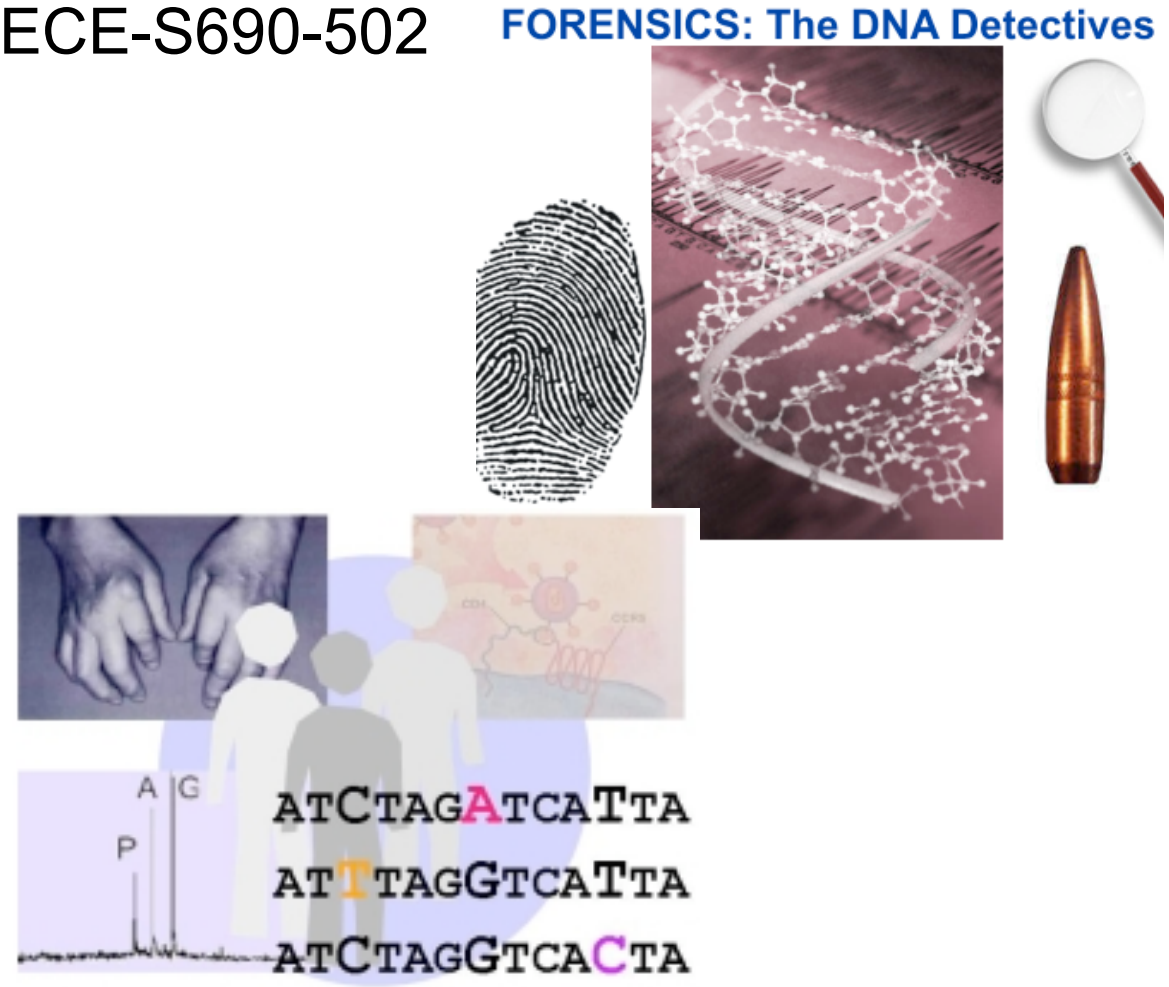
Mutations, Phylogeny, and Tandem Repeats

ECE-S690-502



Today's Objectives

- Types of Mutations
- Reconstructing Phylogenetic Trees
- Brief Communication Theory for Mutations
- Tandem Repeat Detection Algorithms

Some background definitions

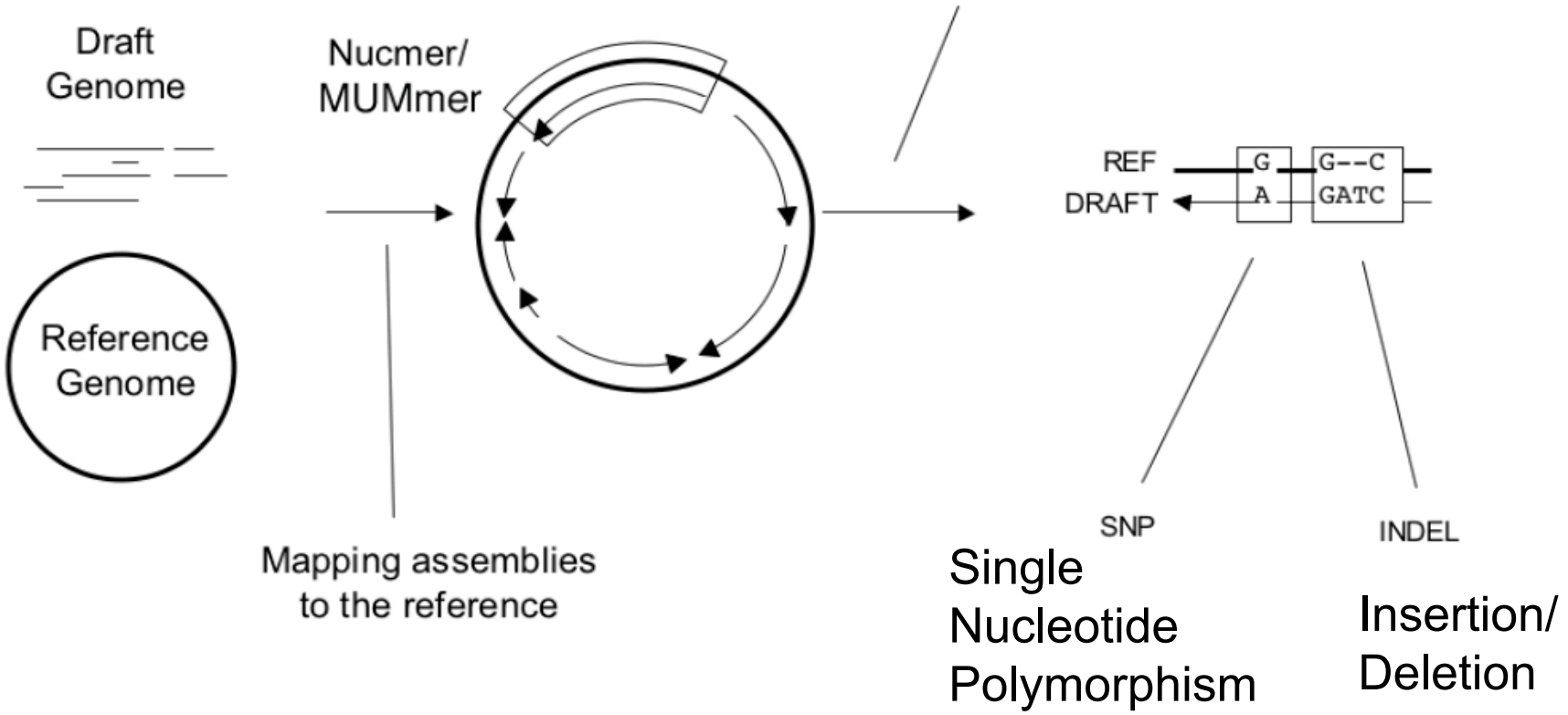
- Locus: The position of a gene on a chromosome.
- Allele: 1) One of the variant forms of a gene at a particular locus, or location, on a chromosome. Different alleles produce variation in inherited characteristics such as hair color or blood type.

  2) May refer to just one pattern …

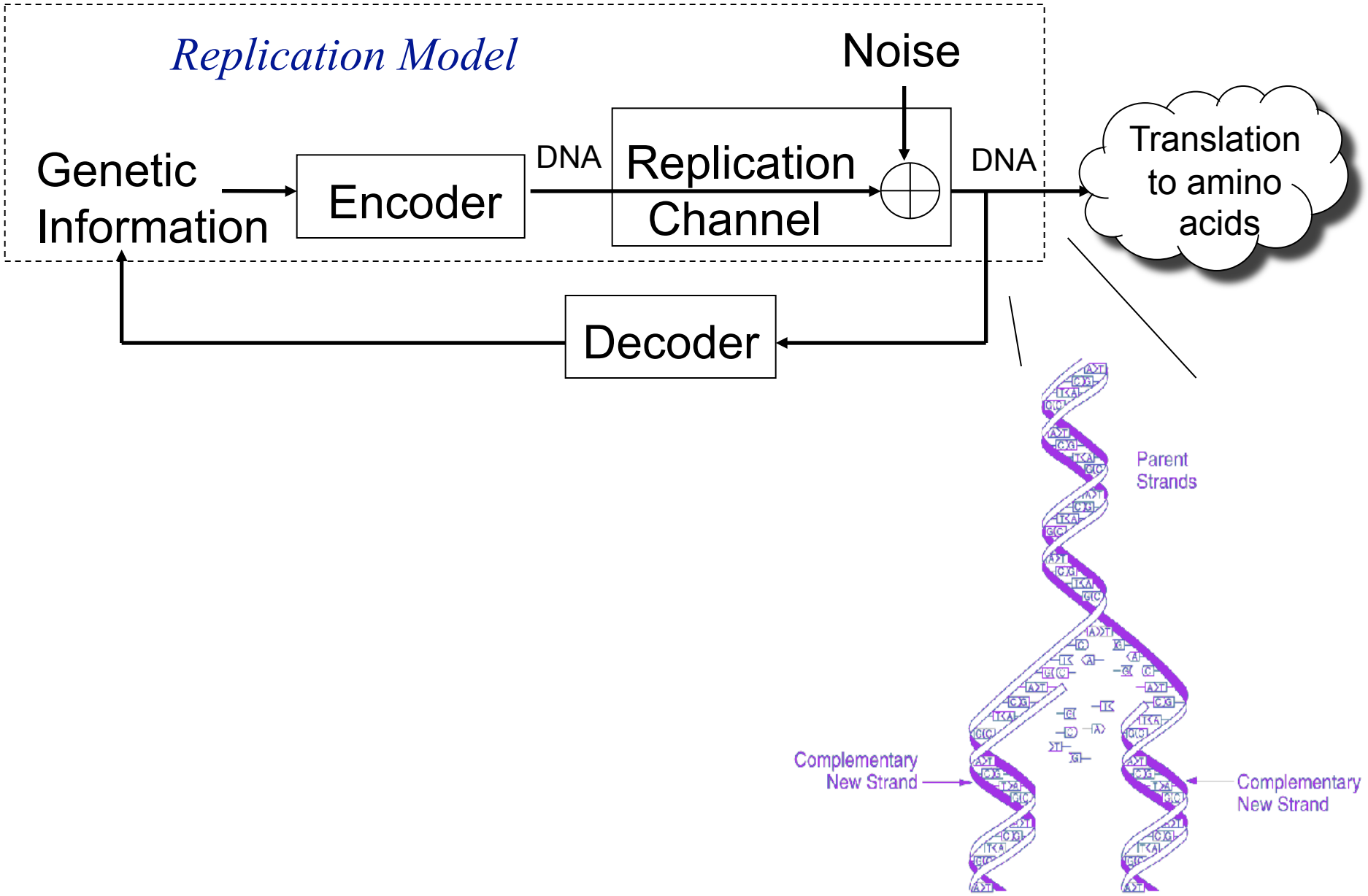
  AATTGG is the allele that is repeated

Polymorphism

• Polymorphism: (literally means -- “having many forms”) A change in the sequence of DNA associated with a large portion of the population. Polymorphisms may or may not be linked to specific diseases.



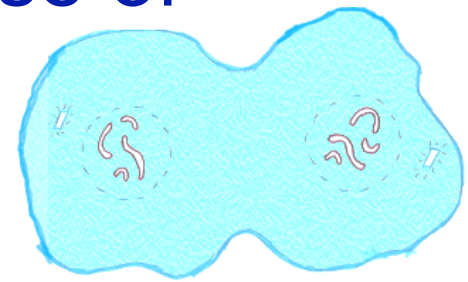
DNA Channel Capacity Model



Error Rates in DNA Replication

- DNA replication has a proofreading mechanism.
- With proofreading: error frequency of approximately 10e-10.
- Without proofreading: error frequency increases to 10e-3 ->10e-5. (viruses do not have proofreading)

Since a human cell division involves 6 x 10^9 bases, there is about one mutation per cell division. (not so bad because of genetic code).

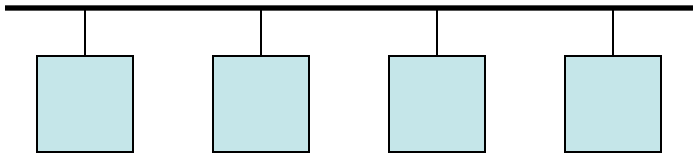


Error and Mutations: Can't live with them, no progress without them

- The many Faces of Mutations
  - Mostly Bad: Death and Disease
  - Occasional Good: New environmentally-robust trait
  - All: Offers insight into genetic variation, genetic trail/markers -- good for analysis

Noisy Channel: Mutations

- Substitutions



- Insertions

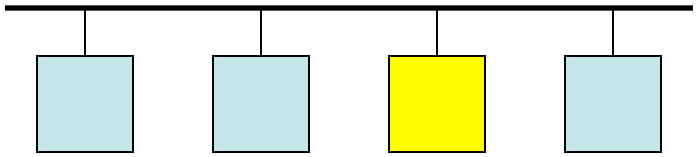
- Deletions

Error rates

- With proofreading: 10e-10
- Without: 10e-3 ->10e-5

Noisy Channel: Mutations

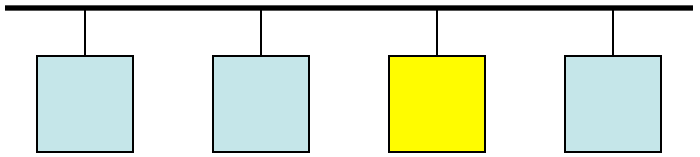
- Substitutions



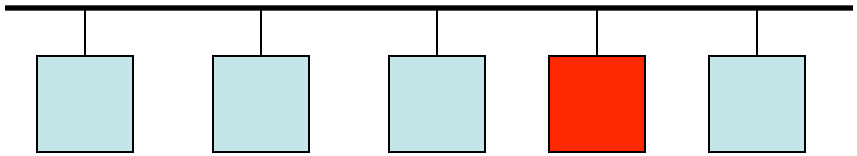
- Substitutions: C->T most common error in replication.
  - Cytosine deamination is 100x faster in single-strand DNA.

