Chapter 12

Molecular Systematics and Evolution
Objectives

Overview of potential uses of molecular tools to answer systematics/taxonomic questions

Brief description of most common tools

Methods, concepts used to classify arthropods undergoing change

Started with formal hierarchical binomial system by Linnaeus in 1758
Systematics

*Systema Naturae*: hierarchical ranking of

- species
- genus
- family
- order
- class
- superclass
- subphylum
- phylum
- kingdom
A Higher Classification of Phylum Arthropoda

Phylum Arthropoda

Subphylum Trilobita (extinct)
Subphylum Chelicerata

Class Merostomata (horseshoe crabs, eurypterids)
Class Pycnogonida (sea spiders)
Class Arachnida (spiders, mites, ticks, scorpions, phalangids)
A Higher Classification of Phylum Arthropoda

Subphylum Crustacea (crabs, shrimp, lobsters)
Subphylum Myriapoda
  Class Diplopoda (millipedes)
  Class Chilopoda (centipedes)
  Class Pauropoda (pauropods)
  Class Symphyla (garden centipedes)
A Higher Classification of the Phylum Arthropoda

Subphylum Hexapoda
Order Protura
Order Collembola
Order Diplura
Class Insecta (insects)
A Classification of the Class Insecta

Apterygota
Paleoptera
Polyneoptera
Paraneoptera
Holometabola
Linnaean System

No provision for evolutionary relationships

Assumed living world limited to ~10,000 species

Darwin’s *The Origin of Species*, 1859, first to assume classification should be based on phylogeny (phylogenetics)
Systematics and Taxonomy

**Systematics**: scientific study of kinds and diversity of organisms and any and all relationships among them

OR

science of the diversity of organisms

**Taxonomy**: theory and practice of classifying organisms
Systematics and Taxonomy

**Taxonomy:** both descriptive and identification

**Systematics:** deals with populations, species and higher taxa, concerned with variation within taxa
Species Diversity

Estimated 1.4 million species identified

Could be only 10% of total species on planet?

A surprise: new ORDER of insects:
Mantophasmatodea discovered in 2002?

All orders thought to have been found
In the 1980s, **cladistics** promoted: based on evolutionary histories of organisms.

In the 1990s and 2000s, some argued that the Linnaean binomial system is obsolete. **PhyloCode** proposed to replace the Linnaean system to make species names stable.

Genus names might be lost; species names may be shortened, hyphenated with the former genus or given numeric ID.
Molecular Evolution

Involves analyzing evolution of DNA and proteins, mechanisms causing such changes

Deciphering evolutionary history of genes and organisms

Comparative genomics compares overall structure and function of genomes

Complete genomes now available (i5K project)
Controversies

In molecular systematics, evolution:

- Molecular vs. Morphological data
- Molecular clock reliability
- Neutral, nearly neutral theory of evolution
- Homology and similarity
- Most appropriate genes to analyze
- Most appropriate analysis methods

Part of the problem is that ALL ONLY ESTIMATE WHAT REALLY HAPPENED
Molecular vs. Morphology

Which is better for phylogenies?

Morphology and molecular changes may be independent, responding to different evolutionary forces.

Question: will characters chosen exhibit variation appropriate to question?

Do characters have a genetic basis?

Are data collected and analyzed appropriately?
Molecular vs. Morphology

Which is better for constructing phylogenies?

Morphology and molecular data each have strengths and weaknesses

Sequence data: clear genetic basis, amount of data limited by genome size, funding, time (except fossils with degraded DNA)

Morphology: obtained from fossils; preserved collections useful; often less expensive
Molecular vs. Morphology

Should not be either/or issue

Combining both types of data likely to provide better results than just one approach

Even so, all ESTIMATE the evolutionary history of the taxa
Molecular Clock

Until 1960s, fossils provided only way to estimate **TIME** when ancestors of extant organisms lived

