Principles of Molecular Systematics

Biodiversity Informatics Research Methods Workshop

National Museums of Kenya

23-27 September 2013

Biological Systematics

- Biological **systematics** is the study of the diversification of life on the planet Earth, both past and present, and the relationships among living things.
- Often confused with **taxonomy**, the science of describing, identifying, classifying, and naming of organisms.
- Systematists build trees (phylogenies) to show how organisms relate to each other.

Study of Systematics

- Traditionally, systematics studied by comparing characters of whole organisms.
- Different theories of systematics… phylogenetic systematics currently favored.
- All theories of systematics based on evolution.
- Theories are used to infer patterns of evolutionary change of organisms by inferring how the characteristics of organisms (genetic and phenetic) change.

Molecular Systematics

- “Stuff” of evolution is change in the genes or chromosomes of organism (ultimate level of evolution), which changes the proteins the genes code for, and ultimately changes the characteristics of organisms.

- Revolution in **molecular biology** has allowed systematists to infer patterns of evolutionary change by examining changes in the protein products of genes or changes in the genes themselves.

Molecular Systematics Methods

- Early studies of molecular systematics were termed *chemotaxonomy* and made use of proteins, enzymes, and other molecules which were characterized using techniques such as chromatography, gel electrophoresis and DNA-DNA hybridization.
- These early methods have largely been replaced today by DNA gene or whole genome, and RNA transcriptome sequencing.

DNA and RNA

- DeoxyriboNucleic Acid (DNA) and (RiboNucleic Acid (RNA) and the main two types of molecules used in molecular systematics today.
- Two types of DNA: nuclear and cytoplasmic.
- Three types of RNA: messenger (mRNA), ribosomal (rRNA) and transfer (tRNA).

Nuclear DNA

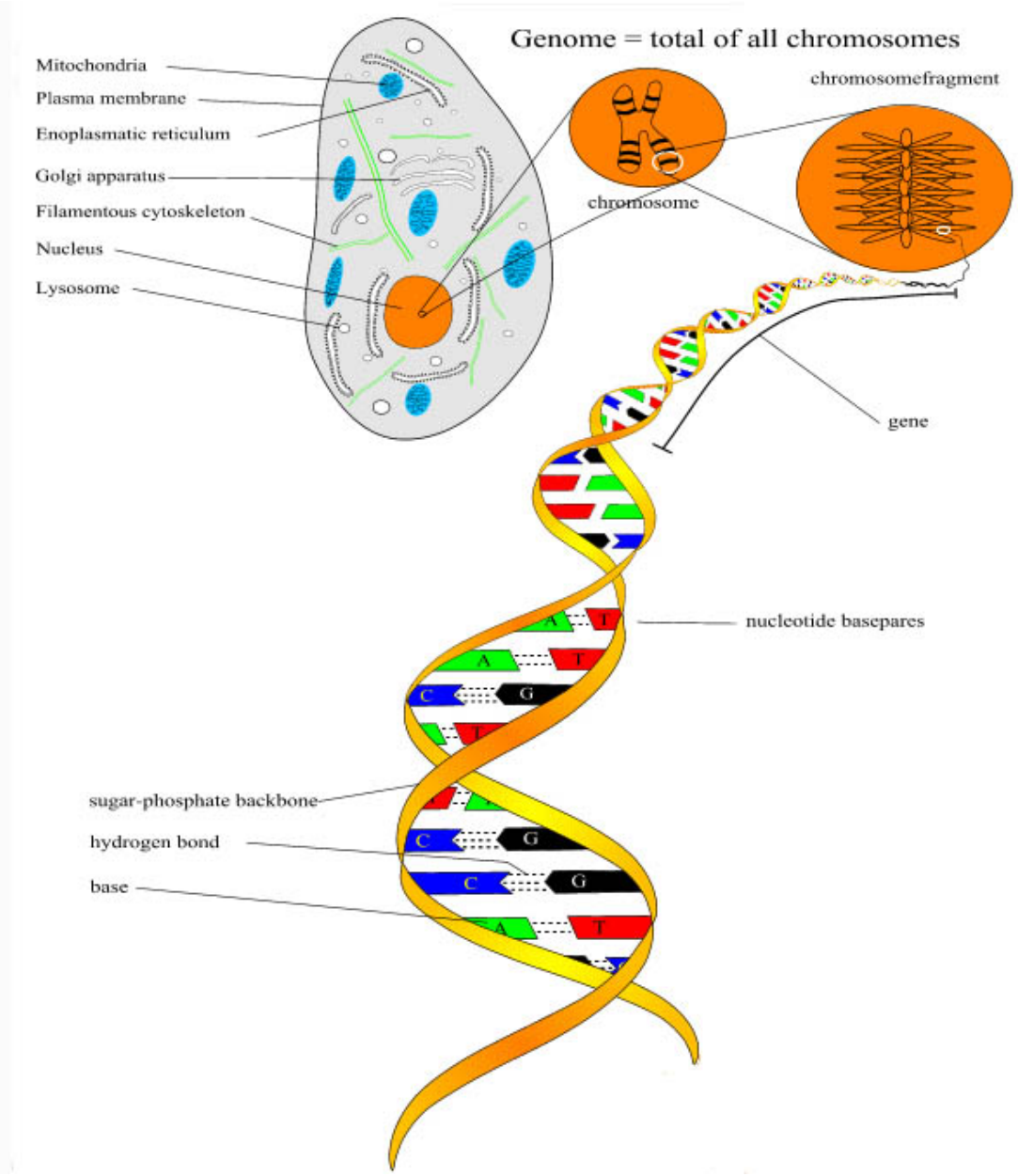
DNA contains the complete genetic information that defines the structure and function of an organism.

Nucleus contains most of the DNA in a cell (**nuclear DNA**).

DNA is double helix with nucleic acids (nucleotides) forming cross braces.

Four types of nucleotides in two classes:
Purines: Adenine (**A**), Guanine (**G**)

Pyrimidines: Cytosine (**C**), Thymine (**T**) (**U**racil replaces Thymine in RNA).



