IB 168 - Lectures 18 & 19 (Molecular systematics) **Spring 2009**

WHAT IS MOLECULAR SYSTEMATICS?

Not a distinct field of systematics; instead, just the use of **macromolecular** (i.e., DNA/RNA) systematic approaches and data in the study of systematic (usually phylogenetic) problems.

WHY MOLECULAR DATA?

- Systematists have been long interested in looking at the <u>fundamental</u> <u>material</u> <u>of</u> <u>heredity</u> as the best way to understand organismal variation and relationships.
- Also, DNA characters are in some ways relatively <u>simple</u> characters for phylogenetic analysis (for example, nucleotide positions are highly <u>discrete</u> characters comprised of 4 alternative character states (ACGT)).
- DNA characters are also much more <u>numerous</u> than morphological characters.

DNA AND CHROMOSOMES:

- Chromosomal theory of heredity was proposed in 1903 (by Sutton); chromosomes have long been known to be associated with inheritance.
- The proteins in chromosomes were long thought to be the likely genetic material (20 amino acids versus only 4 nucleotides in DNA -- DNA was thought to be too simple to encode the complexity of life)
- 1953 discovery that DNA is the genetic material (Watson/Crick)

[Note - we now know that only one DNA double-stranded molecule is present in a single chromosome -- on that basis, chromosomal work could be considered the highest level of molecular (DNA) biology (but called cytogenetics instead]

- DNA STRUCTURE (brief overview relevant to molecular systematic methods):
    - Two strands of alternating phosphates and sugars (deoxyribose) form the backbone of DNA.
    - Nitrogenous bases of the two strands are paired along inside (like spiral staircase), **see handout**.
    - Four nucleotide types (ACGT): pairing is between pyrimidines (CT) and purines (AG).
    - A:T and G:C are each united by hydrogen bonds (weak electrostatic bonds, non-covalent, easily broken by heat).
    - Change (mutation) in one strand forces a compensating change in other (that is, if an A is mutated to a G, then the.T in the other strand at the same position is changed to a C by enzymatic correction).
    - Replication of DNA is semi-conservative -- the double-strand is enzymatically unzipped, with DNA polymerase involved in creating a new mate for each strand (pairing strands are antiparallel; that is, each is the reverse complement of the other)
    - <u>Semi-conservative replication is widely exploited in the methods of molecular systematics</u> (and by molecular biologists in general)

- DNA TYPES

Coding and non-coding DNA generally evolve at different rates and are useful at different levels of study (slowly evolving coding sequences are useful for examining relationships among distantly related plants; rapidly evolving non-coding sequences are useful for examining relationships among closely related plants).

1) Coding DNA in the strict sense codes for proteins (and therefore evolves relatively slowly compared to non-coding DNA -- natural selection maintains the coding sequence). The term "coding" can also be extended to DNA that encodes RNAs that are not translated into proteins but have important metabolic functions; for example, rRNA genes and tRNA genes, which also can be quite conserved in sequence.

The DNA "code": a 3-base stretch of DNA (= a codon) codes for one of 20 amino acids, the building blocks of proteins. 64 codons in total but only 20 amino acids total; DEGENERACY at 3rd position of codons (that is, the 3rd position can change to one of the other nucleotide states without affecting the amino acid that is encoded).

- 3rd positions in codons therefore evolve more quickly than 1st and 2nd positions; natural selection does not act as strongly on 3rd positions.

- Coding DNA in the strictest sense is transcribed and processed into mRNA, which is translated into protein.

- INTRONS (non-coding intervening sequences within genes) are removed during processing of mRNA (primary transcript averages 5X size of mature mRNA in higher eukaryotes -- lots of intron sequence relative to coding (exon) sequence -- rapidly evolving in sequence).

2) Spacer regions between genes and other non-coding sequences are among fastest evolving sequences.

- MUTATION TYPES -- likelihood of fixation is dependent on severity of effect on fitness

1) Substitutions of one nucleotide type for another

2) Structural mutations: insertions/deletions and rearrangements

- PLANT GENOMES

1) Plant nuclear genome can be huge: lots of polyploidy and repetitive DNA in general

- 100,000,000 bp to 100,000,000,000 bp in plants

- NOT ALL GENES ARE EQUALLY USEFUL FOR A GIVEN PHYLOGENETIC QUESTION

- Best to choose a gene region evolving at an optimal rate (fast-evolving gene for young lineage; slowly-evolving gene for ancient lineage)

- DESIRABLE PROPERTIES for a gene region to be used in phylogenetic analyses (in addition to correct rate of change to provide evidence of branching events):

(1) Easily sampled, even from herbarium specimens, (2) Mutations that arise are either fixed or lost at a rate faster than the rate of lineage branching (otherwise get allelic sorting or lineage sorting and can be misled about relationships)

<u>Nuclear ribosomal DNA</u> shows properties "1" and "2" and also includes regions with different rates of evolution and are therefore widely used in plant phylogenetic studies (**see handout**).

<u>Chloroplast DNA</u> is also widely used because of properties "1" and “2” and because of strictly uniparental (clonal) inheritance (no complication from sexual recombination, even if hybridization takes place). Although evolutionary rates are slow overall in chloroplast DNA, considerable variation does exist among genes and between genes and intron and spacer regions (**see handout**). <u>Chloroplast DNA was especially popular for restriction-site studies</u> because of its highly conserved structure (gene order) across land plants (see handout), which allowed using probes from distantly related species to

examine variation within specific regions of the genome (using Southern blot hybridization, see handout).
Mitochondrial DNA evolves too slowly in sequence and too rapidly in structure in most plants to be of much value in phylogenetic studies (unlike in animals, where parts of mtDNA evolve rapidly and are used extensively).