Molecular studies on proteins led to the molecular clock HYPOTHESIS

Useful for species with poor fossil record

Proposed by Zuckerkandl & Pauling, who examined aa substitutions in hemoglobin and cytochrome C in different vertebrates
Molecular Clock

Based on assumption that basic processes such as DNA replication, transcription, protein synthesis, metabolism are ~ among all living organisms and these genes are highly conserved

Over time, mutations occurred and DNA and protein sequences changed, but changes preserved FUNCTION of gene
Molecular Clock

Changes in 3rd codon – rarely cause change in aa (due to the ‘degenerate’ code)

Mutations in housekeeping genes occur at CONSTANT rate → reliable method for estimating divergence time
Molecular Clock

Unfortunately, subsequent protein sequence analyses suggest the clock may **tick at different rates in different lineages**

Rates of substitution vary

Calibration of clock needed with independent data, preferably fossils
Molecular Clock

Molecular clock used to estimate when *Buchnera* endosymbionts colonized aphid hosts:

- 16S ribosomal sequences compared
- Clock nearly constant

Results indicate symbionts in different aphids are distinct, concordant with host phylogeny, due to vertical transfer.
Molecular Clock

Molecular clock used to estimate when *Buchnera* endosymbionts colonized aphid hosts:

Bacterial and aphid radiations occurred at rate of 0.01 - 0.02 substitutions per site per 50 million years

THUS: Association began ca. 160 - 280 mya
Neutral / Nearly Neutral Theory

Mechanism of molecular evolution controversy

Many mutations produce alleles which are equivalent, or nearly so, to each other (neutral/nearly neutral) (WHY?)

These alleles not subject to selection because they do not affect fitness
Neutral / Nearly Neutral Theory

For any gene, a large % of all alleles are deleterious and are eliminated or kept at very low frequency by natural election.

Evolution of morphological, behavioral, ecological traits is governed by natural selection (e.g., neutralists don’t dispute evolution occurs).
Neutral / Nearly Neutral Theory

Neutral alleles do NOT affect morphology, physiology or behavior

MAJORITY of nt substitutions during evolution are due to gradual, random fixation of neutral changes rather than positive Darwinian selection
Neutral / Nearly Neutral Theory

Neutral mutations can spread due to random genetic drift

“Strictly neutral theory has not held up as well as nearly neutral...but useful null $H_0$ for detecting selection”

Why be concerned?

Neutrality is basic assumption of some methods of estimating phylogeny

Affects clock hypothesis
Neutral / Nearly Neutral Theory

Many mutations are under selection and

Much variability is nearly neutral

Debate is over how many and which mutations are neutral or nearly neutral

Each marker should be tested for neutrality
Neutrality less important if many different loci are studied
Homology and Similarity

A terminology issue

**Homology** is an important concept

Had precise meaning: **having a common evolutionary origin**

**Similarity**: protein and DNA sequences may be similar without being homologous

If in doubt, call it % similarity
Molecular Methods

Prior to 1960s, morphology and behavior used

In 1960s, proteins were found to be polymorphic and gel electrophoresis yielded evidence of similar forms (isozymes)

Isozymes are easy, inexpensive, but limited by amount of protein in small insects

Isozymes useful for studying mating systems, heterozygosity, relatedness, geographic variation, hybridization, species boundaries, phylogenetic analyses (< 50 million years)
Molecular Methods

In 1960s, immunological analyses also tried

Less often used today

Cytogenetic analyses: chromosome # and structure

Hybridization, species boundaries studies

DNA - DNA hybridization: single - copy DNA

Species and higher taxa up to family and order
Molecular Methods

Modern methods solve different problems

- RFLP
- RAPD - PCR
- Single-locus microsatellites
- Multilocus fingerprints
- DNA / RNA sequencing