Noisy Channel: Mutations

- Substitutions



- Insertions



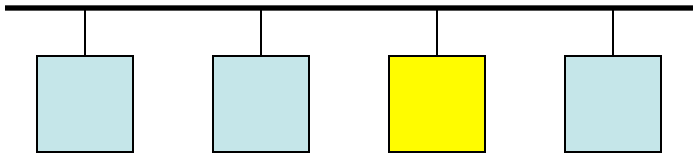
- Deletions

Error rates

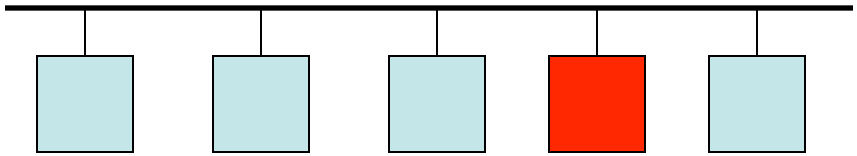
- With proofreading: 10e-10
- Without: 10e-3 ->10e-5

Noisy Channel: Mutations

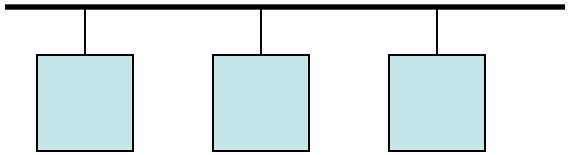
- Substitutions



- Insertions



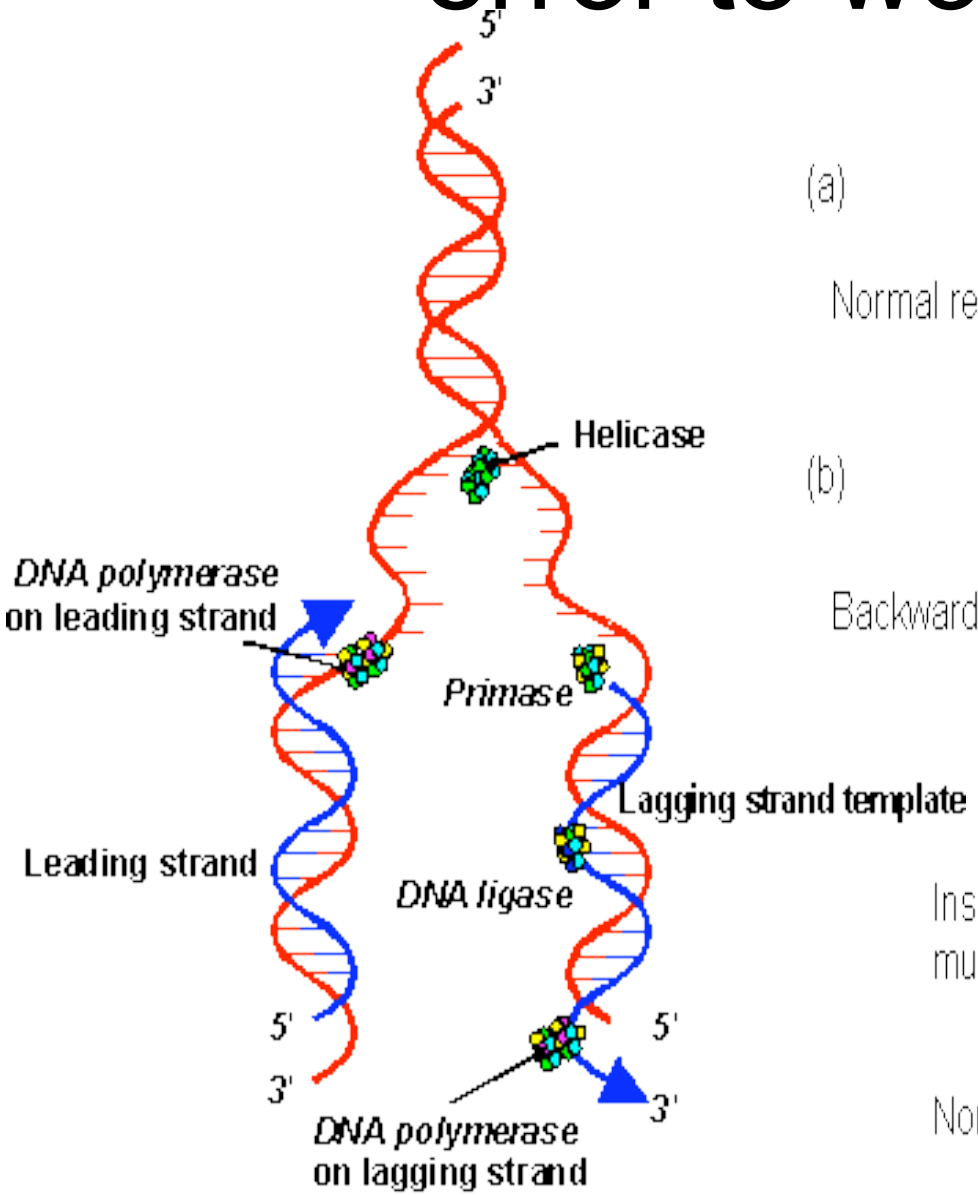
- Deletions

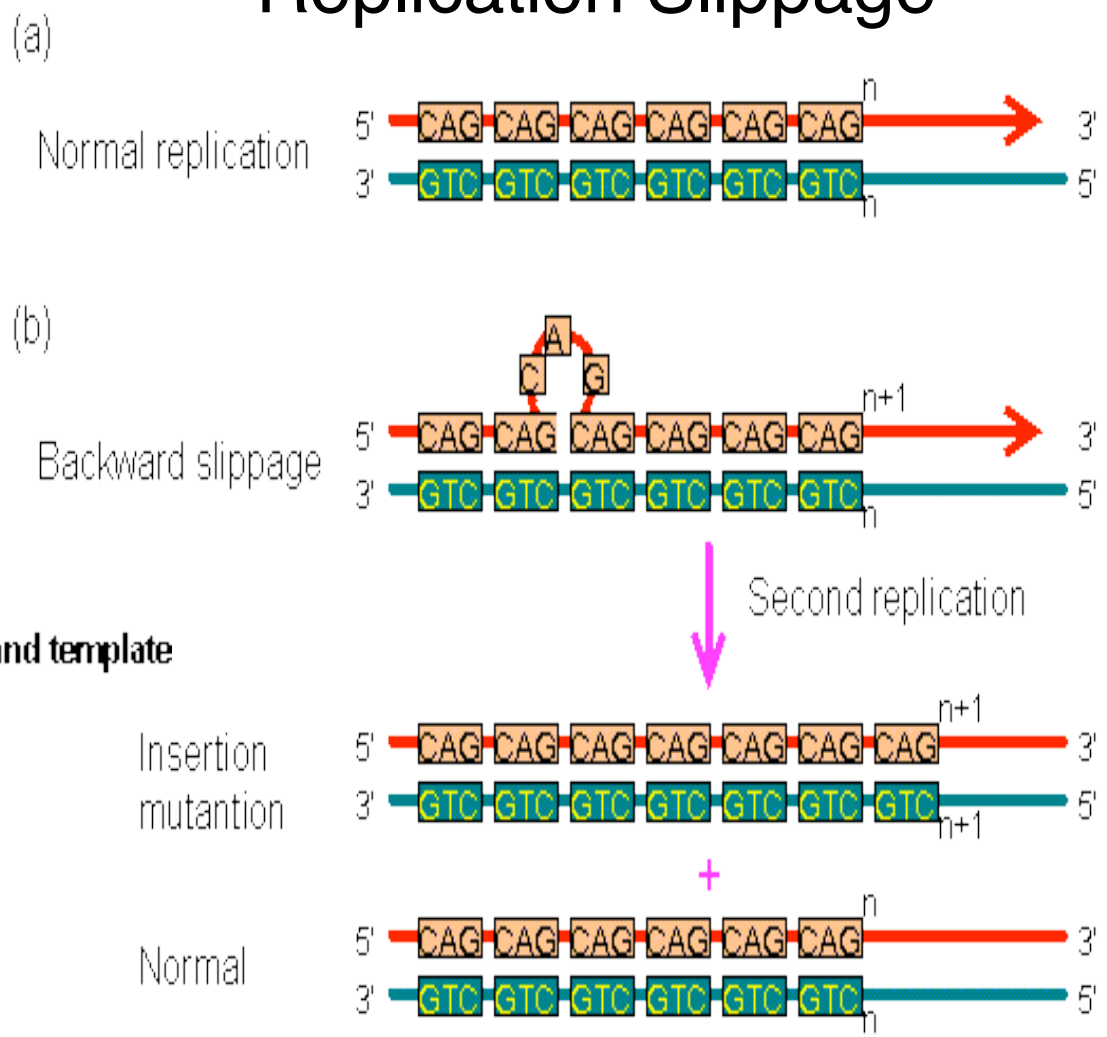


Error rates

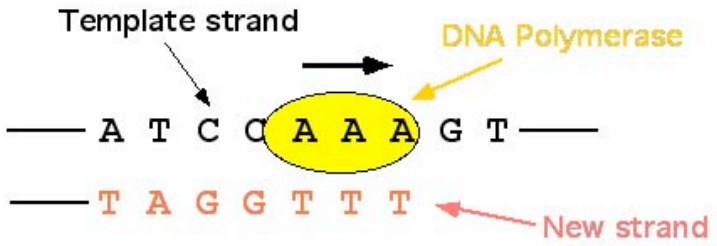
- With proofreading: 10e-10
- Without: 10e-3 ->10e-5

Slippage, one more replication error to worry about

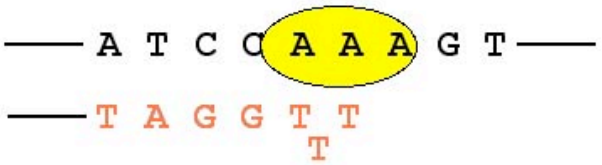




Tandem Repeats are Polymorphisms too



Strand slippage



Final strand

The Aspects of Tandem Repeats

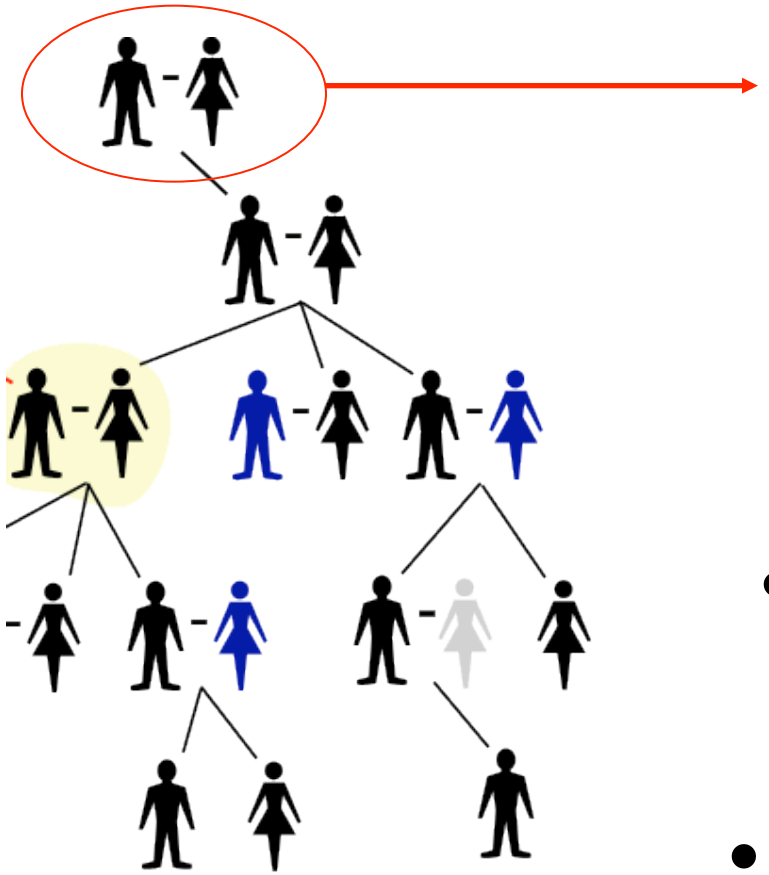
- Remember problems in putting together very similar sequences (which order?!)
- Forensics
- Evolution/Lineage
- Genetic Diseases
  - Example: Myotonic Dystrophy (CTG repeat in DMPK gene)
  - 5-37 repeats: No effect
  - 50-80: Male-pattern baldness
  - >80: Myotonic Dystrophy

Short Tandem Repeats (STR)

• Minisatellites: repetitive, generally GC-rich, variant repeats that range in length from 10 to over 100 bp. Overall length is greater than 0.5kb

• Microsatellites: polymorphic loci present in ***nuclear*** DNA that consist of simple sequence repeats (SSRs) of units of 1-4 base pairs in length, repeated 10 to 100 times..

Tracing Lineage (using SNPs and STRs)



Father: Y Chromosome
Mother: Mitochondrial DNA

- Y Chromosome never "recombines" with Mother's
- Only Men: Y Chromosome

- We get ~100% Mom's Mitochondrial DNA
- Everyone: Mitochondrial

Unique Event Polymorphism -- Genetic Marker

- SNP (Single Nucleotide Polymorphism) introduced) -- 1 in 100 million chance
- Passed now from generation to generation
- Haplogroup: is the group of all the descendants of the single person who first showed that UEP mutation.
- Haplotype: A person's individual footprint of all tested genetic markers. Even the difference of a single genetic marker delineates a distinct haplotype.

Lineage with STR Genetic Markers

- Much higher mutation rates than SNPs
- Results in usage for “smaller” timescales
- Small as in differences between parents and children, siblings (forensics)
- Consensus sequence can also show longer term effect

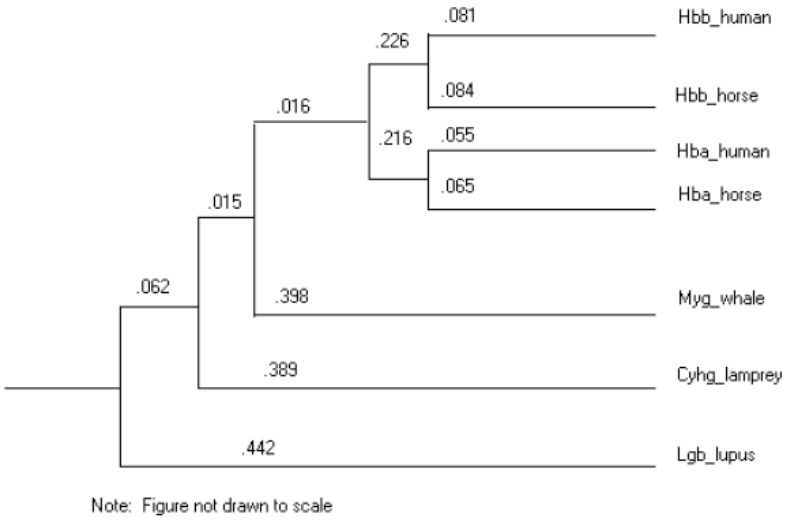
Tracking Human Migration with Genetic Markers

- Spencer Wells - Journey of Man

Constructing Phylogenetic Trees

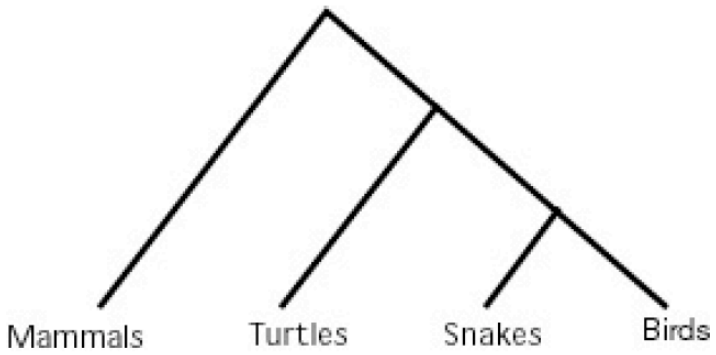
- Phylogenetic trees illustrate the evolutionary relationships among groups of organisms, or among a family of related nucleic acid or protein sequences
- E.g., how might have this family been derived during evolution