- METHODS OF DNA ANALYSIS

Following discovery that DNA coded for proteins, approximately 20 years elapsed before DNA systematic studies began because of technical limitations.

DNA/DNA HYBRIDIZATION (beginning in 1970s) -- see handout.

First approach -- did not require use of electrophoresis or enzymes in the lab and involved comparing the entire single-copy or low-copy (mostly coding) gene regions in different species (championed by Sibley/Ahlquist)

The method is based first-and-foremost on **exploiting the thermal kinetics of double-stranded DNA**: the process involves annealing single-stranded DNA of different organisms into a "heteroduplex" and then heating the double-stranded heteroduplexes until they are denatured into single-strands. By this process, a "melting curve" is obtained; <u>**the higher the similarity between the strands, the higher the denaturation temperature**</u>.

Strength of method:

Whole genome approach (a single gene region can be misleading about evolutionary history; DNA-DNA hybridization maximizes the number of genes examined.

Weakness of method:

Crude comparison of genomes and data difficult to analyze.

Huge amounts of DNA required.

All pairwise comparisons between taxa included in the study must be made to complete the similarity matrix and allow data analysis (logistically difficult to complete the matrix, especially if DNA is limiting). In other words, every time a species is added to the dataset, you must conduct DNA-DNA hybridization between it and every other species in the analysis.

Each comparison between taxa yields a single number -- overall similarity -- rather than a set of characters and character-states. With a similarity value, **the potential for data analysis is limited to distance methods**, such as phenetics.

RESTRICTION SITE ANALYSIS (1980s) -- see handout

- Cloning and commercial availability of restriction enzymes made this method possible (as did refinement of gel electrophoresis = separation of molecules by charge or size in an electric field, through a gel matrix)

- **Restriction enzymes (type 2) cut double-stranded DNA at or near a specific 4, 5, or 6 base-pair recognition sequence** (the enzymes were isolated from bacteria, which use them to cleave and thereby destroy invading viral DNA)

- Restriction enzymes allow an <u>**indirect**</u> assessment of nucleotide differences by <u>**presence or absence of restriction sites**</u>

- <u>Presence or absence of restriction site can be inferred across species by examining small regions of the genome (using probes for specific regions) and looking for differences where Taxon 1 has a large band that is absent from Taxon 2, which instead has two bands that are absent in the Taxon 1 and that add up (in molecular weight) to the same size as the large band in Taxon 1 (**see handout**)</u>.

Strength of the method:

- Like the DNA-DNA hybridization method, restriction site analysis allows for **wide sampling throughout the genome** (**usually the chloroplast genome in plants**).
- Like DNA sequencing (and unlike DNA-DNA hybridization), restriction site analysis yields <u>character-data</u> (restriction site is character; presence or absence of site are the character states) and allows resolution <u>close</u> to the DNA sequence level.

Weaknesses of the method:

- An indirect look at DNA sequence (many ways to lose a restriction site -- 18 mutations can knock out a 6 bp recognition sequence; therefore, high potential for homoplasy or noise in data).
- Only 2 character-states per character (DNA sequence has 4 states); other reason for high potential for homoplasy (noise).
- Difficult to impossible to extend the method across distantly related groups of plants.
- Painful data acquisition and analysis to do it right.

PCR (polymerase chain reaction) -- see handout

- Revolutionized molecular systematics by allowing gene regions with known flanking sequence to be amplified and thereby be effectively isolated from other regions of the genome.

- Involves use of short single-stranded DNA "primers" that bind to conserved DNA sequence on either side (and on each strand) of the DNA region of interest that one wishes to study. A heat-stable DNA polymerase (such as *Taq* polymerase) is then used to duplicate the region of interest over-and-over again, through successive rounds of replication, and ultimately yielding large amounts of the target sequence. Before PCR was invented, isolating a DNA region from a given species required laborious molecular cloning methods, which were the main obstacle to DNA sequencing.

- Like DNA/DNA hybridization and Southern blot hybridization (often used in restriction-site analysis), PCR takes advantage of the denaturation kinetics of DNA; each round of PCR involves denaturation of double-stranded target DNA, annealing of the primers, and then extension of the primers by DNA polymerase (as in natural, semi-conservative replication of DNA).

DNA SEQUENCING (see handout)

DNA sequencing also involves use of short single-stranded primers (same ones as used in PCR of the region to be sequenced oftentimes) and primer extension using DNA polymerase. BUT it also involves incorporation -- at random -- of "stop" nucleotides that end strand extension once the "stop" nucleotide (ddNTP) is incorporated. In the end, the sequencing reactions result in a set of DNA fragments that were terminated at every possible position along the DNA sequence of interest. When electrophoresed on a polyacrylamide gel, the DNA fragments sort by size and the sequence can be read from the bottom to the top of the gel, one base at a time.

Strengths of the method:

- Highest level of resolution of DNA data.
- Relatively easily analyzed character data that can be incorporated into a wide diversity of data-sets and subjected to wide diversity of analyses.
- Four nucleotide states better than restriction site situation (with only 2 states)
- Can choose a DNA region evolving at the right rate for a particular phylogenetic problem.

Weakness of the method: Generally looking at only a tiny region of the genome; therefore, high potential to be misled by unusual evolutionary history of a particular gene (enhanced capacity for rapid sequencing may overcome this weakness soon, though).