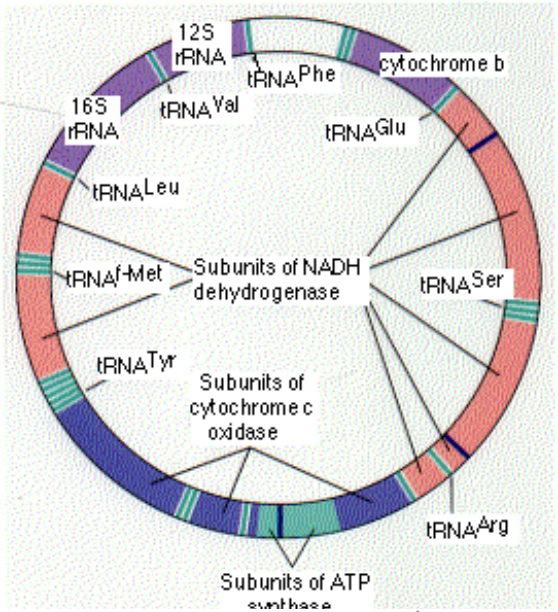
Cytoplasmic DNA

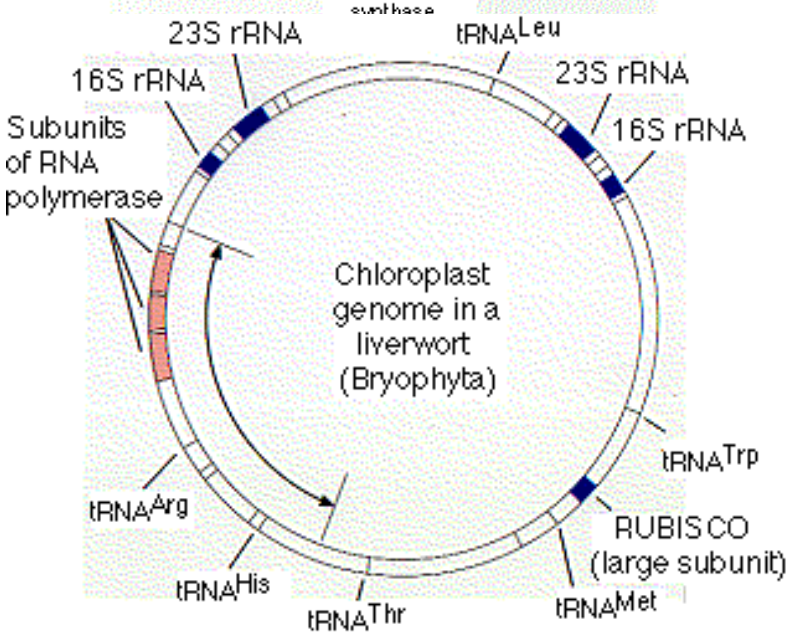
Other types of DNA found in cytoplasm.

Mitochondria, which play a role in the oxidative degradation of nutrient molecules, also contain DNA, called **mitochondrial DNA** (**mtDNA**).

Eucariotic cells that are capable of photosynthesis also contain **chloroplasts** with **chloroplast DNA** (**cpDNA**).

mtDNA and cpDNA double-stranded and circular, like DNA of procaryotes.





http://users.rcn.com/jkimball.ma.ultranet/BiologyPages/

Genetic Code

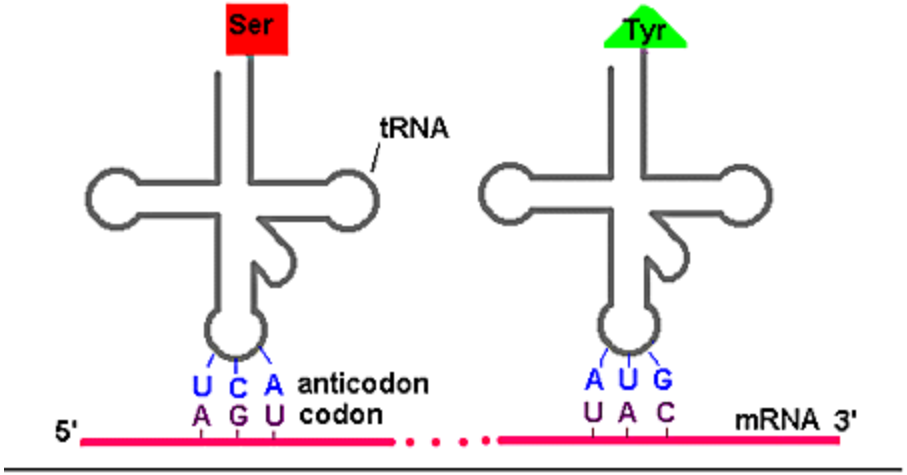
Nucleotides form **codons** in groups of three. Codons code for amino acids, the building blocks of proteins.

64 codons code for 20 amino acids; built in redundancy at **first** and **third** codon positions.

Changes in nucleotides at these positions do not change the amino acid. Such changes are selectively neutral or **synonymous**.

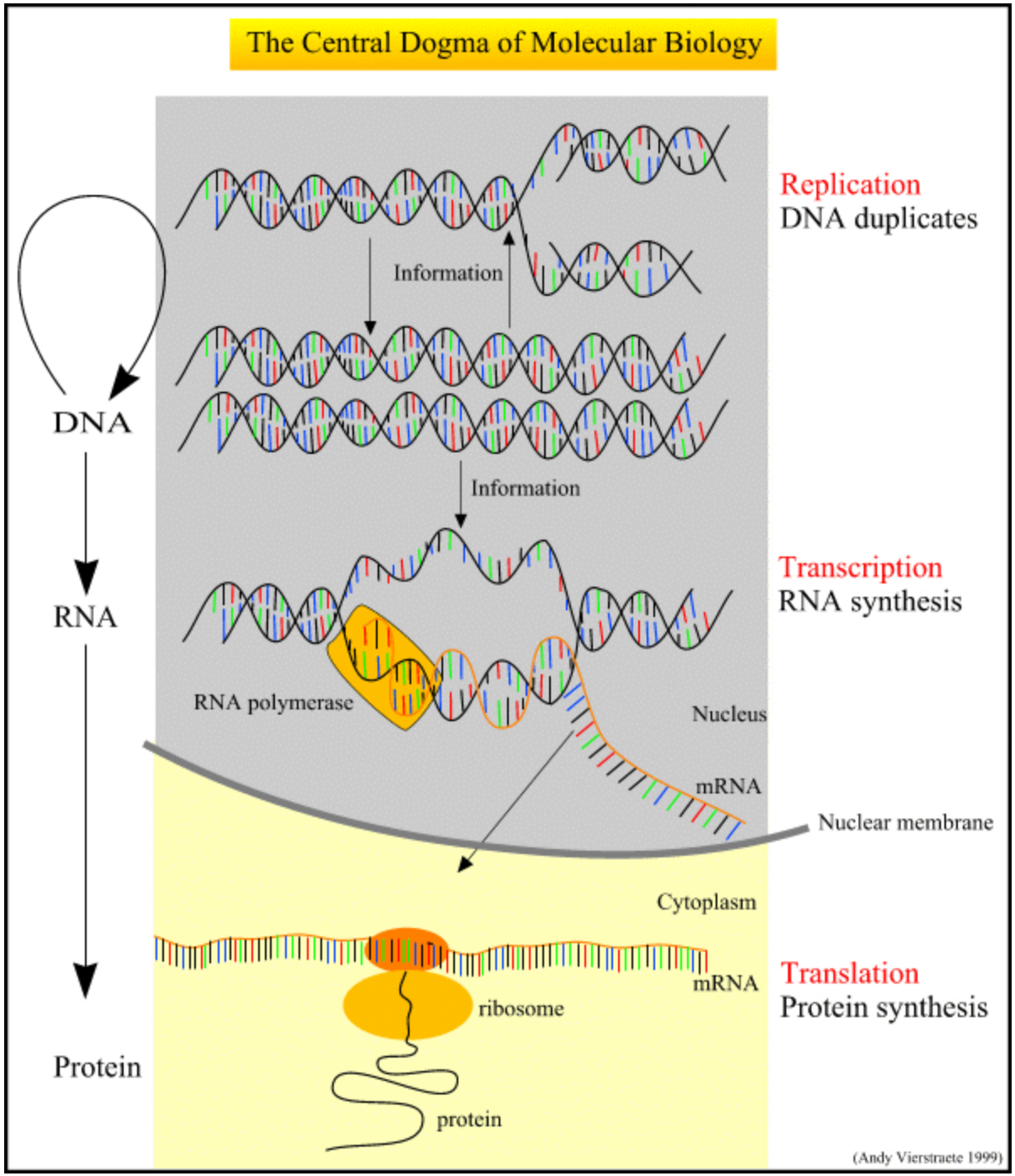
Substitutions that change the amino acid are called **non-synonymous** substitutions.