Methods may be inappropriate, marginally appropriate, not cost-effective, appropriate, and effective for specific questions
<table>
<thead>
<tr>
<th>RFLP Analyses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene evolution</td>
</tr>
<tr>
<td>Pop. structure</td>
</tr>
<tr>
<td>Mating systems</td>
</tr>
<tr>
<td>Clonal detection</td>
</tr>
<tr>
<td>Heterozygosity</td>
</tr>
<tr>
<td>Paternity</td>
</tr>
<tr>
<td>Relatedness</td>
</tr>
<tr>
<td>Geog. variation</td>
</tr>
</tbody>
</table>
## RFLP Analyses

<table>
<thead>
<tr>
<th>Hybridization</th>
<th>Appropriate &amp; effective</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spp. boundaries</td>
<td>Appropriate &amp; effective</td>
</tr>
<tr>
<td><strong>Phylogeny</strong></td>
<td></td>
</tr>
<tr>
<td>0 - 5 mya</td>
<td>Appropriate &amp; effective</td>
</tr>
<tr>
<td>5 - 50</td>
<td>Appropriate &amp; effective</td>
</tr>
<tr>
<td>50 - 500</td>
<td>Marginal</td>
</tr>
<tr>
<td>500 - 3500</td>
<td>Inappropriate</td>
</tr>
</tbody>
</table>
RFLP Analyses

More than 1400 restriction enzymes available

Recognition sequences typically 4 to 6 bp

Changes in # and size of fragments occur if sequences change due to:

- Inversions
- Deletions
- Duplications
- Base substitutions
- Inverted duplications
- Additions
RFLP Analyses

1) Extract DNA: how clean does it need to be? How much do you need?

2) Digest DNA

3) Run fragments on gel with size standard, visualize with EtBr staining or Southern blots

EtBr requires at least 2 ng of DNA per band
Radiolabeled probes for Southern blots require less (1 - 5 pg)
<table>
<thead>
<tr>
<th>DNA / RNA Sequencing</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene evolution</td>
<td>Appropriate &amp; effective</td>
</tr>
<tr>
<td>Pop. structure</td>
<td>Appropriate &amp; effective</td>
</tr>
<tr>
<td>Mating systems</td>
<td>$ (not cost effective)</td>
</tr>
<tr>
<td>Clonal detection</td>
<td>$</td>
</tr>
<tr>
<td>Heterozygosity</td>
<td>Marginal</td>
</tr>
<tr>
<td>Paternity</td>
<td>$</td>
</tr>
<tr>
<td>Relatedness</td>
<td>$</td>
</tr>
<tr>
<td>Geog. variation</td>
<td>Appropriate &amp; effective</td>
</tr>
</tbody>
</table>
# DNA / RNA Sequencing

<table>
<thead>
<tr>
<th>Category</th>
<th>Time Range</th>
<th>Appropriate &amp; effective</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hybridization</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spp. Boundaries</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Phylogeny</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0-5 mya</td>
<td></td>
<td>Appropriate &amp; effective</td>
</tr>
<tr>
<td>5-50</td>
<td></td>
<td>Appropriate &amp; effective</td>
</tr>
<tr>
<td>50-500</td>
<td></td>
<td>Appropriate &amp; effective</td>
</tr>
<tr>
<td>500 -3500</td>
<td></td>
<td>Appropriate &amp; effective</td>
</tr>
</tbody>
</table>
Sequencing

DNA sequencing first done in 1975 (Sanger or Maxam & Gilbert)

Use of PCR makes sequencing less time consuming and expensive

Core facilities save time and resources

Used to construct molecular phylogenies of genes or gene families
Evaluate evolutionary changes within species
Construct phylogenies of different species
Sequencing

DNA sequencing remains relatively expensive, time consuming (typically uses Sanger sequencing but some NextGen sequencing is also being done)

**Targets:** single-copy genes, mt DNA, ribosomal DNA, introns

Limited use? if large numbers of individuals must be analyzed due to cost
Fragment Analyses

RAPD-PCR, single-locus microsatellites, and multilocus DNA fingerprinting can be used in some systematic problems.