Globin Sequences



Note: Figure not drawn to scale

Hypothetical Tree Relating Organisms



Phylogenetic Relationships Among Organisms

- Entrez: www.ncbi.nlm.nih.gov/Taxonomy
- Ribosomal database project:

  rdp.cme.msu.edu/html/
- Tree of Life:

  phylogeny.arizona.edu/tree/phylogeny.html

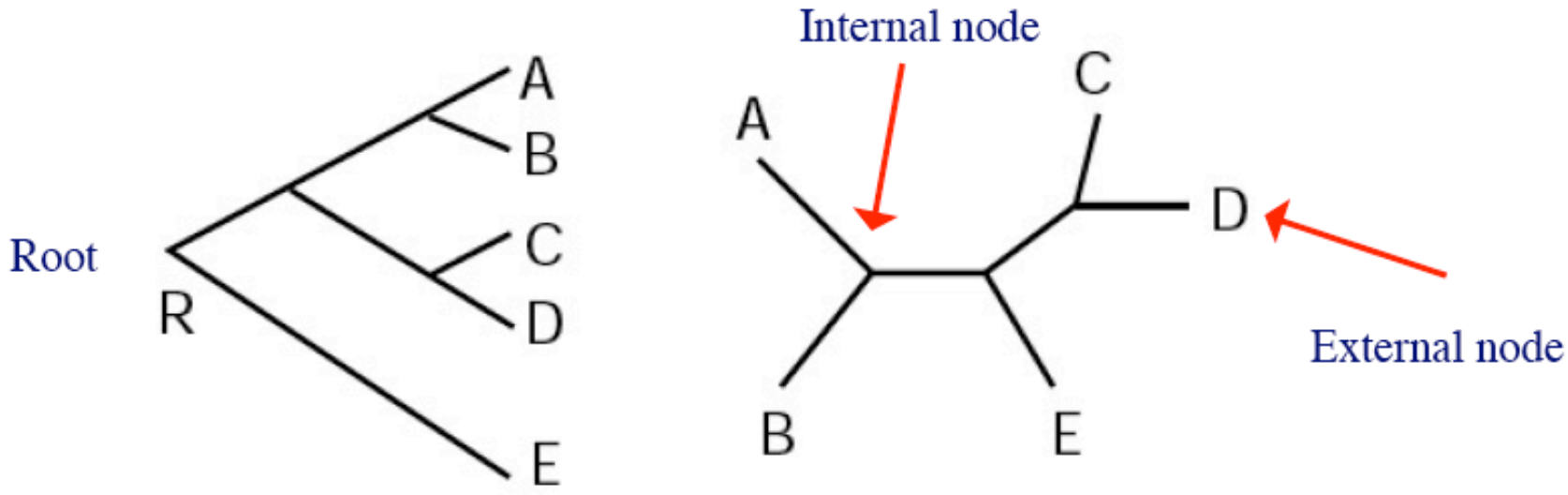
Phylogeny Applications

- Tree of life: Analyzing changes that have occurred in evolution of different organisms
- Phylogenetic relationships among genes can help predict which ones might have similar functions (e.g., ortholog detection)
- Follow changes occuring in rapidly changing species (e.g., HIV virus)

Traditional Methods

- Traditionally: morphological features (e.g.,number of legs, beak shape, etc.)
- Today: Mostly molecular data (e.g., DNA and protein sequences)

Rooted vs. Unrooted Trees



Rooted tree

Unrooted tree

Note: Here, each node has three neighboring nodes

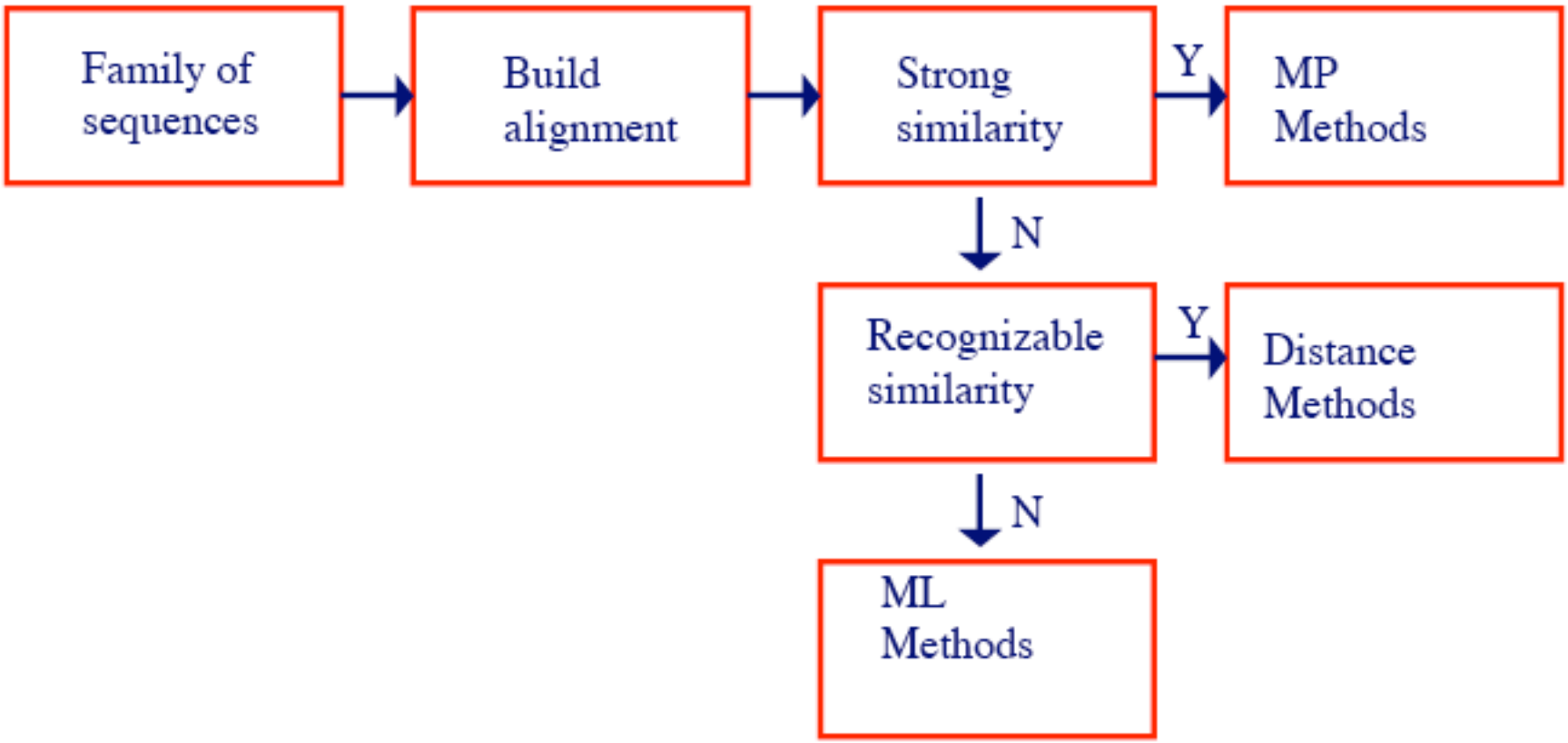
Terminology

- External nodes: things under comparison; operational taxonomic units (OTUs)
- Internal nodes: ancestral units; hypothetical; goal is to group current day units
- Root: common ancestor of all OTUs under study. Path from root to node defines evolutionary path
- Unrooted: specify relationship but not evolutionary path
  - If have an outgroup (external reason to believe certain OTU branched off first), then can root
- Topology: branching pattern of a tree
- Branch length: amount of difference that occurred along a branch

Tree construction methods

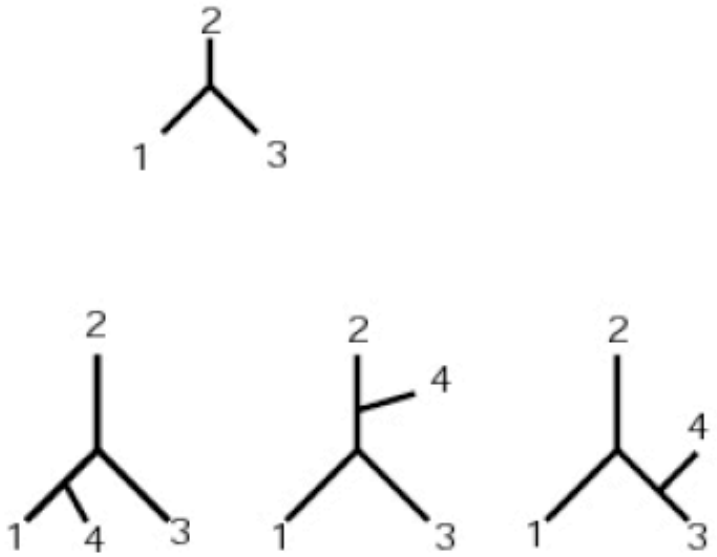
- Distance methods: evolutionary distances are computed for all OTUs and build tree where distance between OTUs "matches" these distances
- Maximum parsimony (MP): choose tree that minimizes number of changes required to explain data
- Maximum likelihood (ML): under a model of sequence evolution, find the tree which gives the highest likelihood of the observed data

Phylogeny Flowchart



Number of Possible Trees

Given n OTUs, there are $\prod_{i=3}^{n}(2i-5)$ unrooted trees



OTUs	unrooted trees
3	1
4	3
5	15
10	2,027,025

Number of possible trees

Given n OTUs, there are $\prod_{i=3}^{n}(2i-3)$ rooted trees

Bottom Line: an enumeration strategy over all possible trees to find the best one under some criteria is not feasible!

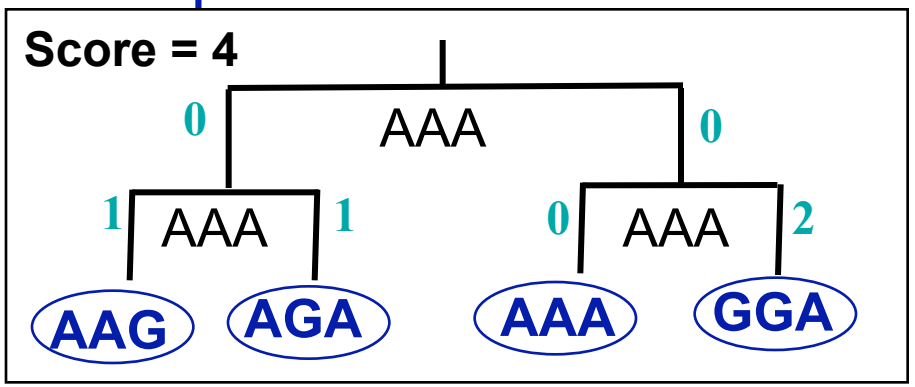
OTUs	Rooted trees
3	3
4	15
5	105
10	34,459,425

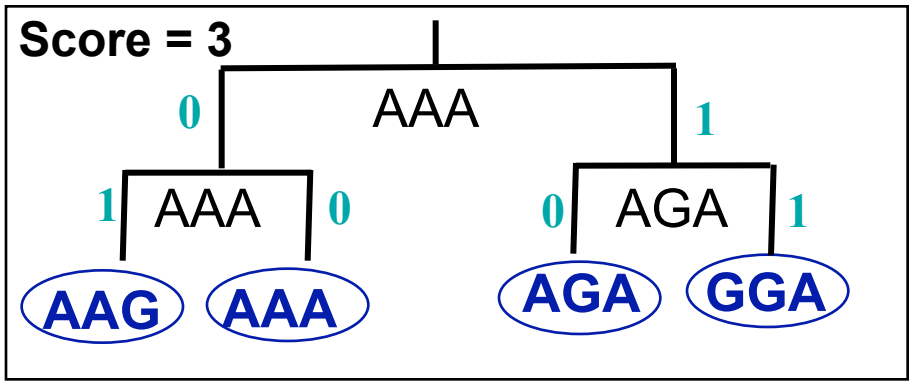
Most Parsimonious Tree

Parsimony-score:

Number of character-changes (mutations) along the evolutionary tree
(tree containing labels on internal vertices)

Example:





Most parsimonious tree:

➔ Tree with minimal parsimony score

Minimal Evolution Principle

Small vs. Large Parsimony

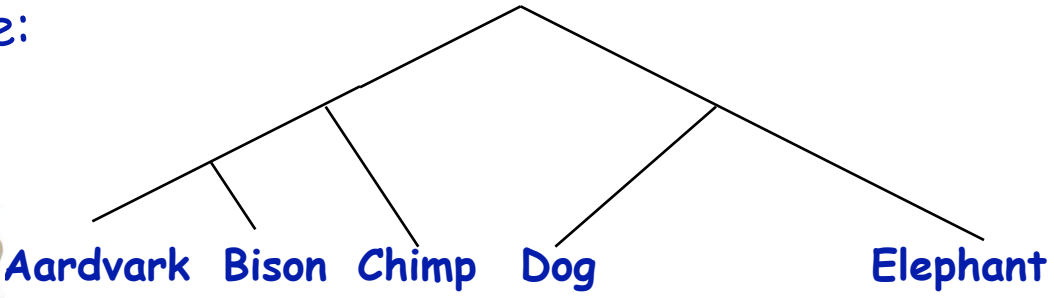
We break the problem into two:

1. Small parsimony: Given the topology find the best assignment to internal nodes
2. Large parsimony: Find the topology which gives best score

→ Large parsimony is NP-hard
→ We'll show solution to small parsimony (Fitch and Sankoff's algorithms)

Input to small parsimony:
tree with character-state assignments to leaves

Example:



A: CAGGTA
B: CAGACA
C: CGGGTA
D: TGCACT
E: TGCGTA

Parsimony

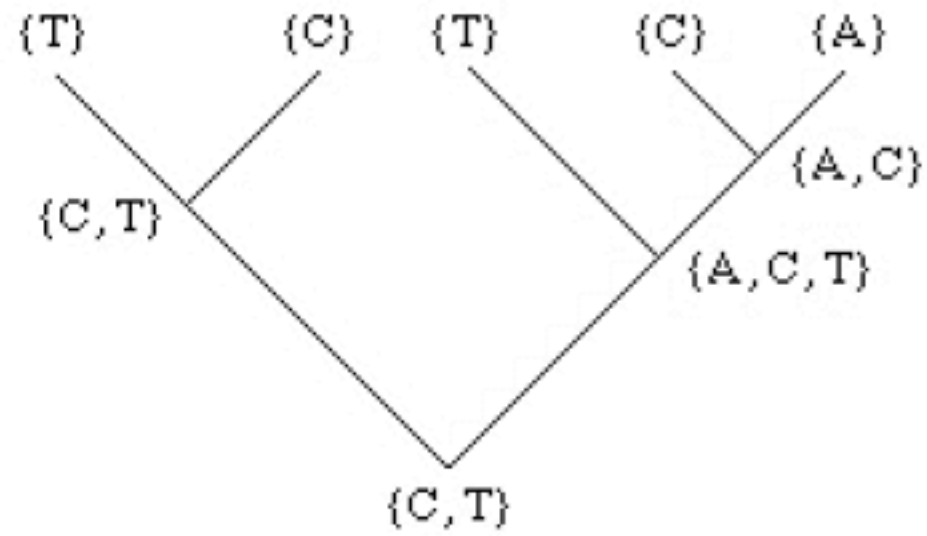
- For given example tree and alignment, can do this for all sites, and get away with as few as N changes
- Changing the tree (either the topology or labeling of leaves) changes the minimum number of changes need
- Two computational problems
  - (Easy) Given a particular tree, how do you find minimum number of changes need to explain data? (Fitch)
  - (Hard) How do you search through all trees?

