edwards et al. (2002; Systematic Biology, 51:599).

This method is based on taking large amounts of sequence data that is assumed to be a random sample from each respective genome, and calculating the frequency of \( n \) bp words in each taxon. If \( n = 1 \) there are 4 words, G, A, T, C (so the data are the base frequencies). If \( n = 2 \) there are 16 possible dinucleotide words (4’).
Edwards et al. (2002) use 5 bp words, so there are $4^5 = 1024$ possible words, and the frequency of each word is calculated from the genome sample for each OTU.

The Euclidian distance between each pair of genomes is calculated to generate a distance matrix.

$$d_{ij} = \left\{ \sum_{x=1}^{1024} (f_{xi} - f_{xj})^2 \right\}^{1/2}$$

where $f_{xi}$ is the frequency of word $x$ in taxon $i$ and $f_{xj}$ is the frequency of word $x$ in taxon $j$.

This matrix is then subjected to any of a number of tree estimation methods (that we’ll cover later).

Deep split in bird phylogeny (Paleognathous birds) is reflected in the genomic signature.
It remains to be seen to what extent such approaches will find application in phylogenetics. There are a number of potential problems (calculating distances based on non-homologous characters), and these are discussed in the paper. Even though it’s over a decade old, it’s still a good read as an introduction to alignment-free genomic approaches.

B. Potential Character Data – Many molecular approaches are amenable to erecting a character-by-taxon matrix as you’ll do in lab tomorrow night.

1. **Allozymes** (isozymes) represent one of the first types of molecular character data. Allozymes are simply allelic forms of enzymes that differ in electronic charge (because of an amino-acid substitution), and have different electrophoretic mobility. {For the math/c.s. types, electrophoresis is a standard lab procedure in which molecules (e.g., proteins or DNA fragments) in a solution are subjected to an electromagnetic field, usually through some type of gel matrix. These molecules migrate through the matrix at a rate that is dependent on size and ionic charge. A reasonable tutorial on electrophoresis is available at: [http://www.resonancepub.com/electropho.htm](http://www.resonancepub.com/electropho.htm)}

Throughout the 70’s and 80’s, allozymes were the most widely used molecular data type, and a ton was written on how best to use the data this approach provided. It’s not used much for phylogeny anymore, but is still used some in population genetics.

A good review was provided by Murphy et al. (1996; pp. 51-120 in *Molecular Systematics*, 2nd Edition [Hillis et al., eds.] Sinauer).

2. Another data type, that’s seeing a bit of a resurgence, is **chromosomal data**. These data ranged from simple karyotypes (which included things like counts and gross morphology of chromosomes), to FISH (fluorescent in-situ hybridization), which can be used to identify the locality of genes of interest on the complement of chromosomes.

Usually some inference is made with respect to inversion events, and the events are used as characters. A great early example is Carson (1983. Genetics 103: 465), who applied inferred inversions of polytene chromosomes to the phylogeny of Hawaiian *Drosophila*. 
A good overview of these techniques was provided by Sessions (1996. pp121-167 in *Molecular Systematics*, 2nd Edition [Hillis et al., eds.] Sinauer).

A more recent example is from Kamail et al. (2012. PLoS Pathogens), in which the authors use sequences flanking break points, to produce a very well supported reconstruction of inversions among several species of *Anopheles*, which are the vectors of human malaria.

They then use these in a classical Hennigian analysis that indicates the major vectors are not sister species.
3. **Fragment Data** – DNA sequence variation can be assayed indirectly by using enzymes called restriction enzymes. These enzymes cleave DNA at particular sites, and recognize particular sequences.

For example, the enzyme EcoR1 will cleave DNA anywhere there is the following sequence occurs. This is called a restriction site.

```
... G - A - A - T - T - C ...
```

The application of restriction enzymes to phylogenies is usually conducted on either chloroplast DNA in plants or mitochondrial DNA in animals. This approach allows one to survey the entire organellar genome for variation, and several of enzymes are used.

AFLP’s are a related approach.

The **characters** that are used are **presence/absence of restriction sites** (or fragments). Because we have no way to tell what the actual substitutions were that may have generated a change (either a gain or a loss of a site), parsimony methods are the usual mode of analysis and weighting schemes are typically employed. However, there have been some attempts to model restriction site data.

4. The most common type of molecular data is **sequence data**. These can obviously be from DNA or from proteins.

   a. **Gene sequences** (DNA sequences) are the most common.

      This is quite straightforward; nucleotide positions are the characters, and the four nucleotides are the character states (there are only four possible states).

      We’ll spend the majority of the semester with this type of data.

   b. **Protein sequences** (amino-acid sequences) are also used, particularly to estimate old phylogenetic relationships.

      So, amino-acid positions are the characters and the 20 amino acids are the character states. There’s a much bigger character state space, and this is seen by some as an advantage.
However a disadvantage is that there may be more than one path of nucleotide substitutions from one amino acid to another, and this has led some to view this as an inferior character type.

Typically, nucleotide sequences are collected and amino-acid sequences are inferred. This is due to the fact that the DNA sequencing technology is far superior.