**RAPD - PCR** useful preliminary screen for **cryptic species**: rarely do different species have same banding patterns.
Targets

Mitochondria: multiple copies per cell

- Microbial symbiont (related to Rickettsia, Anaplasma, Ehrlichia?)
- Low gene content: loss or transfer of genes occurred early in evolution
- Many mt genes moved horizontally to nuclear genome (mosaic genome)
Mitochondria

Respiratory power plant of cell → ATP

Mutation rates variable: mammals with rate 50-fold greater than plants

7 to 10 independent losses of each gene over evolutionary time

No evidence once genes transferred to nucleus they were regained by mt

No evidence for widespread lateral transfer into or between mt
Mitochondria

Maternally transmitted (cytoplasmic)

Some regions **diverge rapidly**, some **conserved**

Different regions useful for different systematics questions

How do you know which ones are appropriate for your project?
Mitochondria

Animal mt are small (16 to 20 kb), circular, lack introns, have compact gene arrangement

Usually contain 37 genes coding for small, large subunit rRNAs, 13 proteins, and 22 tRNAs
Mitochondria

Recombination does not occur in mt: fixation of advantageous mutations by selection will fix all other polymorphisms by ‘genetic hitchhiking’

Even rapidly evolving noncoding origin of replication (OR) region may not have neutral allele frequency due to linkage
Mitochondria

Used as molecular clocks to time divergences: may be problematic when non-neutral evolution is combined with altered rates of evolution in sister taxon: neutrality $H_0$ should be evaluated in phylogenetic analyses

Molecular clock appears to tick at different rates in different lineages and at different times within a lineage
Mitochondria

Problems:

- lack of recombination → single heritable unit
- recombination may rarely occur
- heteroplasmy (2 types of mt in single individual) may sometimes occur
- biparental inheritance can rarely occur
Mitochondrion of \textit{A. mellifera}

Many other mt genomes are available
Barcode of Life Project

A method of identifying organisms based on a short, standardized fragment of DNA (648-bp region of the COI gene of mitochondrion)

Used by taxonomists, ecologists, etc.

Technique proposed in 2003 with the promise

“…it would allow identification of all animal species”
Barcode of Life Project

Enthusiastically adopted as having the potential to “accelerate our discovery of new species, improve the quality of taxonomic information and make this information readily available to nontaxonomists and researchers.”

Many entomologists have used DNA barcoding (more than any other taxonomic group)

BarCode of Life projects for Formicidae, Trichoptera and Lepidoptera
Barcode of Life Project

However, COI sequences are not informative in algae, protists and plants (not universal)

Use of Barcoding to ID new species is controversial if it is the ONLY data provided

DNA barcoding should only create the $H_0$ that there is a new species; additional data should be obtained to confirm
Barcode of Life Project

DNA barcoding requires a library of data, voucher specimens authenticated by qualified taxonomists

**Major issue:** nuclear mitochondrial pseudogenes (also known as numts)

**Numts** are found in many species and may confuse the taxonomists, leading to inaccuracies such as belief that there is a ‘new’ cryptic species
Barcode of Life Project

If both mitochondrial and nuclear COI sequences are sequenced and there is a 3% difference, the erroneous conclusion may be reached that there is more than one ‘species’

This makes it challenge to the use of barcoding as an accurate method for species identification
Ribosomal RNA

Ribosomes: many per cell, involved in translating mRNA into proteins

Ribosomes = rRNA plus proteins, can be disassociated into 2 subunits, each containing rRNA and proteins

Ribosomal DNA used to evaluate evolution/systematics because universal rRNAs have variable and conserved regions
Ribosomal RNA

In eukaryotes, genes encoding 18S and 28S rRNA are clustered as tandem repeats in nucleolus-organizing regions of nuclear chromosomes.

Plus 2 ribosomal genes in mitochondria.