Parsimony: Fitch's Algorithm

Ex:

```
    123456
A   GTCGTA
B   GTCACT
C   GCGGTA
D   ACGACA
E   ACGGAA
```

Site 5



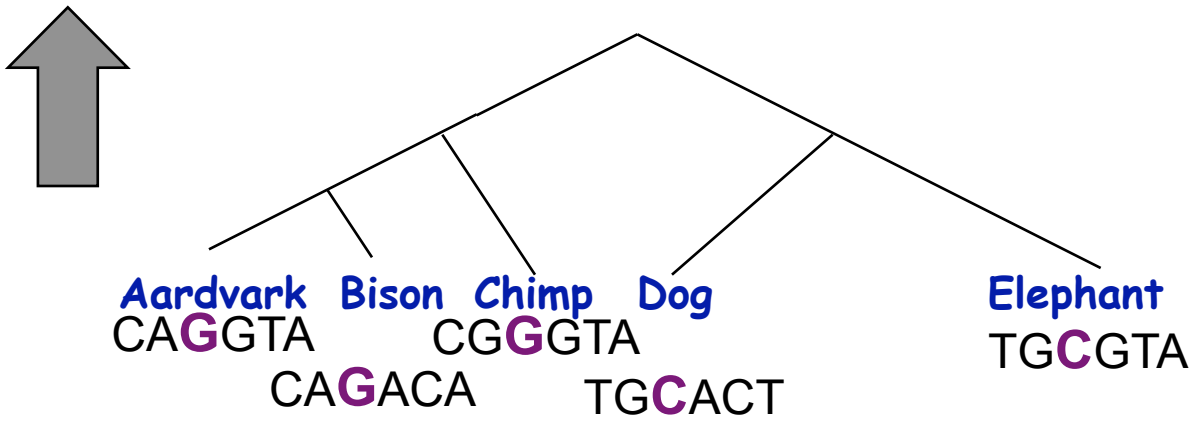
Idea: construct set of possible nucleotides for internal nodes, based on possible assignments of children

Fitch's Algorithm

Execute independently for each character:

Bottom-up technique: Determine set of possible states for each internal node

Dynamic Programming framework



Parsimony: Fitch's Algorithm

- For each site:
  - Each leaf is labeled with set containing observed nucleotide at that position
  - For each internal node i with children j and k with labels S_j and S_k

$$S_i = \begin{cases} S_j \cup S_k \text{ if } S_j \cap S_k \text{ is empty} \\ S_j \cap S_k \text{ otherwise} \end{cases}$$

- Total # changes necessary for a site is # of union operations

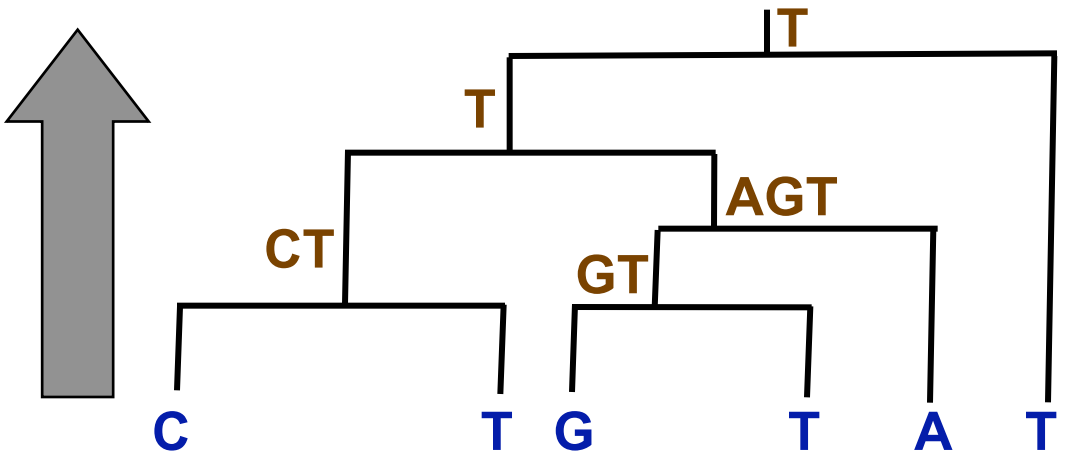
Fitch's Algorithm

Bottom-up phase

Determine set of possible states for each internal node

- Initialization: $R_i = \{s_i\}$
- Do a post-order (from leaves to root) traversal of tree
  - Determine R_i of internal node i with children j, k:

$$R_i = \begin{cases} R_j \cap R_k & if \;\; R_j \cap R_k \neq \phi \\ R_j \cup R_k & otherwise \end{cases}$$



Parsimony-score = # union operations

score = 3

Exploring the Space of Trees

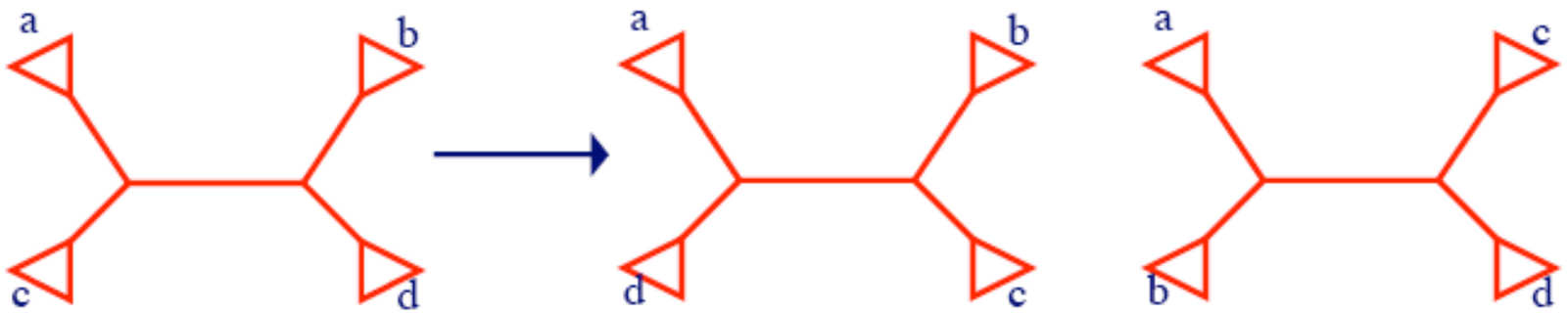
- We saw how to find optimal state-assignment for a given tree topology
- We need to explore space of topologies
- Given ***n*** sequences there are **(2*n*-3)!!** possible rooted trees
  and **(2*n*-5)!!** possible unrooted trees

$$n!! = 1 \cdot 3 \cdot 5 \cdot \ldots \cdot n \approx 2^{n/2} \cdot \left(\frac{n}{2}\right)!$$

taxa (n)	# rooted trees	# unrooted trees
3	3	1
4	15	3
5	105	15
6	945	105
8	135,135	10,395
10	34,459,425	2,027,025

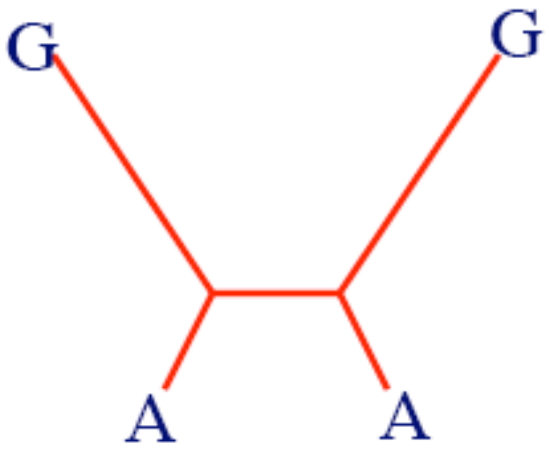
Parsimony

- How do you search through all trees?
  - Enumerate all trees (too many…)
  - Can use techniques to try to limit the search space (e.g., branch and bound)
  - or use heuristics (many possibilities)
    - E.g., nearest neighbor interchange. Start with a tree and consider neighboring trees. If any neighboring tree has fewer changes, take it as current tree. Stop when no improvements

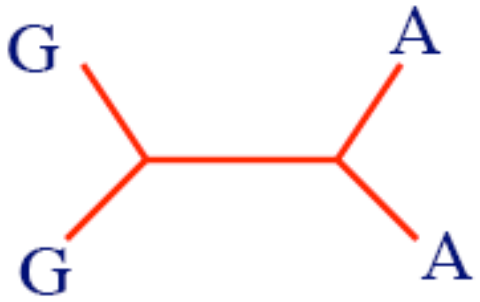


Parsimony Weakness

Parsimony analysis implicitly assumes that rate of change along branches are similar



Real tree: two long branches where G has turned to A independently



Inferred tree

Computing Distances between two sequences

Hamming distance – rudimentary method

Could compute fraction of mismatches between two sequences; however, this is an underestimate of actual distance

A simple clustering method for building a ROOTED tree

UPGMA (Unweighted Pair Group Method using Arithmetic averages)

Or the **Average Linkage Method**

Given two disjoint clusters S_i, S_j of sequences,

$$d_{ij} = \frac{1}{|S_i| \times |S_j|} \Sigma_{\{p \in Si,\ q \in Sj\}} d_{pq}$$

Claim that if $S_k = S_i \cup S_j$, then distance to another cluster S_l is:

$$d_{kl} = \frac{d_{il}\,|S_i| + d_{jl}\,|S_j|}{|S_i| + |S_j|}$$

Algorithm: Average Linkage

Initialization:

Assign each x_i into its own cluster S_i

Define one leaf per sequence, height 0

Iteration:

Find two clusters S_i, S_j s.t. d_{ij} is min

Let $S_k = S_i \cup S_j$

Define node connecting S_i, S_j,

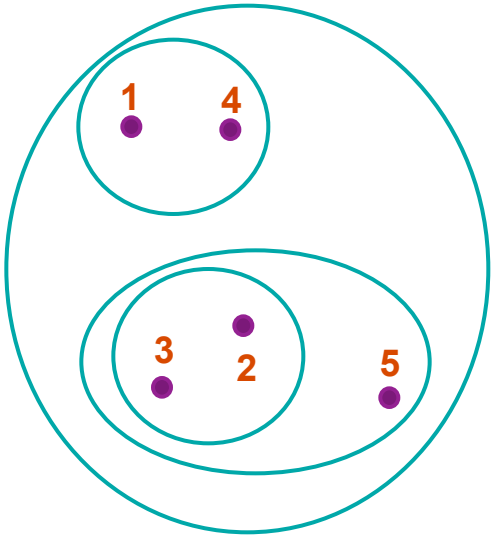
& place it at height $d_{ij}/2$

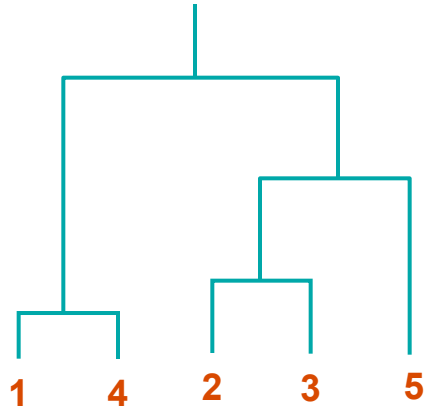
Delete S_i, S_j

Termination:

When two clusters i, j remain,

place root at height $d_{ij}/2$





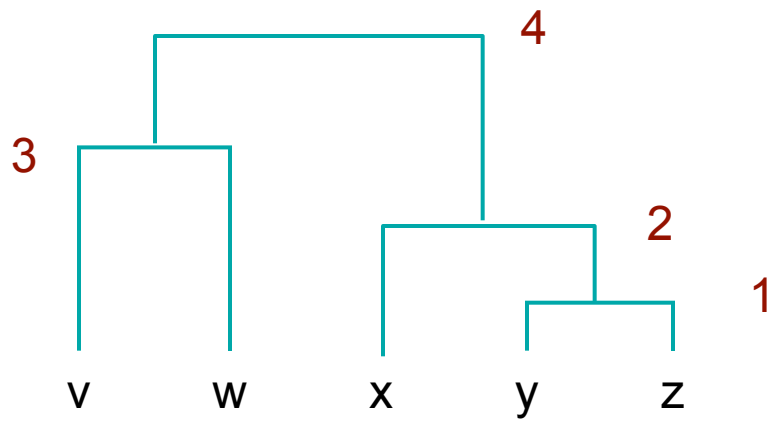
Example

	v	w	x	y	z
v	0	6	8	8	8
w		0	8	8	8
x			0	4	4
y				0	2
z					0

	v	w	xyz
v	0	6	8
w		0	8
xyz			0

	vw	xyz
vw	0	8
xyz		0

	v	w	x	yz
v	0	6	8	8
w		0	8	8
x			0	4
yz				0

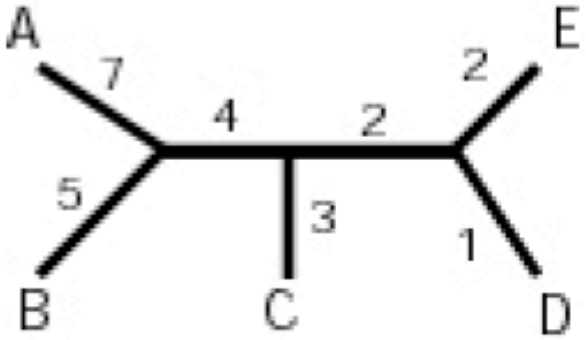


Distance Methods (Unrooted)

- Input: given an n x n matrix M where $M_{ij}>=0$ and M_{ij} is the distance between objects i and j
- Goal: Build an edge-weighted tree where each leaf (external node) corresponds to one object of M and so that distances measured on the tree between leaves i and j correspond to M_{ij}

Distance Methods

	A	B	C	D	E
A	0				
B	12	0			
C	14	12	0		
D	14	12	6	0	
E	15	13	7	3	0



A tree exactly fitting the matrix does not always exist.