Most DNA is non-coding.



2nd base in codon

1st base in codon	U	C	A	G	3rd base in codon
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	STOP	A
	Leu	Ser	STOP	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G


The Central Dogma of Molecular Biology
Replication
DNA duplicates
Information
DNA
Information
Transcription
RNA synthesis
RNA
RNA polymerase
Nucleus
mRNA
Nuclear membrane
Cytoplasm
mRNA
Translation
Protein synthesis
ribosome
Protein
protein
(Andy Vierstraete 1999)

Molecular Systematics Laboratory Techniques

Polymerase Chain Reaction

Proteinase K is used to dissolve the cell and nuclear membranes, and dissociate proteins from the DNA. After this step, **phenol** is added to separate tissue components into an aqueous phase with the **DNA**, which can be precipitated with ethanol, a phenol phase with the dissolved proteins and fats, and an interface with denatured proteins. Very small quantities of DNA needed.

Polymerase chain reaction (PCR) is a technique used to amplify a number of copies of a specific gene for sequencing. In order to use PCR, one must already know the exact sequences of DNA flanking either side of the gene region of interest.

First step in PCR is to synthesize "***primers***" representing about 20 nucleotides of known sequence flanking either side of gene of interest on both strands of DNA.

(5')TTAACGGGGCCCTTTAAA.....TTTAAACCCGGGTTT
AATTGCCCCGGGAAATTT.....>
and:
<.....TTTAAACCCGGGTTT
AATTGCCCCGGGAAATTT.....AAATTTGGGCCCAAA (3')

Three steps to PCR:

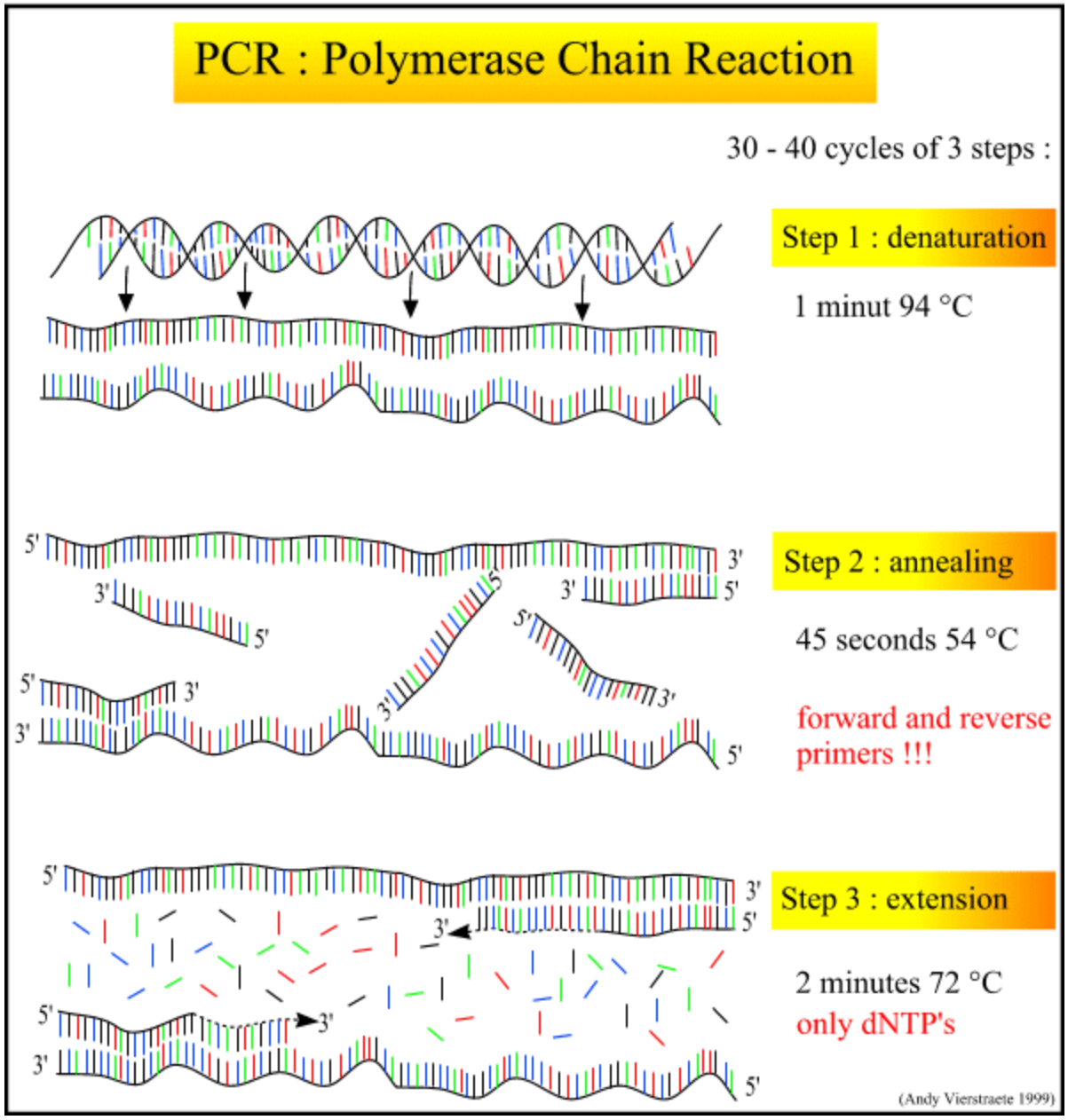
1. **Denaturation**
   heat, separate DNA