Most animals have 100 to 500 copies of rDNA in nuclear genome in tandem repeated transcription units.
Ribosomal RNA

As few as 45 copies in *Sciara coprophila* to > 3000 in *Locusta migratoria*

30 spp. of mosquitoes: 39 to 1023 copies
Ribosomal RNA

Repeated transcription unit consists of:

- leader promoter region = External Transcribed Spacer (ETS)
- 18s rDNA
- Internal noncoding Transcribed Spacer region (ITS)
- 28S rDNA
- InterGenic nontranscribed Spacer segment (IGS)
Ribosomal genes or fragments: Often are used in systematics studies

Transcription unit: repeated up to thousands of times
Ribosomal RNA

Ribosomal RNA genes undergo concerted evolution: sequence similarity of family more similar within a species than between species:

Achieved by unequal crossing over
Gene conversion
Illegitimate recombination

Ribosomal gene families considered “quite uniform”
Ribosomal RNA

28S rRNA genes often contain 2 retrotransposons, R1 and R2, in most insects.

Association > 500 million years?

Usually precisely located, many mutated

Block production of functional rRNA

Elements kept from inactivating all rRNA genes by unknown mechanisms
Satellite RNA

May comprise a large fraction of total DNA in an insect

Microsatellites usually species specific

Possibly because it mutates at very high rate

Can be used for species diagnosis or population studies
Introns

Introns within single-copy genes may be used in systematics

Are often highly variable

May be used for population studies
Nuclear Protein - Coding Genes

A variety of loci have been used

- alpha amylase
- acetylcholine esterase
- actin
- alcohol dehydrogenase
- arylphorin
- cecropin
- chorion genes
- dopa decarboxylase
Nuclear Protein - Coding Genes

A variety of loci have been used

- guanylate cyclase
- globin family genes
- histones 1 and 4
- hunchback
- kruppel
- luciferase
- lysozyme
- prune

- nullo
- opsin
- period
- phosphoglucone isomerase
- snail
- timeless
- triosephosphate isomerase
Nuclear Protein - Coding Genes

A variety of loci have been used:

- white
- wingless
- xanthine dehydrogenase
- yolk protein 1 and 2
- zeste

Now that multiple arthropod genomes have been sequenced completely, more genes are available.
Nuclear Protein - Coding Genes

Problems: loci may be heterozygous, are present in low copy number

Loci may be heterozygous
Are present in low copy number → difficult to amplify by the PCR
Genes contain large introns making it difficult to amplify more than one exon unless RT-PCR is used on mRNA
Many single-copy loci are ACTUALLY present in > one copy
MicroRNAs

Most protein-coding genes are highly conserved throughout evolution.

However, microRNAs have been added over evolutionary time as more complex organisms have evolved.

MicroRNAs code for regulatory RNAs that are 22 nt in length that affect translation of target mRNAs.

H₀: the addition of microRNAs can be used as characters to evaluate deep evolution.
MicroRNAs

Once gained, microRNAs usually remain functional, so the “signal stays intact for hundreds of millions of years”

Do microRNAs hold the secret to morphological complexity?

778 microRNA families have arisen during the 600 million or so years of animal evolution and only 48 have been lost
Rare Genomic Changes

Changes in genome structure may be useful in systematics / evolution

INDELS = INsertions or DELetions
Retroposon integration
Signature sequences
Mt gene order changes
Gene duplications
Codon changes
Phylogenetic Analysis of DNA Sequence Data

Which genes informative for which questions?
- Species, genera, families, orders, kingdoms

May not know in advance which gene/sequences are informative
- Literature reviews sometimes helpful, but not always
Phylogenetic Analysis of Sequence Data

Gene Trees or Species Trees: Analysis of particular locus may NOT agree with species phylogeny

Due to horizontal transfer of genes
Duplication and extinction of one of gene pair
Lineage sorting
Mt genes may be more reliable for some recent divergences
Phylogenetic Analysis of Sequence Data

Rooted or Unrooted Trees?

Most phylogenetic methods produce UNROOTED trees.

Information on evolutionary rate or most ancient relationship needed to ROOT inferred trees.
For any 4 taxa, there are 3 different unrooted trees; each can be rooted on any of its 5 branches.