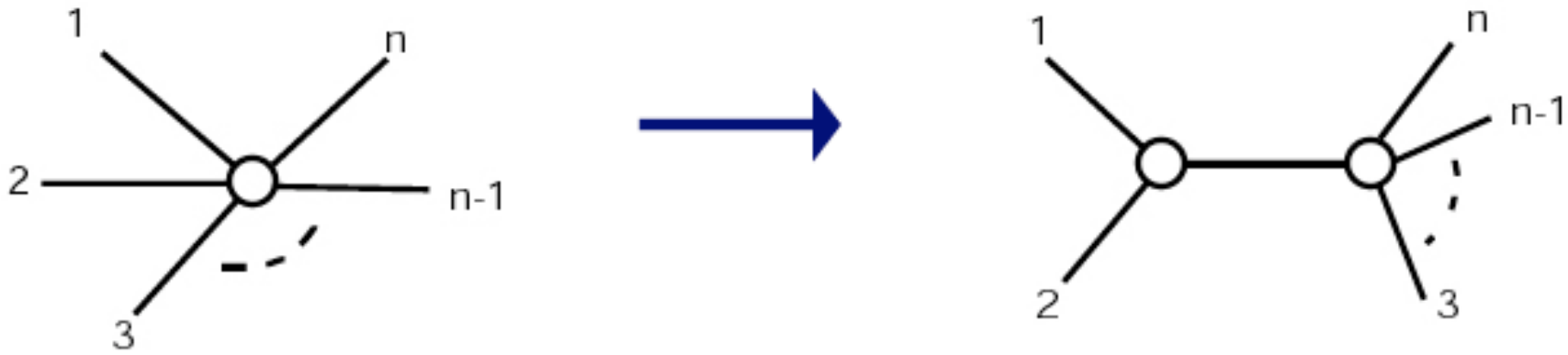
Distance Method Criteria

- Try to find the tree with distances d_{ij} which "best fits" the distance data M_{ij}
- Different possibilities for "best"
  - Cavalli-Sforza criterion: minimize
    $$\sum_{i,j}(M_{ij} - d_{ij})^2$$
  - Fitch-Margoliash criterion: minimize
    $$\sum_{i,j} \frac{(M_{ij}-d_{ij})^2}{M_{ij}^2}$$
- Unfortunately, both lead to computationally intractable problems (e.g., enumerating)

Distance Method: Neighbor Joining Method

- Most widely-used distance based method for phylogenetic reconstruction
- UPGMA illustrated that it is not enough to just pick closest neighbors
- Idea here: take into account averaged distances to other leaves as well
- Produces an unrooted tree

Neighbor Joining Method



Start off with star tree; pull out pairs at a time

NJ Algorithm

Step 1: Let $u_i = \sum_k \frac{M_{ik}}{n-2}$

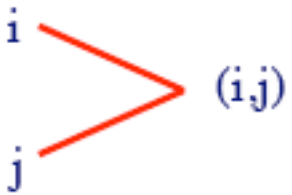
- (Almost) "average" distance to other nodes

Step 2: Choose i and j for which $M_{ij} - u_i - u_j$ is smallest

- Look for nodes that are close to each other, and far from everything else
- Turns out minimizing this is minimizing sum of branch lengths

NJ Algorithm

Step 3: Define a new cluster (i, j), with a corresponding node in the tree



Distance from i and j to node (i,j):

$$d_{i,(i,j)} = 0.5(M_{ij} + u_i - u_j)$$

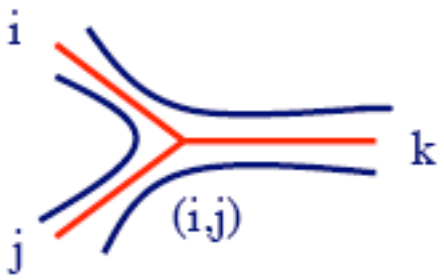
$$d_{j,(i,j)} = 0.5(M_{ij} + u_j - u_i)$$

Default: split distance but if on average one is further away, make it longer

NJ Algorithm

Step 4: Compute distance between new cluster and all other clusters:

$$M_{(ij)k} = \frac{M_{ik} + M_{jk} - M_{ij}}{2}$$



Step 5: Delete i and j from matrix and replace by (i, j)

Step 6: Continue until only 2 leaves remain

NJ Algorithm

- Works well in practice
- If there is a tree that fits the matrix, it will find it
- Can sometimes get trees with negative length edges (!)

Ultrametric Distances and Molecular Clock

Definition:

A distance function d(.,.) is ultrametric if for any three distances $d_{ij} \le d_{ik} \le d_{jk}$, it is true that

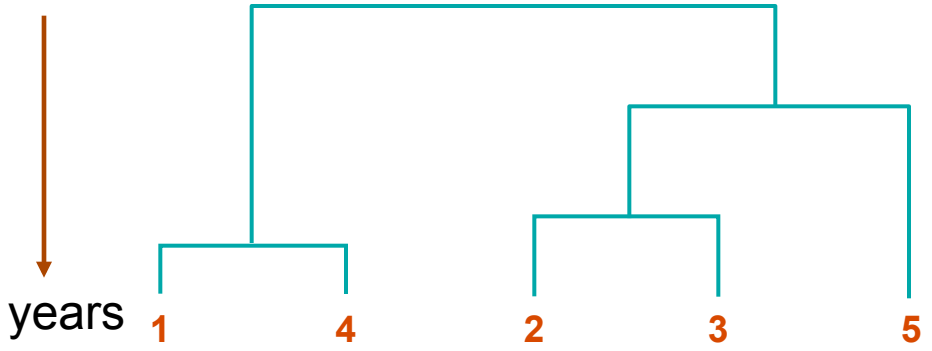
$$d_{ij} \le \max(d_{ik}, d_{jk})$$

Different from Triangle Inequality: $d_{ij} \le |d_{ik}|+|d_{jk}|$

The Molecular Clock:

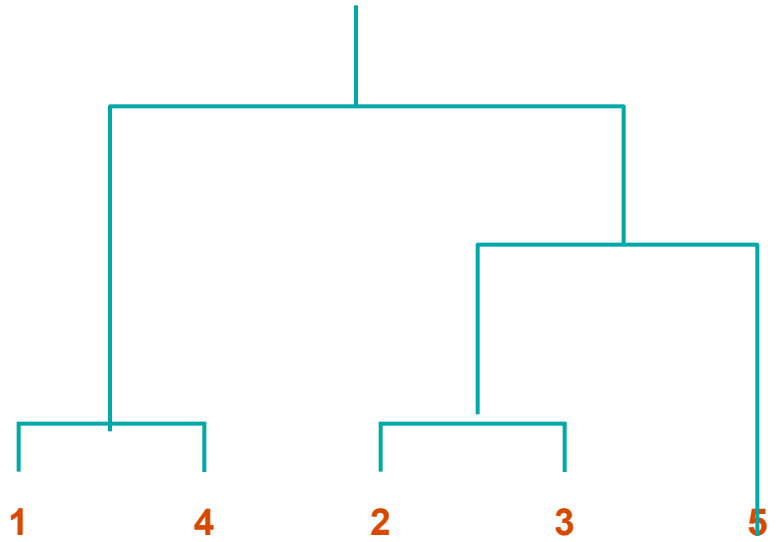
The evolutionary distance between species x and y is 2× the Earth time to reach the nearest common ancestor

That is, the molecular clock has constant rate in all species



The molecular clock results in ultrametric distances

Ultrametric Distances & Average Linkage



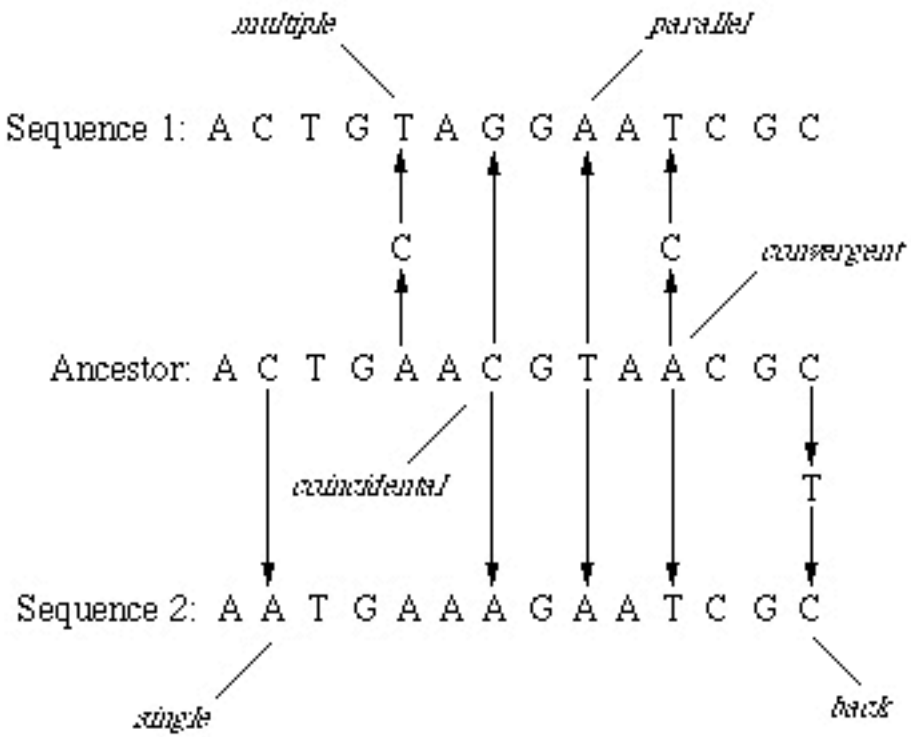
Average Linkage (e.g. UPGMA) is guaranteed to reconstruct correctly a binary tree with ultrametric distances

Computing Distances Between Sequences

("Undersampling" of sequences for long times)

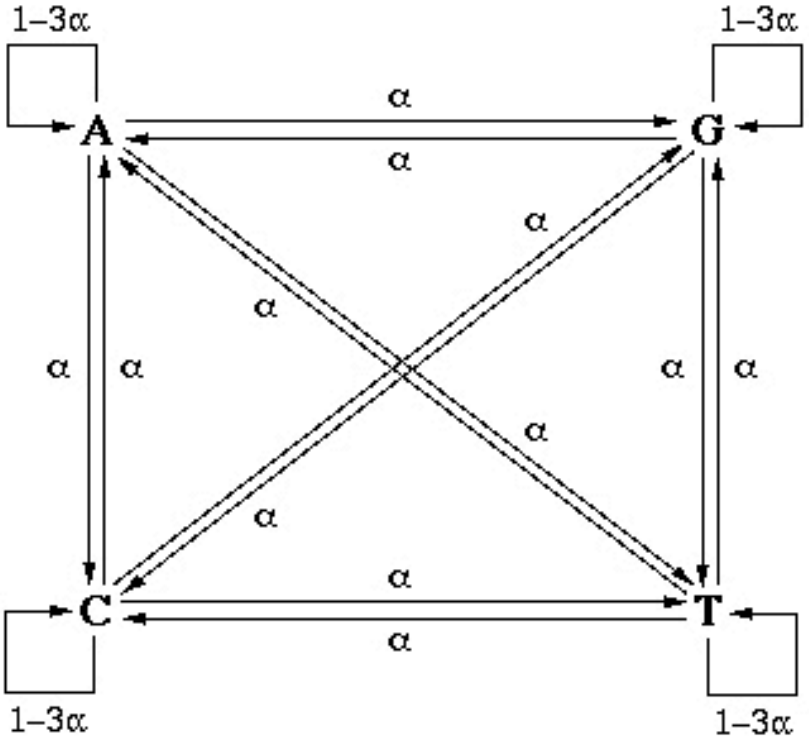
E.g., many underlying substitutions possible

Use models of substitution to correct these values



How estimate true number of substitutions?

Computing Distances Between Sequences



Jukes & Cantor model
- -Each position in DNA sequence is independent
- -Each position can mutates with same probability to any another base

Correction to observed substitution rate (see notes):

$$-0.75\left(\ln\left(1-\frac{4}{3}\left(\frac{\text{observed \# differences}}{\text{length}}\right)\right)\right)$$

Derivation in Cristianini and Hahn

Transition Probabilities

$$M = \begin{pmatrix} 1-3\alpha & \alpha & \alpha & \alpha \\ \alpha & 1-3\alpha & \alpha & \alpha \\ \alpha & \alpha & 1-3\alpha & \alpha \\ \alpha & \alpha & \alpha & 1-3\alpha \end{pmatrix}$$

$$M(t) = M^t = \sum \lambda_i^t v_i v_i^T =$$

$$1 + \sum (1-4/3\alpha)^t \bullet 1/4(\text{eigenvectors})$$

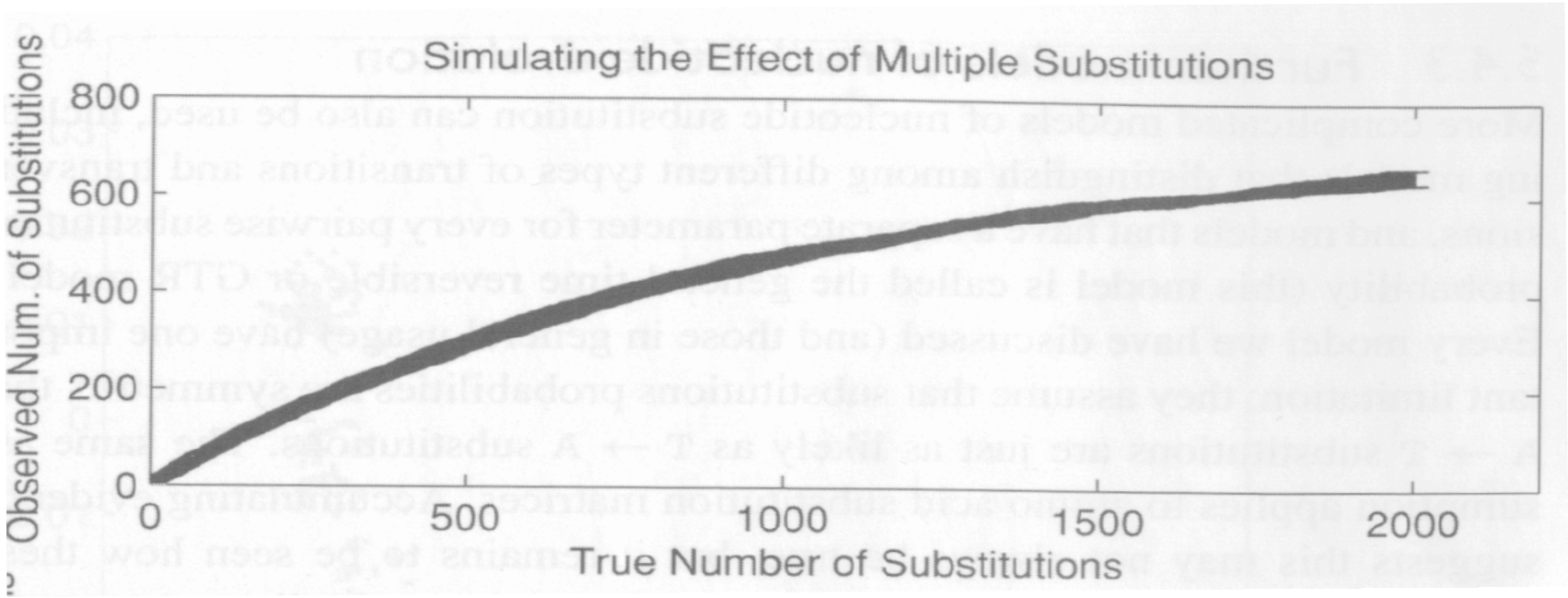
$$M(t) = \begin{pmatrix} 1-3\alpha & \alpha & \alpha & \alpha \\ \alpha & 1-3\alpha & \alpha & \alpha \\ \alpha & \alpha & 1-3\alpha & \alpha \\ \alpha & \alpha & \alpha & 1-3\alpha \end{pmatrix}^t$$