2. **Annealing**
   Cool, add primers

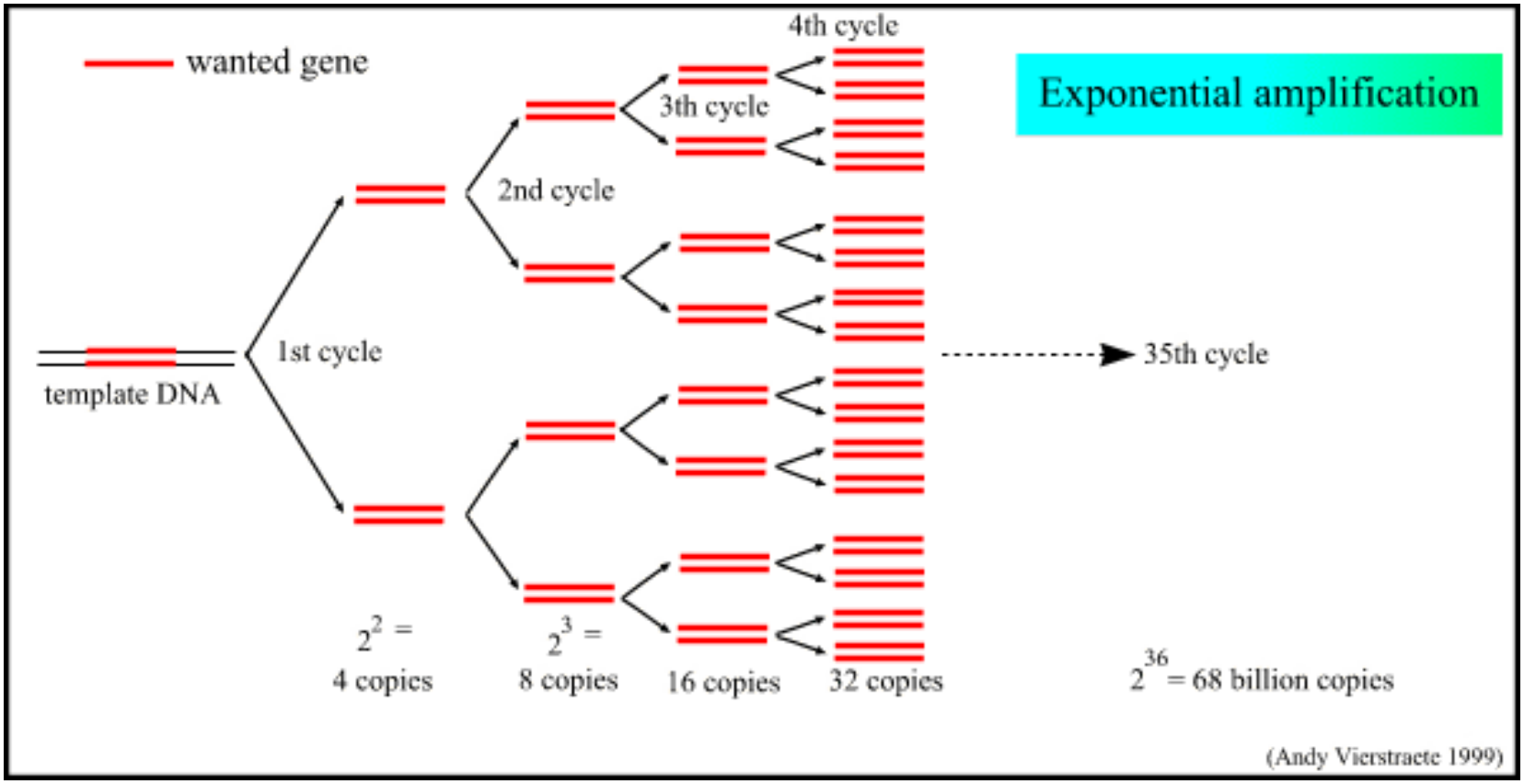
3. **Extension**
   ***Taq* polymerase** adds **dNTP's** to 3' end of primers.

Cycle 30-40 times

End result is multiple copies of gene of interest



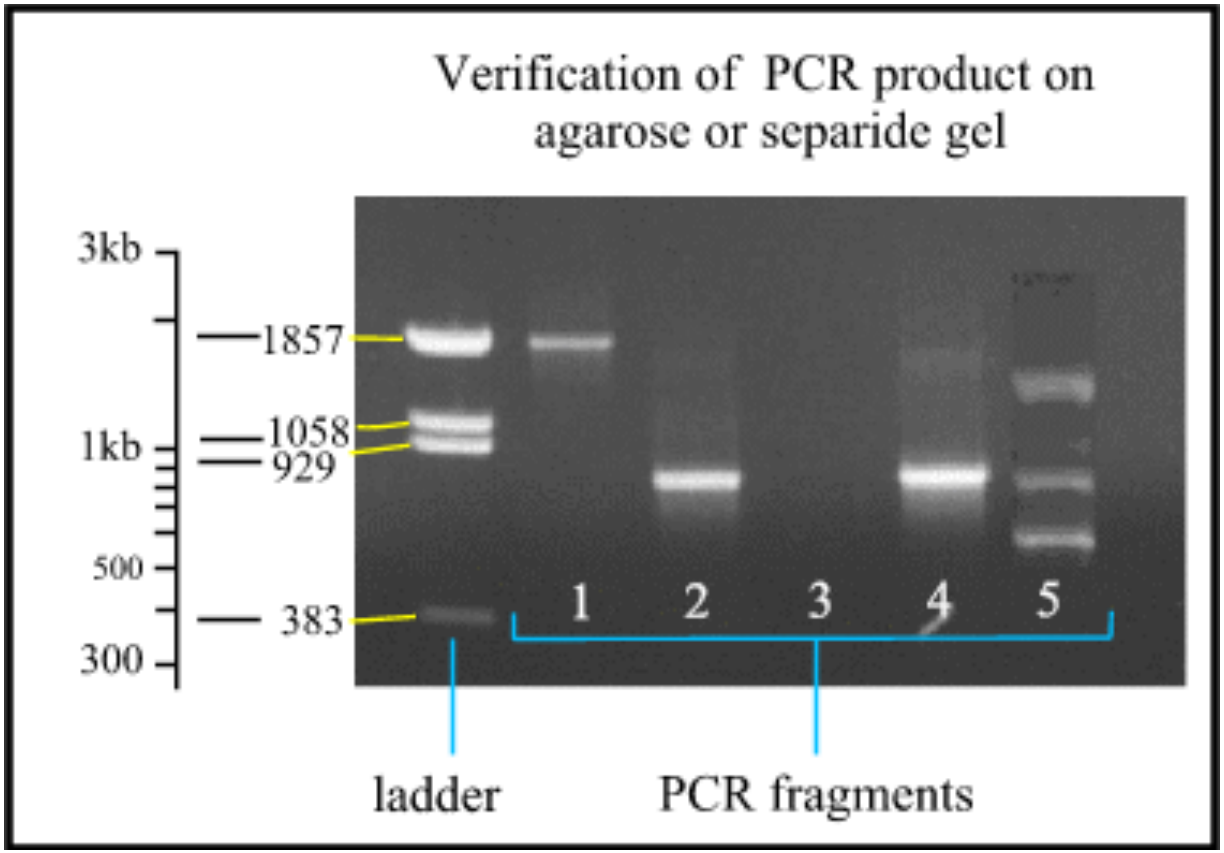
PCR Amplification



Exponential increase in number of copies of gene with each PCR cycle

Check PCR Product

Run PCR product out on gel to verify that it is the right size.