Unrooted trees of 4 taxa

Two possible rooted trees from II above

IIA

IIB
Analysis of Sequence Data

Speciation methods make analysis complex

**Phyletic speciation:** gradual change through thousands of years

**Cladogenic speciation:** two populations of a species become isolated and diverge as a result of mutation, natural selection, genetic drift

Speciation through hybridization & polyploidy

Speciation through regulatory gene modification
Analysis of Sequence Data

Characters: attributes used to establish relation to other organism

Based on morphology, physiology, ecology, behavior, biochemistry, genetics

How to use characters is controversial

Cladistic (phylogenetic) systematics: uses only cladistic relationships as basis for classification

Phenetic systematics: focuses on overall similarities among organisms
Analysis of Sequence Data

Phenetic Systematics: (phenotype)

Involves all possible characters

Calculates average similarities with all characters assumed equally useful

Sometimes reflects phylogeny of taxa because those that are most ~ may share most recent ancestor, but not always so because of convergence
Analysis of Sequence Data

Cladistic Systematics: (phylogeny)

Rate of change not considered

Only monophyletic taxa used

Focuses on order of origin of lineages

Deals with amount and nature of change which occurs after cladogenesis

Characters assumed equal, are weighted
Analysis of Sequence Data

Cladistic Systematics: (phylogeny)

Often difficult to determine which character is primitive (ancestral, or PLESIOMORPHIC) and which is derived (APOMORPHIC).
Analysis of Sequence Data

Classification often presented in graphical form as treelike dichotomous branching graphs = DENDOGRAMS

A dendogram produced from cladistic information is a CLADOGRAM

Cladograms show sequence of origin of clades and indicates times at which various cladogenic events occurred

PHYLOGRAM (PHYLOGENETIC TREE) = dendrogram containing both phenetic and phylogenetic data
Analysis of Sequence Data

Project Goals and Appropriate Data

Evaluate published data to identify POTENTIALLY informative genes/sequences for your project.

Different goals (orders vs. genera) require different degrees of variability.

Conserved vs. variable regions in rRNA.

DNA extraction methods important, especially with dried museum specimens.
Analysis of Sequence Data

Project Goals and Appropriate Data

Sequences obtained from PCR products or by cloning and sequencing

Sequencing should be done of both strands to detect sequencing errors

Analysis can yield information as to structure, function of protein
Steps in Phylogenetic Analyses

1. Define systematics problem
2. Evaluate potential target genes or sequences based on literature review or preliminary laboratory analysis
3. Develop primers for PCR or clone; purify DNA for sequencing
4. Sequence DNA; Sequencing may be conducted in a core/commercial facility using automated equipment
5. Obtain related sequences by a BLAST search of databases, such as GenBank
6. Create multiple alignment of sequences; analyze by ClustalW (or latest version); refine alignment
7. Conduct phylogenetic analysis: use PAUP to create a tree, root the tree, analyze reliability of tree by bootstrapping. Tree methods include: Distance and character-based methods, Parsimony, Maximum likelihood, and Bayesian analysis
Form in which sequences are usually obtained with Sanger sequencing
Form of data after analysis to show open reading frame (ORF) of coding strand (●), codons, region at end that is vector sequence.
Analysis of Sequence Data

BLAST Comparisons of Sequence

Sequences in GenBank, EMBL, DNA Data Bank of Japan should be analyzed with YOUR sequences: growing exponentially

Most journals require depositing YOUR sequences to publish

BLAST = Basic Local Alignment Search Tool used to search large databases of DNA sequence data (or proteins)
Analysis of Sequence Data

BLAST Comparisons of Sequence

Goal is to find sequences that have regions of similarity to your sequence (query sequence)

Results of search orders sequences, provides e-values, which are the probability the hits would occur by chance if no true matches are present in the database (e^-5 often used as the cut off)
Analysis of Sequence Data

Aligning Sequences

Sequences aligned either with other sequences obtained in project or from GenBank

Aligning requires computer analyses using one of 3 major methods

Each method has specific assumptions

Usually based on PARSIMONY: minimal number of changes to transform one sequence to another
Analysis of Sequence Data

Aligning Sequences

Computer program called CLUSTALW is commonly used to align DNA sequences.