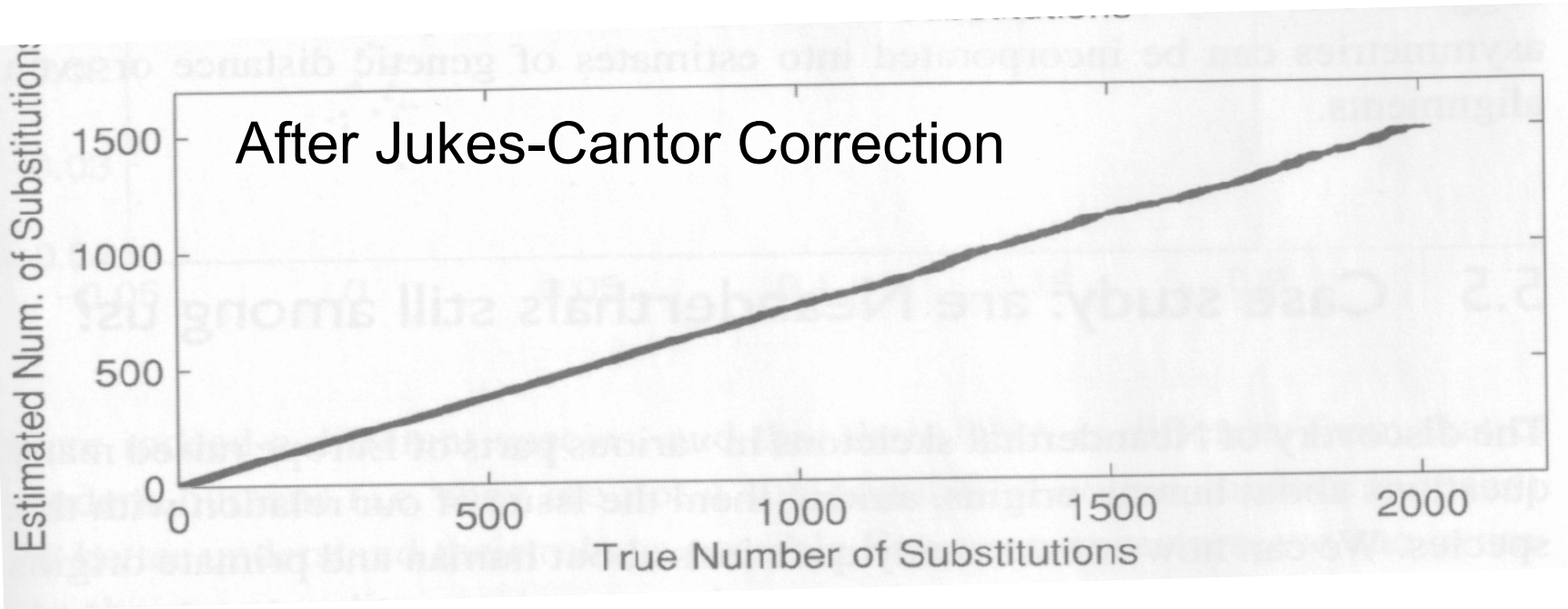
$$s(t) = 1/4 - 1/4(1 - 4/3\alpha)^t$$

Pr of substitution observed

$$\begin{pmatrix} r(t) & s(t) & s(t) & s(t) \\ s(t) & r(t) & s(t) & s(t) \\ s(t) & s(t) & r(t) & s(t) \\ s(t) & s(t) & s(t) & r(t) \end{pmatrix}$$

Lots of approximations:
$t \sim -3/(4\alpha) \ln(1 - 4/3\ d)$
d: differences btw two sequences
# of subs = $t * \alpha$


Simulating the Effect of Multiple Substitutions
Observed Num. of Substitutions
800
600
400
200
0
0
500
1000
1500
2000
True Number of Substitutions


After Jukes-Cantor Correction
Estimated Num. of Substitutions
1500
1000
500
0
0
500
1000
1500
2000
True Number of Substitutions

Computing Distances Between Sequences

- Alignment of two DNA sequences
  - Length of alignment (non gapped positions): 100
  - Number of differences: 25
- Naïve distance calculation = 25/100 = ¼
- Correction

$$-0.75(\ln(1 - \tfrac{4}{3}(1/4)) = -.75\ln(\tfrac{2}{3}) = .304$$

- Other models for DNA, also protein (e.g., PAM)

Maximum Likelihood

- Given a probabilistic model for nucleotide (or protein) substitution (e.g., Jukes & Cantor), pick the tree that has highest probability of generating observed data
  - I.e., Given data D and model M, find tree T such that $Pr(D|T, M)$ is maximized
- Models gives values $p_{ij}(t)$, the probability of going from nucleotide i to j in time t

Maximum Likelihood

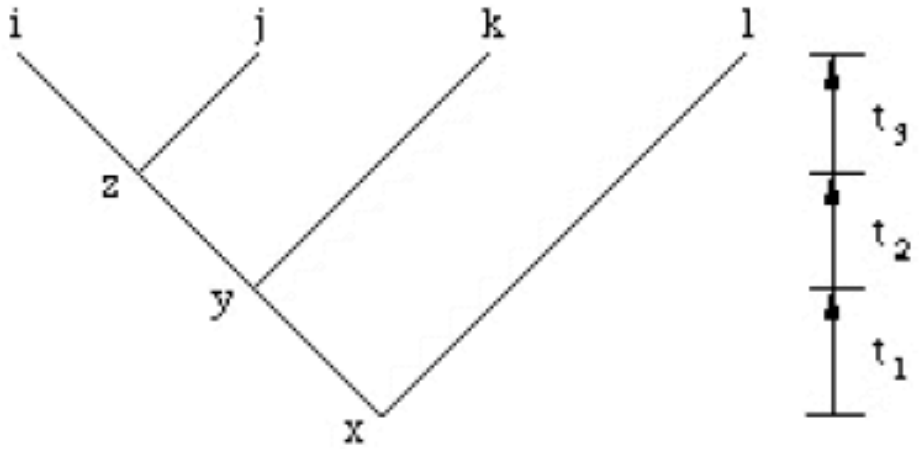
- Makes 2 independence assumptions
  - Different sites evolve independently
  - Diverged sequences (or species) evolve independently after diverging
- If D_i is data for ith site

$$Pr(D|T, M) = \prod_i Pr(D_i|T, M)$$

Maximum Likelihood

How to calculate $Pr(D_i|T,M)$?

$p_{xy}(t) \sim$ prob of going from x to y in time t



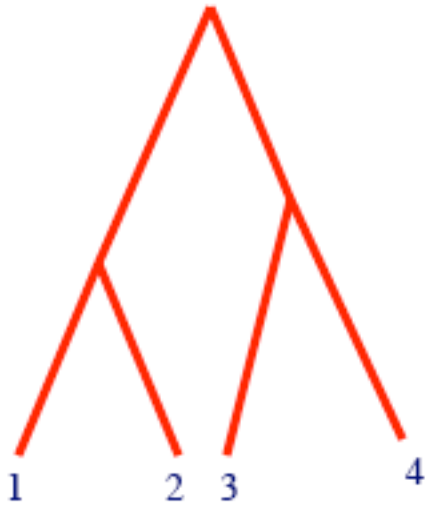
$$Pr(i,j,k,l|T,M) =$$
$$\sum_x \sum_y \sum_z pr(x)(p_{xl} \cdot (t_1 + t_2 + t_3) \cdot p_{xy}(t_1)$$
$$\cdot p_{yk}(t_2 + t_3) \cdot p_{yz}(t_2) \cdot p_{zi}(t_3) \cdot p_{zj}(t_3))$$

Maximum Likelihood

- Given tree topology and branch lengths, can efficiently calculate $Pr(D|T, M)$ using dynamic programming
  - I.e., don't have to enumerate over all internal states
- Finding best maximum likelihood tree is expensive
  - Must consider all topologies
  - Find best edge lengths for each topology
    - Idea: use some search procedure, e.g., EM, to optimize these lengths

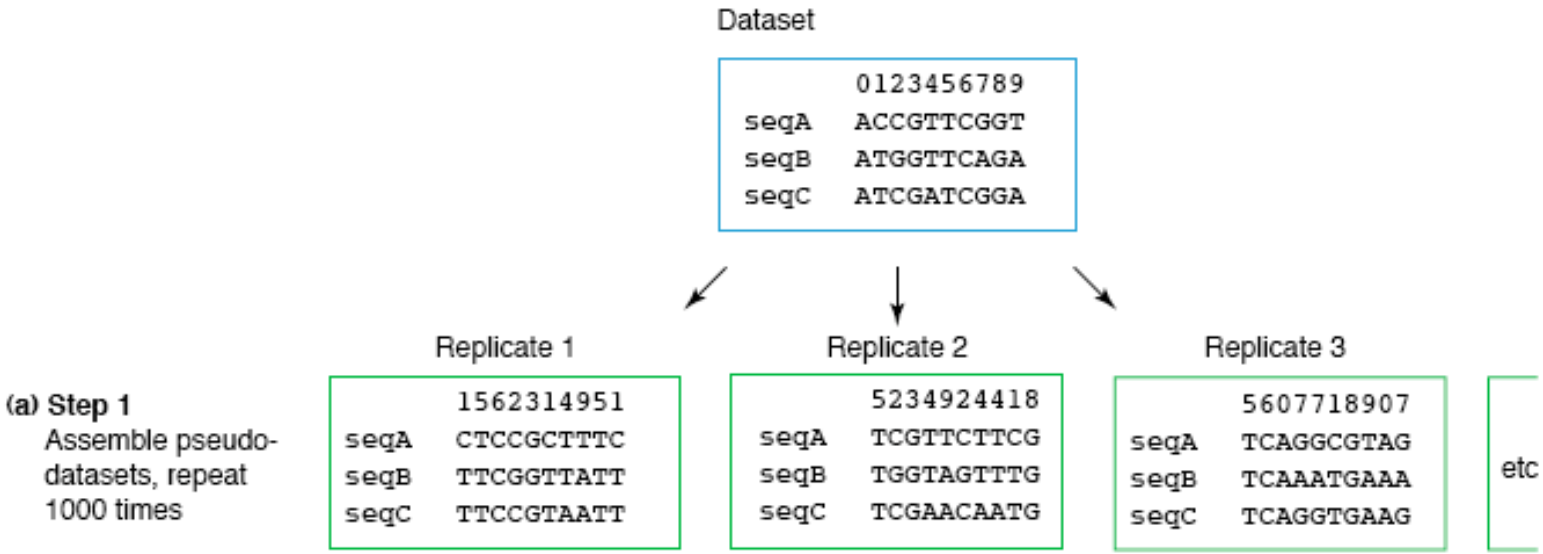
Assessing Reliability -- The Bootstrap

Say we've inferred the following tree

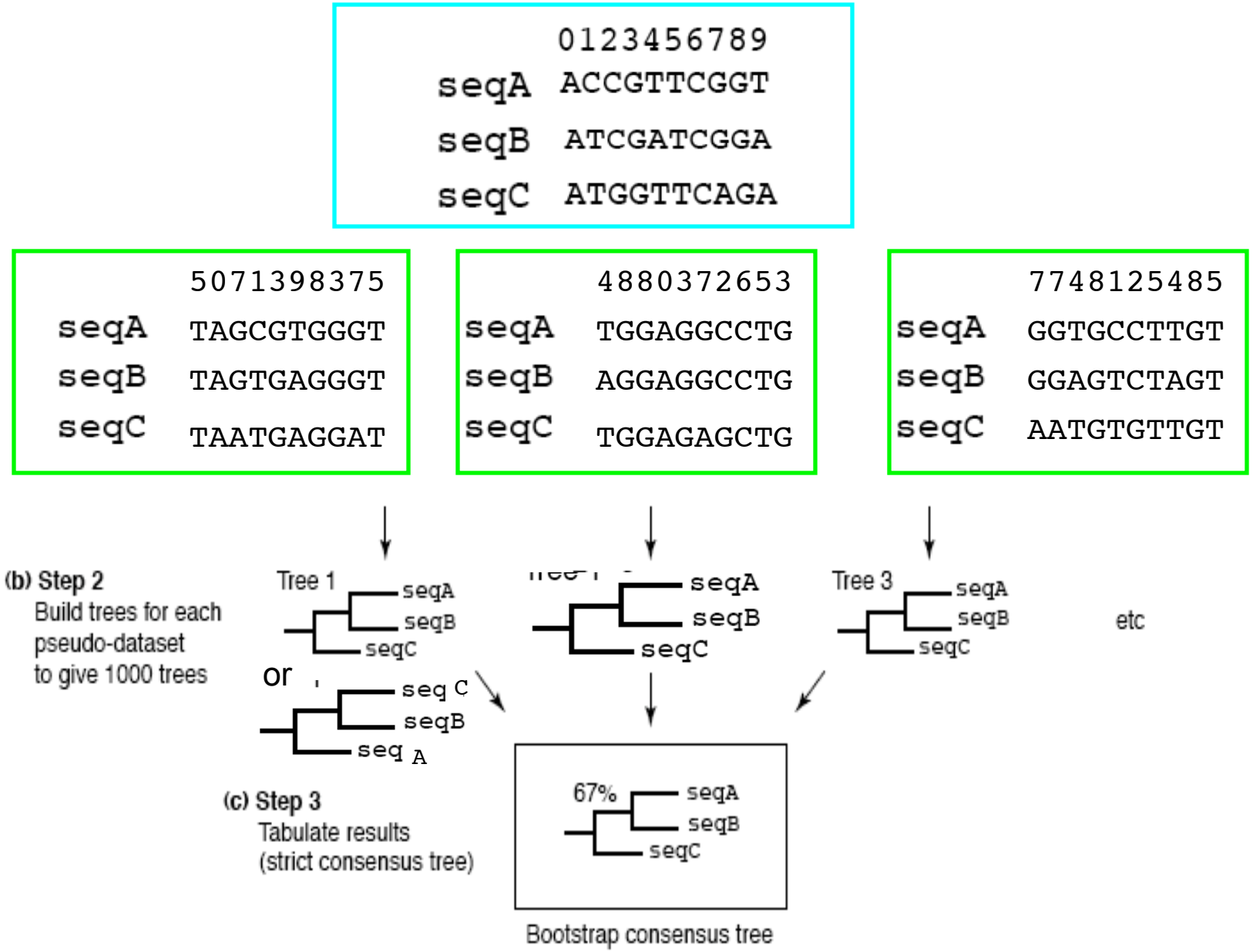


Would like to get confidence levels that 1 & 2 belong together, and 3&4 belong together

Bootstrapping



Bootstrapping



Assessing the Reliability - The Bootstrap

Say we're given following alignment:

```
   12345678
1  GCAGTACT
2  GTAGTACT
3  ACAATACC
4  ACAACACT
```

We'll create a pseudosample by choosing sites randomly until N sites are chosen (N is length of alignment)