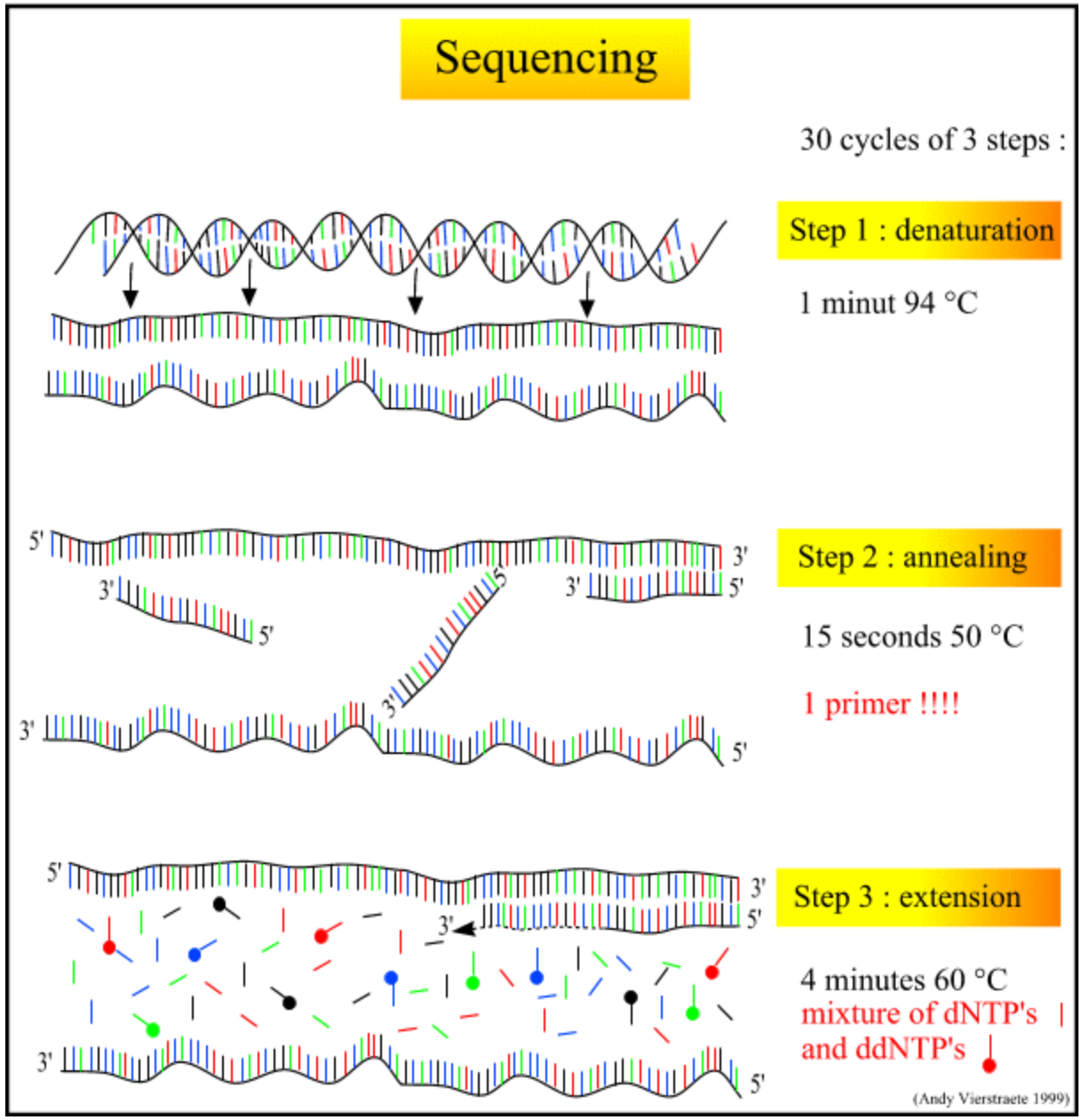
Three steps in PCR *sequencing* reactions:

1. **Denaturation**
   of double stranded *gene copies*

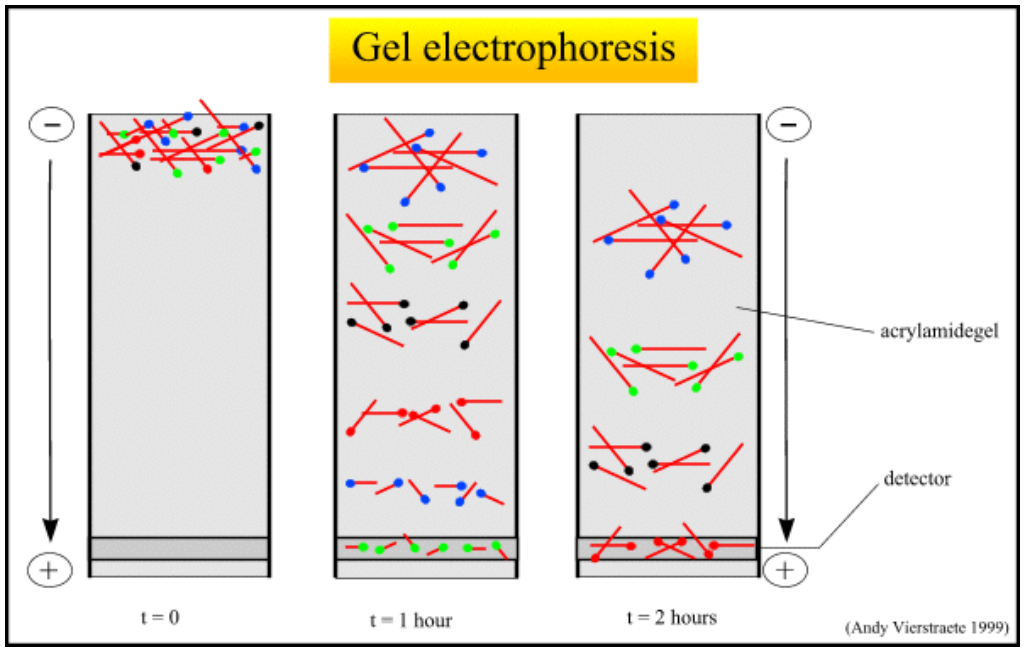
2. **Annealing**
   Cool, add primer for only one strand

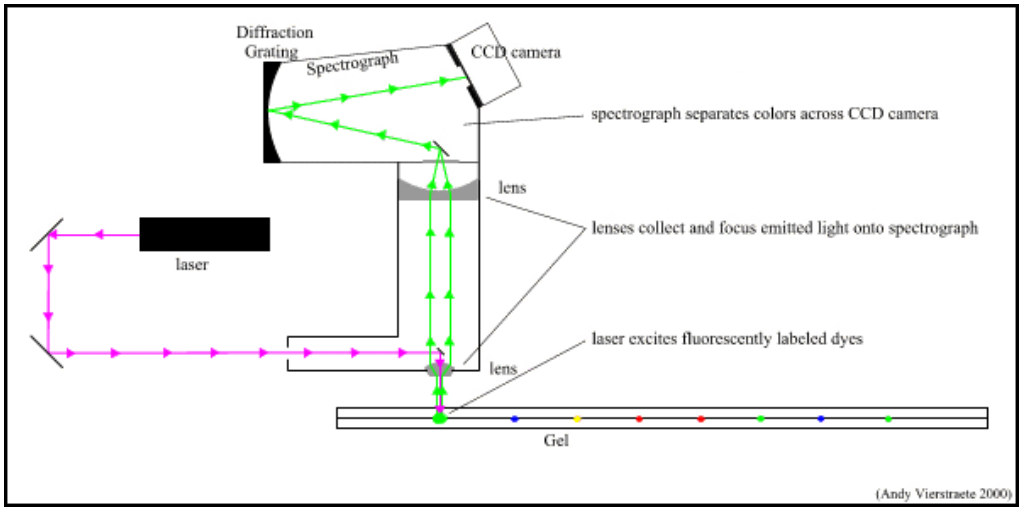
3. **Extension**
   Add mixture of **dNTP's** and **ddNTP's** (**labeled**).