Inserts gaps into one or other of sequences to maximize number of residues that match.

Gaps assumed due to insertions or deletions over evolutionary time.
Analysis of Sequence Data

Constructing Phylogenies: 4 main methods

- Parsimony
- Distance
- Likelihood
- Bayesian

Discussing these in detail is beyond scope of this course

Ideally, you use more than one method to ESTIMATE relationships
Artifacts of Analysis

Inaccuracies in trees occur for a variety of reasons

Alignments may be poor

Are sequences orthologous, paralogous or xenologous?

Ortholog = homologous sequence produced by speciation derived from a common ancestor, often with ~ functions

Paralogs = homologous sequences by gene duplication that diverged subsequently
Artifacts of Analysis

Inaccuracies in trees occur for a variety of reasons

Xenologs = horizontal transfer of a gene occurred between 2 organisms

Sometimes divergent lineages are morphologically similar (homoplasy) due to a reversal to an ancestral trait in a lineage or to convergence or parallelism
Artifacts of Analysis

Inaccuracies in trees occur for a variety of reasons

**Long-branch attraction** results in poor trees when long branches are in close proximity to short branches on trees.

Long-branch attraction can be eliminated by excluding faster evolving third-codon positions, and sampling more taxa to break up long branches, and to sampling more characters.
Artifacts of Analysis

Inaccuracies due to

Improper alignments

Species chosen to represent each group: different species sometimes result in different trees

Increasing number of species can improve accuracy but there is a trade off in difficulty of analysis
Software Packages

Variety are available and change rapidly

- MrBayes
- PHYLIP
- PAUP
- Hennig86
- MacClade

Geneious a use-friendly addition
Universal Tree of Life

Traditional view: ‘animals’ and ‘plants’

Bacteria and fungi often studied in botany departments

A later view: eukaryotes and prokaryotes

Fungi are not plants

Single-celled eukaryotes originally put into phylum Protista, but this is heterogeneous group (algae, protozoa, water molds, etc.)
Universal Tree of Life

Three Domains (Woese et al. 1990) based on ribosomal RNA analyses

Archaeabacteria
Eubacteria
Eukaryota

This increasingly accepted
Origin of Eukaryota about 2 billion years ago
Origin of Eukaryotes

Eukaryotic cells acquired mitochondria from bacteria

Eukaryotes are chimeric, derived from both archaeabacterial and eubacterial lineages

Chimeric genomes occurred through lateral transmission of genes

Gene swapping: “you are what you eat”

Cells engulfing bacteria → genes transferred into nuclear genome (Doolittle 1998)
Origin of Eukaryotes

A consequence of horizontal gene transfer

Analyses of different genes can result in conflicting phylogenies

Some eukaryotic genes more ~ to archaeal

Some eukaryotic genes more ~ to eubacterial

Analyses challenge traditional view that vertical transmission is predominant force in evolution
Molecular phylogenies support ‘web of life’ concept: gene exchange and horizontal gene transfer occurred.

Evolution is not linear.
Origin of Eukaryotes

More data needed to understand

Genome duplication probably important component of evolution

Source of new gene functions

Some duplicate genes lost (pseudogenes)

Many protein-coding genes belong to multigene families: due to gene duplication?
Fossil Record of Arthropods

Extensive

1263 families of fossil insects known

A few (Collembola) known from lower Devonian

Most known from early Carboniferous (> 325 mya)

Highly diverse, ancient

No wonder systematics is difficult!
Molecular Phylogeny of Arthropods

Providing new understanding, but not fully resolved yet

Because the origin of arthropods is ancient (> 550 million years ago) it is hard to resolve relationships
Congruence of Molecular and Morphological Phylogenies

A single method is unlikely to yield evolutionary patterns.