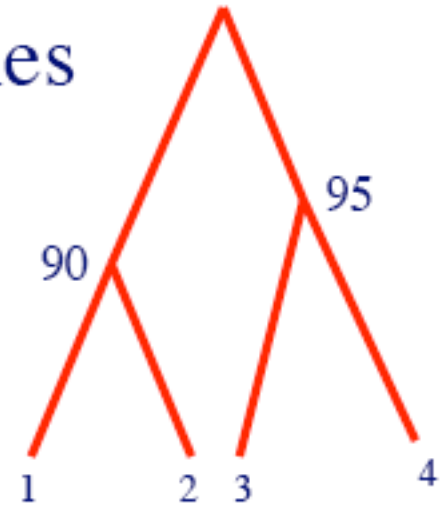
Assessing Reliability - The Bootstrap

Say chose 6^{th}, 1^{st}, 6^{th}, 8^{th}, …

	12345678		6168 …
1	GCAGTACT		AGAT …
2	GTAGTACT	⟶	AGAT …
3	ACAATACC		AAAC …
4	ACAACACT		AAAT …

Assessing the Bootstrap

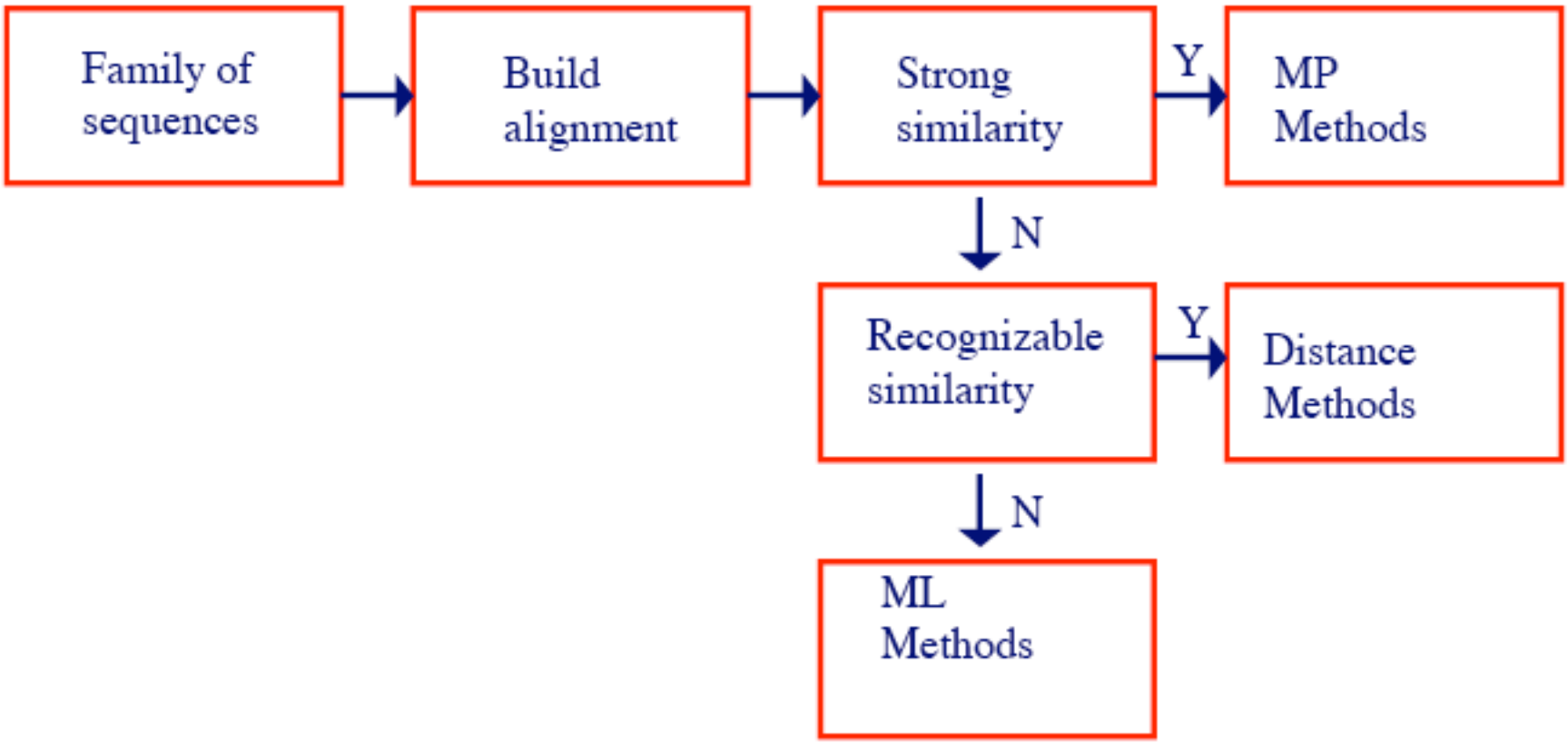
- Use pseudosample to construct tree
- Repeat many times
- Confidence of (1) and (2) together is fraction of times they appear together in trees generated from pseudosamples



Difference in Methods

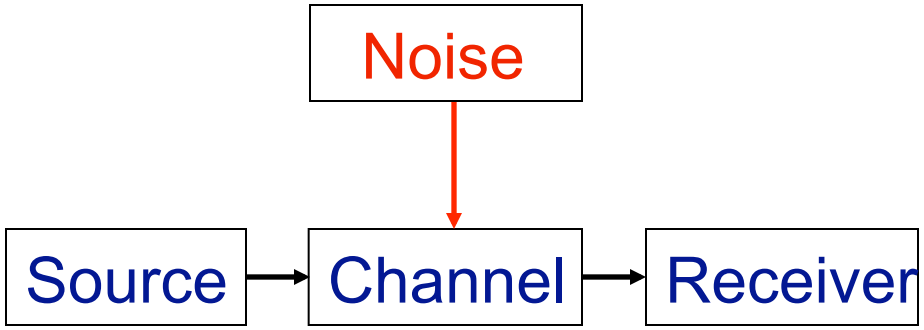
- Maximum-likelihood and parsimony methods have models of evolution
- Distance methods do not necessarily
  - Useful aspect in some circumstances
    - E.g., trees built based on whole genomes, presence or absence of genes
- Religious wars over which methods to use
  - Most people now believe ML based methods are best: most sensitive at large evolutionary distances – but also most time-consuming & depend on specific model of evolution used
- Most commonly used packages contain software for all three methods: may want to use more than 1 to have confidence in built tree

Phylogeny Flowchart

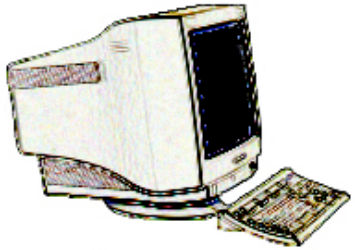


Mutations as Communication Errors

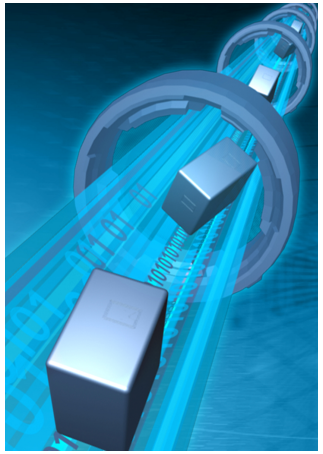
Communication Systems



Source



Channel



010010 (Alphabet)

Receiver

Entropy - Measure of Info Content and Redundancy

- Information content:

$p(m)$ - probability of a message

$$I(m) = -\log_2 p(m)$$

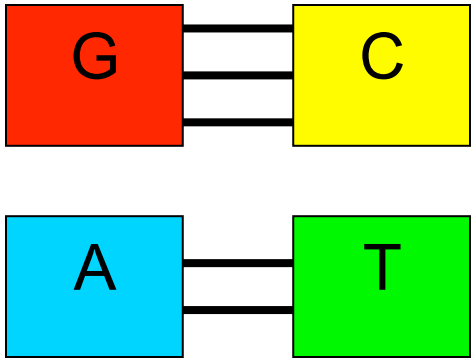
(places higher importance on unlikely message)

- Entropy measure (Average information content of all messages):

$$H = -\sum_{m \in M} p(m) \log_2 p(m)$$

DNA Entropy

DNA Maximum Entropy



$$H = -\sum_{i} p_i \log_2(p_i)$$

$$= -\sum_{i=1}^{4} \frac{1}{4} \log_2(\frac{1}{4}) = 2 \textit{bits}$$

Most organisms: 1.8->1.99 bits

Basic Info Theory

- $\log_2(M)$ - uncertainty
  - Examples for equally likely symbols
  - 1 possible symbol: $\log_2(1) = 0$ bits
  - 2 possible symbols: $\log_2(2) = 1$ bits
  - 4 possible symbols: $\log_2(4) = 2$ bits
- What happens when unequal?

$$\begin{aligned} \log_2(M) &= -\log_2(M^{-1}) \\ &= -\log_2\left(\frac{1}{M}\right) \end{aligned}$$

$1/M$ is the probability that any symbol appears

Entropy Example

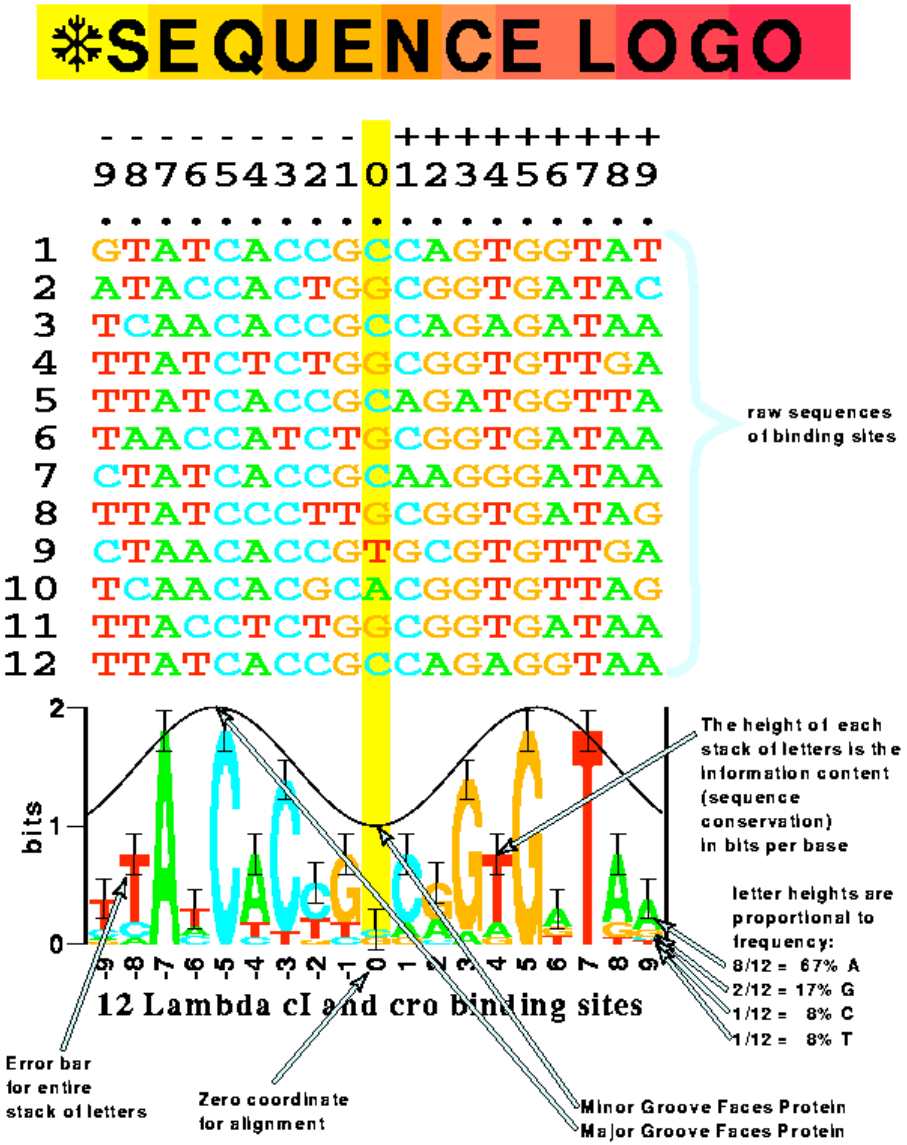
Micrococcus Lysodeikticus has the following base frequencies:

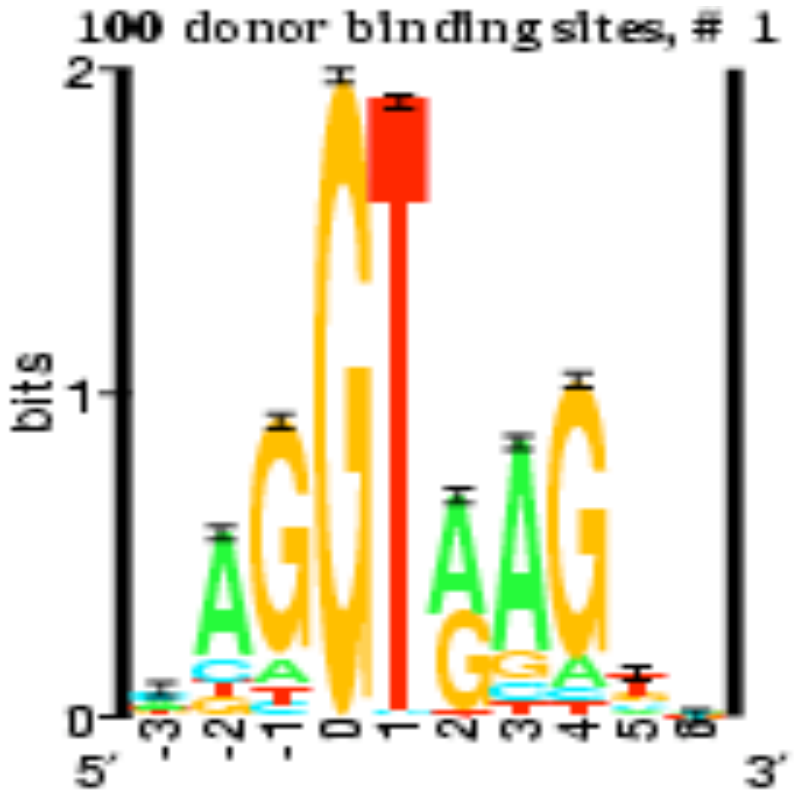
Pr(C)=Pr(G)=.355

Pr(A)=Pr(T)=.145

The entropy for this organism is 1.87 bits, which implies redundancy from this imbalance.

Example App: DNA Info Content





Tom Schneider: National Cancer Research Institute

Sequence Logos

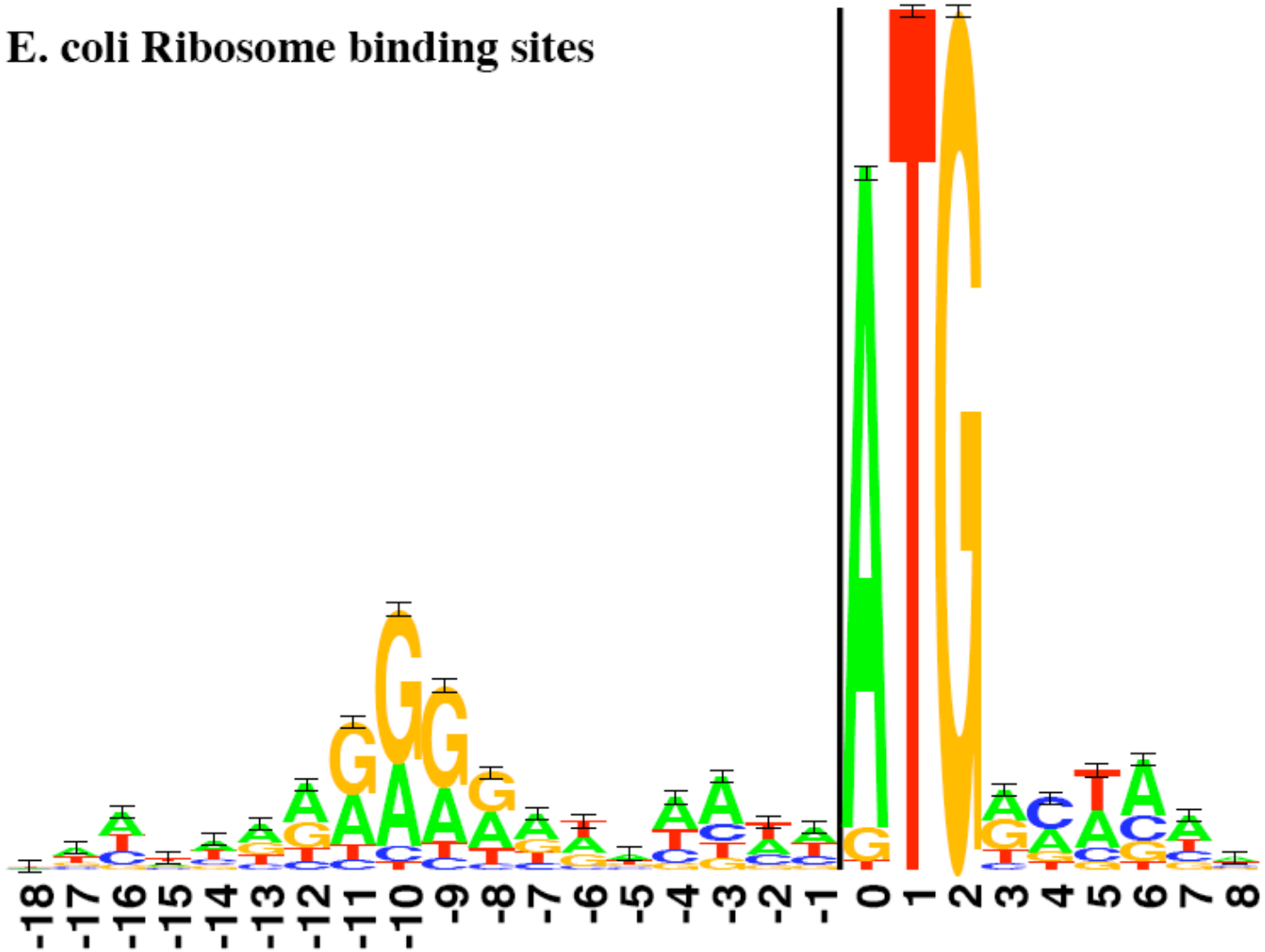
- $H(l) = -\sum_{b=a}^{t} \mathrm{p}(b,l) \log \mathrm{p}(b,l)$ (bits per position)
- $H(l)$ is the uncertainty at position l
- $\mathrm{p}(b,l)$ is the probability of base b at position l
- $R_{sequence}(l) = 2 - (H(l) + e(n))$