There have been controversies over which approach is better. Even if nucleotide data are generated, it’s easy enough to translate them into protein sequences and analyze these data.

5. Higher order molecular characters – Rare Genomic Changes.

There are several different types of higher order molecular characters. A good (but aging) review is provided by Rokas and Holland (2000. TREE, 15:454).
a. Insertions/Deletions in/of introns.

Most of the examples I’ve seen apply these characters to already existing phylogenetic hypotheses.

For example, Murphy et al. (2007. Genome Res., 17: 413) collected comparative genomic data on several mammals to try to test alternative placements at the root of placental mammals.

They identified lots of these sites to support a particular resolution in the deepest part of the tree, that is, that Afrotheria & Xenartha form a clade that is the sister-taxon to the rest of the placental mammals.

b. Transposable elements – Another type of higher order character that has received a fair amount of attention are transposable elements (e.g., SINES & LINES).

These are blocks of DNA that move around in the genome over evolutionary time.
Once an element inserts into the genome, it spawns daughter copies, which also insert into the genome.

In the example shown, cetaceans and Hippos share a unique SINE insertion suggesting Hippos are the sister-group of whales (an hypothesis derived by sequence data).

c. **microRNA Profiles** – miRNAs are small (~70 bp) non-coding RNA that form hairpins. These will bind to mRNA and inhibit translation; this is an important regulatory mechanism during development in metazoans (Reviewed by Tarver et al., 2013. MB&E, 30:2369).

They’re said to be lost only rarely, and not susceptible to convergence, so these are the current “next big thing.”

In fact, a science writer at *Nature* has gone all in with these (Dolgin 2012. *Nature*, 486:460)

However, critical examination of the utility of miRNA leads to a more measured conclusion (Thompson et al. 2014. doi 10.1073/pnas.1407207111; Penny. 2013. GB&E, 5:819).

Losses are more frequent than reported, there is large heterogeneity in rates of gains and losses, there’s ascertainment bias, and silly mistakes have been made in interpretation (i.e., with respect to rooting).
d. Gene order – lots of work has been done on inferring phylogenies from gene order.

These have often focused on organellar genomes (e.g., rearrangement of genes in mtDNA genome), and here again, shared rearrangements can be treated as Hennigian synapomorphies. (Example – Webster & Littlewood. 2012. Int. J. Parasit. 42:313-321.) Similar to the chromosomal inversion approaches we just addressed.

Increasingly, gene-order data have been applied to the growing database of prokaryotic genomes.

Here, the methods are distance based (and the term “genomic distance,” as opposed to genetic distance, is used). They’re based on HSPs (High Scoring Sequence Pairs) identified by pairwise BLAST searches.

They measure the number of base-pairs shared in a pair of genomes in these putative homologous genes.

These papers don’t tend to show up in the phylogenetics literature.

More and more genomicists are trying to get away from alignment-based phylogenies (e.g., Jun et al. 2010. PNAS. 107:133). This paper uses a word frequency approach on proteomes, alters the word length until the tree stabilizes. In many ways Scott Edwards was ahead of the curve, and the genomicists don’t cite the phylogenetic literature.

So among many molecular systematists, there’s a great deal of enthusiasm for these RGC approaches. The idea is that many of them are caused by changes in complex molecular genetic machinery (e.g., DNA repair mechanisms) and are unlikely to be subject to homoplasy (convergence and/or reversal). Therefore they should provide great classical Hennigian synapomorphies.

My opinion is less enthusiastic for two reasons. First, by their nature RGC’s are rare and we shouldn’t expect to find enough of them to provide resolution of very large phylogenies (and phylogenies are growing all the time). That is, for $n$ taxa, there are $n-3$ internal branches in an unrooted tree; for a modern analysis of moderate size (say 80 taxa), we would need 77 synapomorphies and have them distributed perfectly across the tree (i.e., one on each internal branch).

Second, because they’re rare, we don’t have enough information to model them and it’s therefore difficult to apply statistical phylogenetics.

II. Homology of Molecular Characters – We need to focus on issues of homology, just as we did for morphology. For molecular characters, there are two levels of homology that we have to deal with.

A. Alignment and homology of characters (e.g., nucleotide sites).

Because of its centrality to molecular phylogenetics, alignment will be addressed separately in Lecture 5. Alignment of sequences determines positional homology (or, more precisely, provides hypotheses of positional homologies).
B. At a higher level, we have to worry about homology of the source of the sequence data (i.e., homology of the genes).

Remember that homology is similarity due to descent from a common ancestor.

Paralogy is homology due to gene duplication.  
Orthology is homology due to speciation,  
Gene duplication has been an incredibly important and widespread evolutionary process.

It’s easy to demonstrate that if a mixture of orthologous and paralogous genes are compared, the true history of the genes (a.k.a. the gene tree) won’t correspond to the species phylogeny (a.k.a. the species tree) that we’re trying to estimate.

Because gene duplication is such a widespread phenomenon, there have been several methods devised to try to take advantage of them a character data. The community is just starting to model the process now, so statistical approaches should be forthcoming.