Gene-tree / species-tree problems can occur with molecular data.

Morphological and molecular data can lead to different conclusions in some cases but produce congruent conclusions in others.
Analyses of Arthropod Phylogenies

Evolution of the Ecdysozoa (Superphylum)

Includes, Insecta, Crustacea, Myriapoda, Chelicerata, Onychophora, Tardigrada and 5 phyla of worms including the Nematoda

Includes > 4.5 million living species in diverse ecological niches (Telford 2008)

Data indicate the Ecdysozoa are a natural monophyletic group
Analyses of Arthropod Phylogenies

Relationships among the Arthropoda by Regier et al. (2010) used 75 species and 62 single-copy nuclear protein-coding genes and Likelihood, Bayesian and parsimony methods.

Results support the Pancrustacea (Hexapoda plus Crustacea) hypothesis and the Mandibulata (Myriapoda plus Pancrustacea).
Analyses of Arthropod Phylogenies

Phylogeny of the **Holometabola** studied by Wiegmann et al. (2009) looked at 11 orders, all of which are thought to have originated in the late Carboniferous (318-300 mya), using 6 nuclear protein-coding genes.

All orders are monophyletic and the Hymenoptera are the basal-most lineage of the Holometabola. Strepsiptera are a sister group to the Coleoptera. Orders appear to have diverged rapidly between 274-213 mya.
Genomes and Arthropod Phylogenies

Will complete genome sequences lead to an understanding of what happened ~ 530 mya in the Cambrian explosion?

Repositories for alignments of whole-genome sequences have been set up

Next-Gen sequencing will soon “transform ecology and evolution by fundamentally changing the ranges and types of questions that can be addressed”
Speciation involves genetic differences that result in \textit{prezygotic isolating factors} (mating discrimination, habitat preferences) and \textit{postzygotic isolating factors} (hybrid inviability, sterility).

Reproductive isolation, along with selection and genetic drift, creates and expands any differences between species.
Species Concepts

Physical isolation (allopatry) leads inevitably to evolutionary change through natural selection or drift and pre- and postmating isolation mechanisms evolve as a by-product of the genetic changes.

Sympatry (living together in the same area) often leads to increased reproductive isolation.

Sometimes, speciation can occur sympatrically.
Species Concepts

Taxonomists produce different results based on different models or assumptions

**Biological species:** what most learn in introductory biology, but not sole concept

**Evolutionary species:** emphasizes continuity through time (lineages evolving separately from others)

**Phylogenetic species:** smallest diagnosable cluster of individuals with a parental pattern of ancestry and descent
Speciation Genes

Speciation assumed to be due to changes in more than one gene

- Unknown number usually (often polygenic)
- At least 2 required?
- Few genetic analyses conducted on number of speciation genes

Molecular tools offer method for analysis of speciation

See examples of Drosophila species
Cryptic Species

Molecular tools often ONLY way to ID cryptic species
- Often important pests
- Has implications for pest-management programs

RAPD-PCR often useful for preliminary screening
- Varroa and Ageniaspis
Cryptic Species Identity: RAPDs

2-6 Thailand, 7-16 Australian, 17-16 Taiwan *Ageniaspis*

UPGMA, Clustal4 analysis

Hoy et al. 2000
Conclusions

Molecular tools provide taxonomic and phylogenetic answers to both basic and applied problems.

Molecular methods are relatively young and are still undergoing development.

At some point, most questions become a 'judgment call' as to what is a species, genus, etc., because some taxonomists are 'splitters' and others are 'lumpers'.
Conclusions

How much would it cost to describe all animals?

About 1.4 million of the 6.8 million animals have been described.

One estimate (Carbayo and Marques 2011) suggests it will cost $263 billion to describe all animals and take 360 years, assuming descriptions occur at the current rate.

As a result, much work remains!