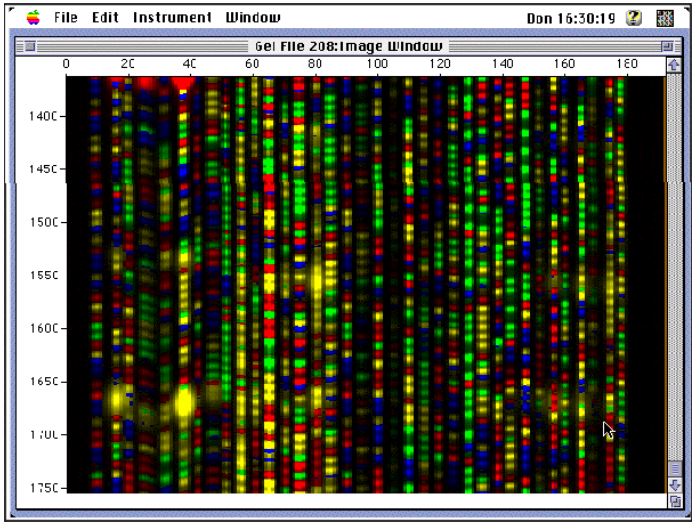
Cycle 30 times.



Automated Sequencing







ABI PRISM

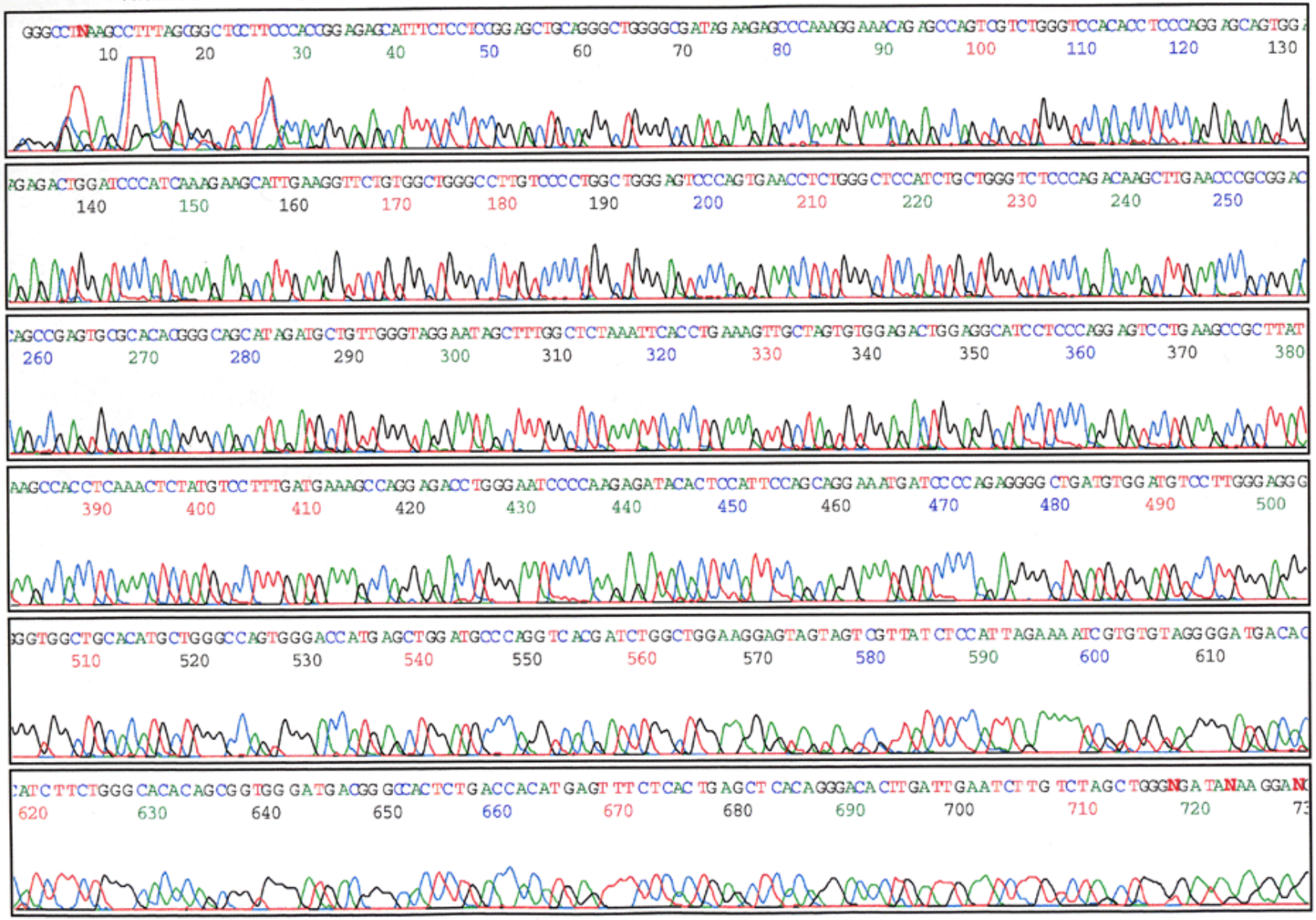
Model 377
Version 3.4.1
ABI100
Version 3.2

10•JEO433

JEO433
Lane 10

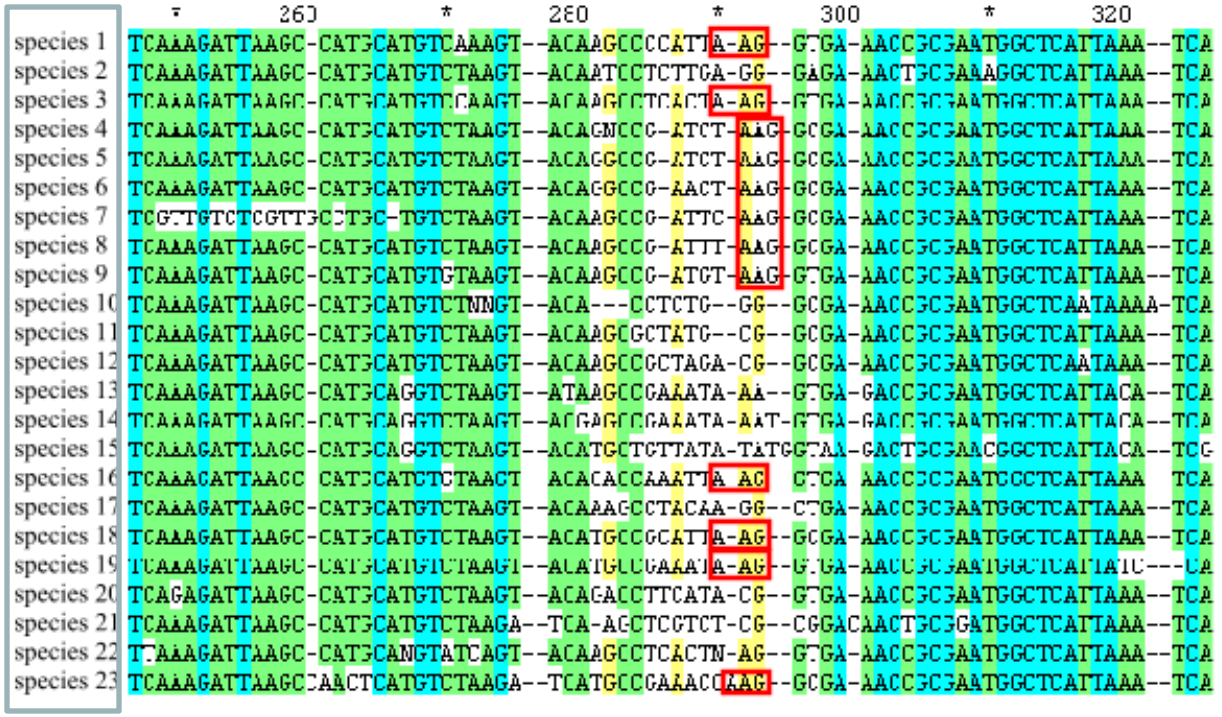
Signal G:254 A:158 T:126 C:141
DT377{BDv3}v1.mob
Big Dye3
Points 1277 to 10624 Pk 1 Loc: 1277

Tue, 2 Jul 2002 02:28 AM
Mon, 1 Jul 2002 05:01 PM
Spacing: 10.91{10.91}



Sequence alignment

Specimens tissues were taken from (**should be vouchered**).



To compare two or more sequences, it is necessary to first align them and identify locations of differences (insertions, deletions, substitutions).

Variety of computer programs available for aligning sequences. Can also align them by eye. Customary to ignore poorly aligned areas, but some programs allow these regions to be aligned as well.

Nucleotide substitutions

Substitutions are of two types: **transitions** (purine-purine, pyrimidine-pyrimidine) and **transversions** (purine-pyrimidine). Because bases on opposite strands of DNA typically match, transitions occur more frequently than transversions (**substitution bias**).

By one theory (**neutral theory**), **synonymous substitutions** assumed to occur at constant rate such that accumulation is clock like: the more substitutions, the longer taxa have been diverging.

Similarities in sequence *assumed* to be **homologous** (i.e., similarity due to common ancestry). With just four possible nucleotides, over time substitutions can change nucleotide at a particular position back to ancestral state (analogous similarity or **homoplaisy**).

Sequences become **saturated** with changes as a function of evolutionary time; sequence data can be adjusted (corrected) to account for this.

Distance methods

Differences in sequence among taxa can be viewed simply as pairwise divergence in a distance matrix (genetic or p distance = proportion of site differences).

A dendrogram can be calculated from the matrix with **clustering algorithms:**

UPGMA clustering (Unweighted Pair Group Method using Arithmetic averages).

Neighbor Joining: which corrects UPGMA method for its assumption that the rate of evolution is the same in all taxa.

Phylogenetic methods

Phylogenetic methods derive trees by considering the various pathways evolution could take to produce a set of sequence data.

They are based on **parsimony** or **resampling** methods. The resulting tree is called a **cladogram**.

Parsimony method evaluates all possible trees for **each position** in the alignment. Trees are given scores based on the number of evolutionary changes needed to produce the observed sequence changes. The **most parsimonious tree** is the one requiring the **fewest evolutionary changes** to derive all sequences from a common ancestor.

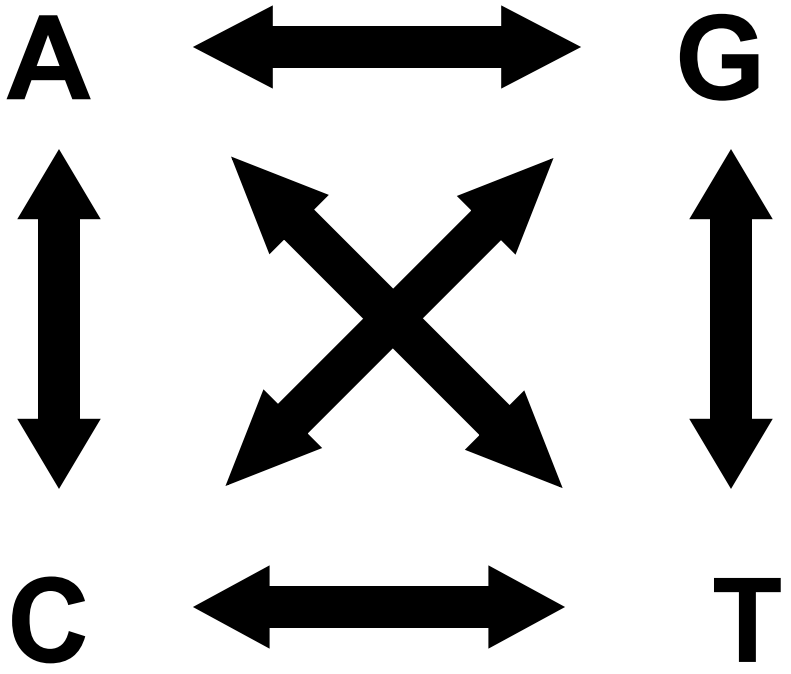
Resampling methods also use **each position** in an alignment, and evaluate all possible trees, but calculate the **support for each tree** using an explicit **model of evolution** (result not necessarily the most parsimonious). The support values for each aligned position are then multiplied to provide a combined support value for each tree. The tree with the most combined support is considered the most probable tree.

Models of Molecular Evolution

Nucleic Acid Models
- Jukes-Cantor
- Kimura
- Felsenstein
- Hasegawa

Amino Acid Models
- Codon
- Dayhoff
- Henikoff & Henikoff

Advanced Models
- Combined Models
- Heterogeneous Rates
- Structural Biology



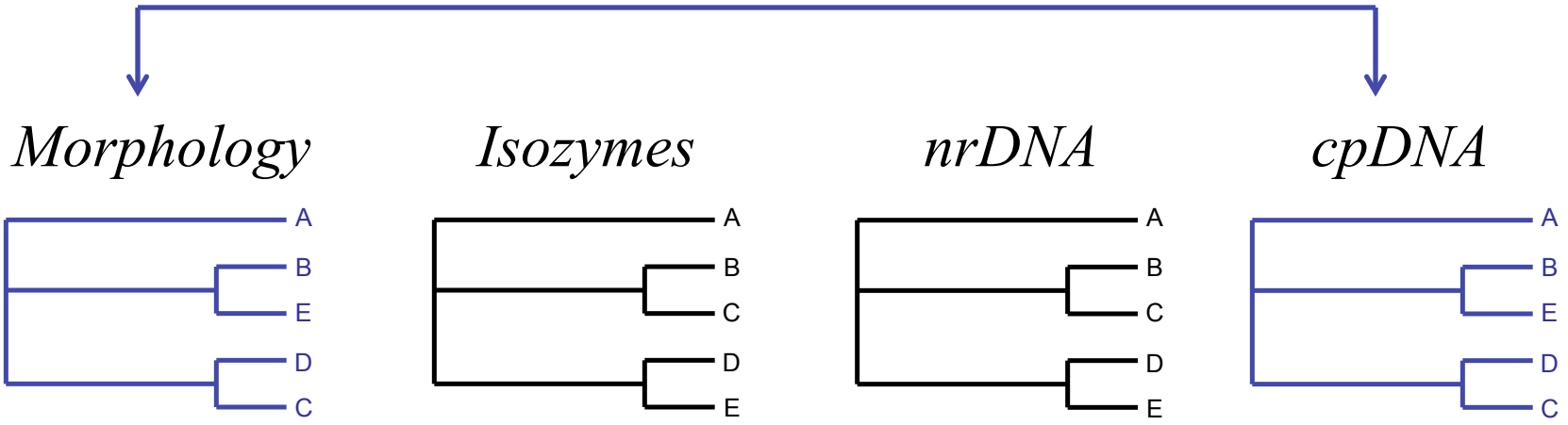
Phylogenetic Trees

One to many phylogenetic trees can result from phylogenetic analysis of molecular data (as many trees as characters).

Can either produce a **consensus tree**, or have to decide which topology has most support (**branch support methods**).

Also problems of congruence among different kinds of data (molecular or otherwise)…could perform a **total evidence** analysis i.e. combine all data into a single matrix.

Incongruence of trees based on different types of data

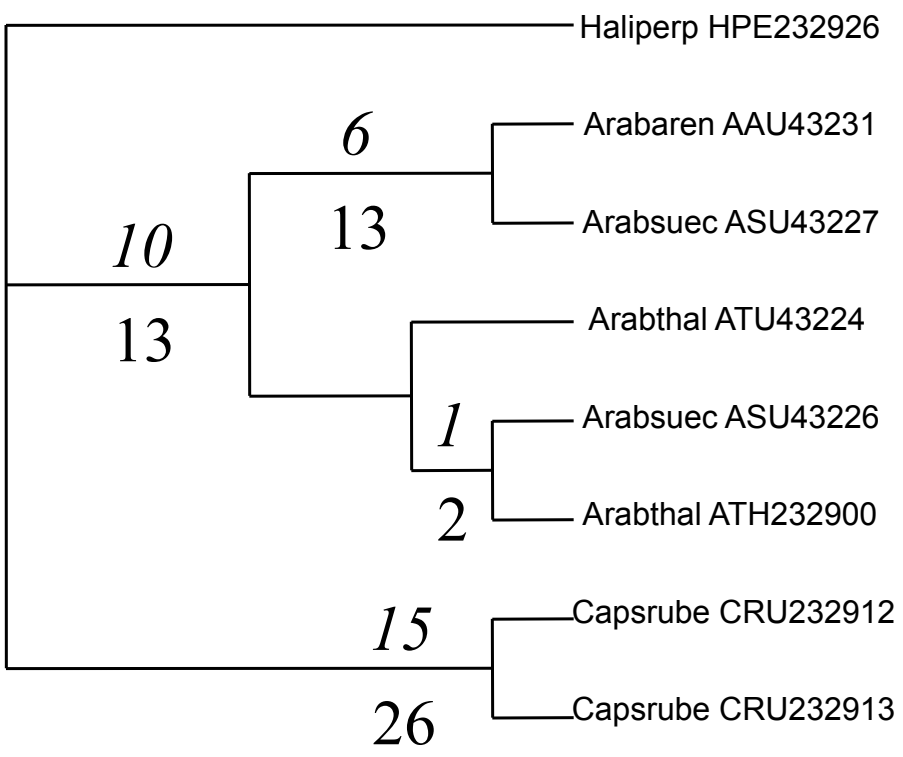


Common tree support measures

- Congruence of ”independent” data sets
- Bremer support
  - *Parsimony framework*
- Resampling methods (bootstrap, jackknife)
  - *Any method applicable to discrete data*
- Monte Carlo methods
  - *model-based methods*

Bremer Support (decay index)

- Number of steps (character state transformations) it would take to collapse the group.
- Can't be larger than the branch length (character support).
- No direct connection to branch length otherwise
- Quantifies branch support within a parsimony framework.



Resampling Methods

- Statistical methods for resampling data and inferring an unknown error distribution.
- Draws a random subset (’pseudoreplicate’) of data (characters) many times.
- Infer the error distribution from all these random subsets.
- Assumes data set is large enough to accurately reflect the true error distribution.
- Assumes data at hand are identically and independently distributed (i.e., sampled at random).
- Assumptions violated for most (if not all) data sets used in phylogenetics.

Jackknife

- Random sampling of original data ***without replacement*** (leaving out one observation at a time).
- Proportion drawn affects result.

Bootstrap

- Random samples of original data ***with replacement*** until one gets a data size as large as the original.
  - sensitive to the number of characters in the matrix. When $n>\infty$, jackknife-delete-1/e and the bootstrap coincide, but only when there are no conflicts in data.

Resampling statistics in phylogenetics

- ”…provides us with a confidence interval… [of] the phylogeny that would be estimated on repeated sampling of many characters from the underlying pool of characters” (Felsenstein 1985)
- True? We don’t know. The exact statistical interpretation remains unclear.

Computation

- For large data sets (many taxa) exact solutions for any method employing an optimality criterion (parsimony, likelihood, minimum evolution) are not possible
- Errors due to false solutions add to the random error and increases uncertainty about the interpretation of resampling frequencies, but…..

Thoroughness of tree searches

- ”...no branch-swapping is required for finding well-supported groups, and jackknifing eliminates poorly supported groups” (Farris et al. 1996)...

Resampling statistics: Cautionary remarks

- Can be employed with any tree-building method
- Give reasonable measures of support
- Jackknife and bootstrap give different, but similar results
- The exact probabilistic meaning of resampling frequencies is unknown
- Error introduced by heuristic tree searches is largely unexplored

Monte Carlo methods

- Simulation from a theoretical model
  - Parametric bootstrapping
  - Bayesian phylogenetic inference

Parametric bootstrapping

- Repeated simulation of a phylogeny using a particular evolutionary model. Allows:
  - evaluation of specific, competing hyptheses: ”Is tree A significantly better than tree B?”
  - examining potential sources of errors, i.e. recombination and long branch attractions
  - conduct power analyses, i.e. how much data is needed to resolve a given tree?

Bayesian phylogenetic inference

- Prior beliefs (model) are incorporated in probability calculations
- No attempts are made to find the ”best” tree
- The parameter space (e.g. tree topologies) is investigated using Markov Chain Monte Carlo (MCMC) algorithms

Bayesian phylogenetic inference

- The frequency with which a given clade is encountered in the MCMC iterations approaches its posterior probability!
- Computable for large data sets
- Results vary according to the prior beliefs (i.e. the evolutionary model)

Internet Sources

Internet sites I used to put together this presentation (please acknowledge them if you decide to use any of this):

- http://allserv.rug.ac.be/~avierstr/index.html
- http://morphbank.ebc.uu.se/systbio/Literature/supportslides/support.ppt
- http://people.ku.edu/~jbrown/pcr.html
- http://www.bioinf.org/molsys/lectures.html
- http://www.may.ie/academic/biology/james/
-