"low-data correction"

amount of information present at position l

- height of base b at position $l = \mathrm{p}(b,l) R_{sequence}(l)$

Ribosomal Binding Sites

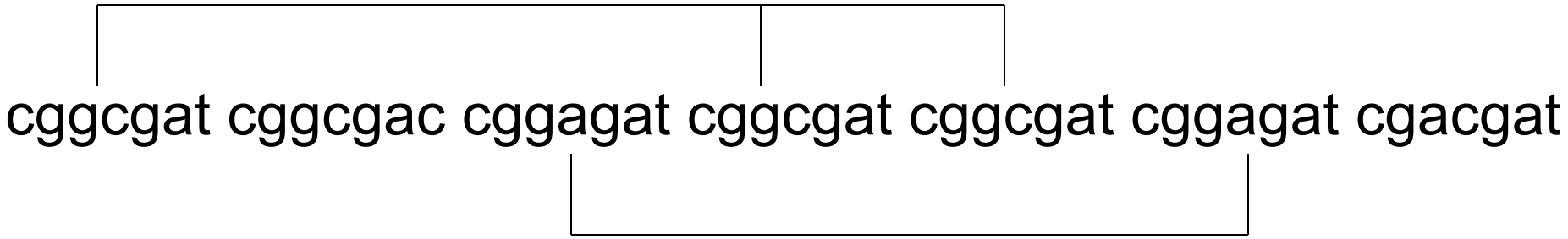


Classifications of Tandem Repeats

- Some Types:
  - Simple Sequence Repeat (Microsatellites): CTG CTG CTG
  - Variable Number Tandem Repeat (Minisatellites): CATG CATG CATGTG CATGTG CATG CATG
  - Multi-period Tandem Repeats: CAG CAT TAG CAT ACG CAT TAG CAT TTG CAT TAG

Minisatellites

Tandem Variant Repeats



- Unstable region in the human genome
- Genetic Markers (DNA fingerprinting/Evolutionary analysis)
- Regulator of Gene Expression?

Microsatellites (simple tandem)

Microsatellites (also known as SSR – Simple Sequence Repeats)

- Mononucleotide SSR (A)11
AAAAAAAAAAA
- Dinucleotide SSR (GT)6
GTGTGTGTGTGT
- Trinucleotide SSR (CTG)4
CTGCTGCTGCTG
- Tetranucleotide SSR (ACTC)4
ACTCACTCACTCACTC

Microsatellites

- Majority are in non-coding region
- Unstable region in the human genome
- Genetic Markers (DNA fingerprinting/ Evolutionary analysis)
- Associated with 18 neurodegenerative diseases

- Tandem repeats in genes -- 14 neurodegenerative disorders (Fragile X syndrome, Huntington's disease).

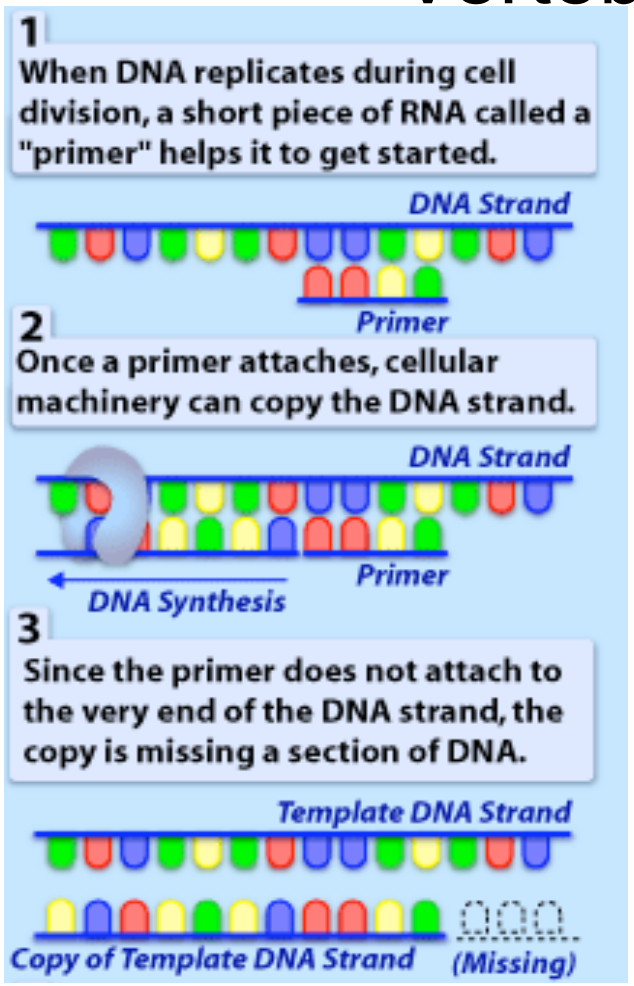
Genbank annotations

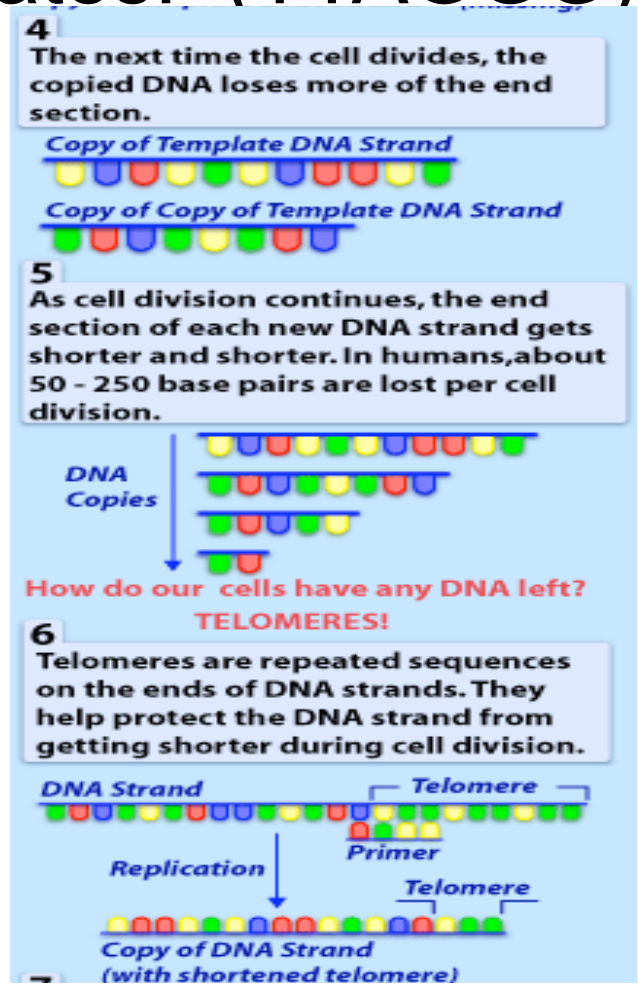
- Repeat_region
  - Repeat_unit
  - LTR (long terminal repeat)
- Satellite
  - minisatellite
  - microsatellite

Example: Telomeres

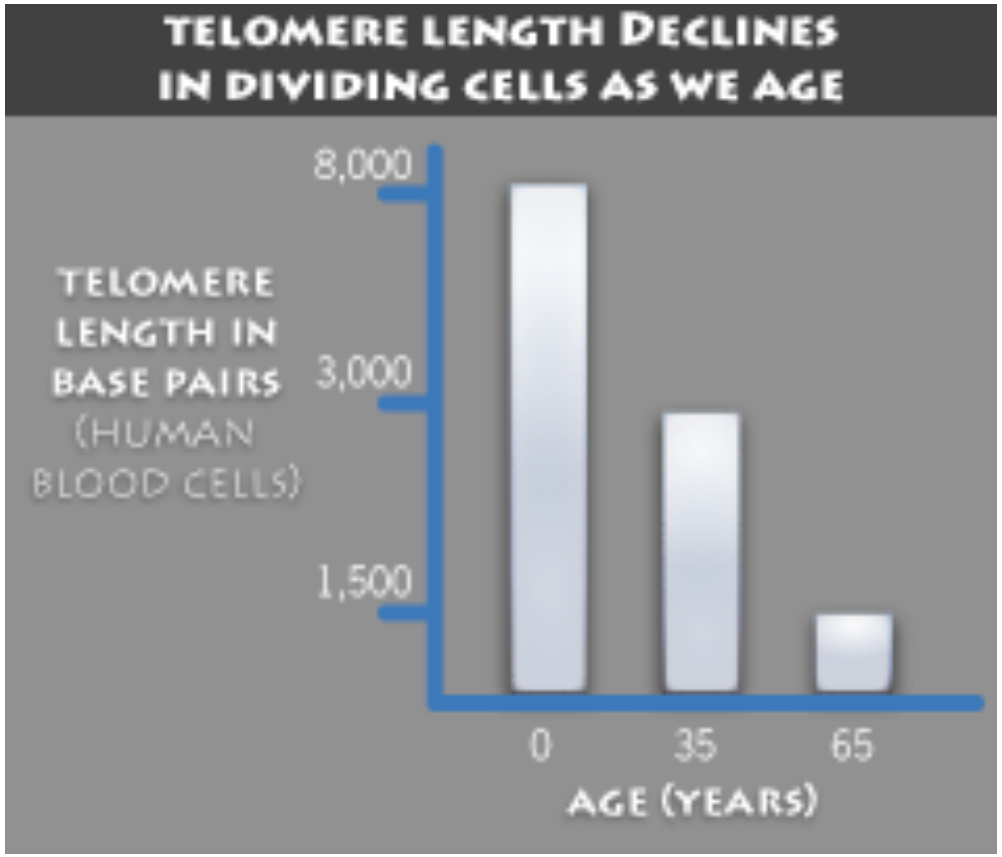
Telomeres
(eukaryotic)

Insects: (TTAGG)n
Plants: (TTTAGGG)n
Vertebrates: (TTAGGG)n





Telomeres and Aging



In germline cells (egg and sperm), an enzyme called telomerase is responsible for adding more repeat sequences to the end of the DNA, thus making them "immortal".

Germline cell DNA strand

Telomerase adding repeats

8

In somatic cells, the telomerase enzyme functions at much lower levels, making these cells "mortal".

DNA from old somatic cell (no telomere left)

9

When the telomeres in a somatic cell shorten to a critical level, that cell no longer divides. This phenomenon contributes to some of the changes we see in aging.

Genbank

- http://www.ncbi.nlm.nih.gov/BankIt/examples/repeat.html
- Relevant feature information for sequences containing repeat regions:
  * repeat region intervals
  * repeat family, if known (eg, Alu, Mer)
  * repeat type (tandem, inverted, flanking, terminal, direct, dispersed, or other)
  * repeat unit description/intervals, if region contains more than one repeat

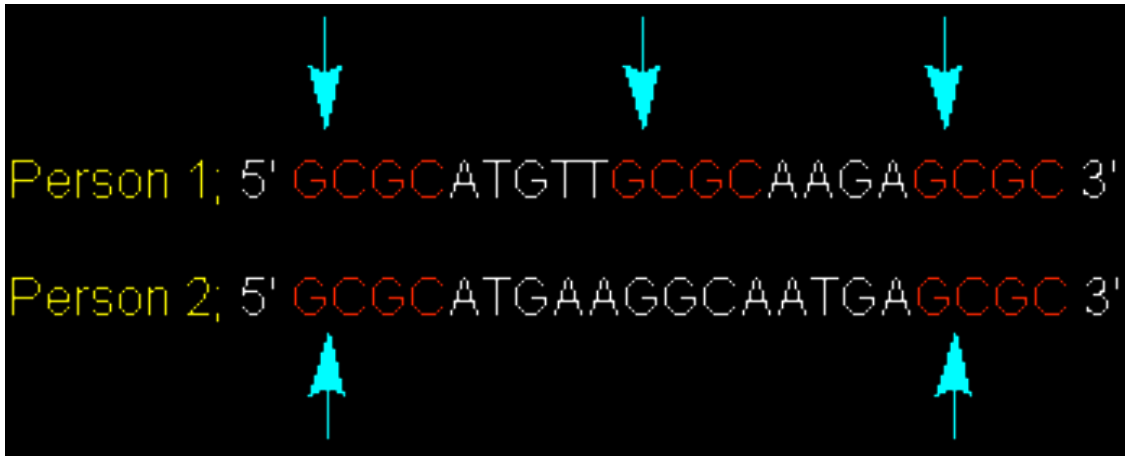
Genbank -- Repeat family

- http://www.genome.org/cgi/content/full/14/11/2245
- Alu repeats are the most abundant family of repeats in the human genome, with over 1 million copies comprising 10% of the genome.

Short Tandem Repeat Database (STRBase)

For Human Forensics: http://www.cstl.nist.gov/biotech/strbase/

Nice overview of DNA forensics typing: http://www.biotechnology.uwc.ac.za/teaching/Honours/Forensics/Topic_2%20.html



Major Tandem Repeat Detection Algorithms

- Alignment
- Direct via Similarity Measures
- Indirectly via Compression Methods

Tandem Repeat Detection (algorithmic compression method)

- Different algorithm to detect repeats
- Maximal run of k-mismatch tandem repeats, with period p:
  - –A maximal string such that any substring of length 2p is a tandem repeat with at most k mismatches

TRD: Periodic Transform (Buchner)

- Periodic subspace (example of 3), P_3:
  - Periodic subspace basis vectors:

    δ_3^0={1,0,0,1,0,0,1,0,0,1,0,0}

    δ_3^1={0,1,0,0,1,0,0,1,0,0,1,0}

    δ_3^2={0,0,1,0,0,1,0,0,1,0,0,1}
- Have sequence, *x*={3,2,1,3,2,1,3,2,2,4,2,1}
- Get contribution of nucleotides to each subspace:

$$\alpha_s = p\left\langle x, \delta_p^s \right\rangle$$

TRD: Periodic Transform (Buchner)

$$\begin{aligned} \alpha_1 &= 3.25 \\ \alpha_2 &= 2 \\ \alpha_3 &= 1.25 \end{aligned}$$

Periodic Projection:

$$\pi(x, P_p) = \sum_{s=0}^{p-1} \alpha_s \delta_p^s$$

$$\{3.25, 2, 1.25, 3.25, 2, 1.25, 3.25, 2, 1.25, 3.25, 2, 1.25\}$$

Non-fraction, perfect projection

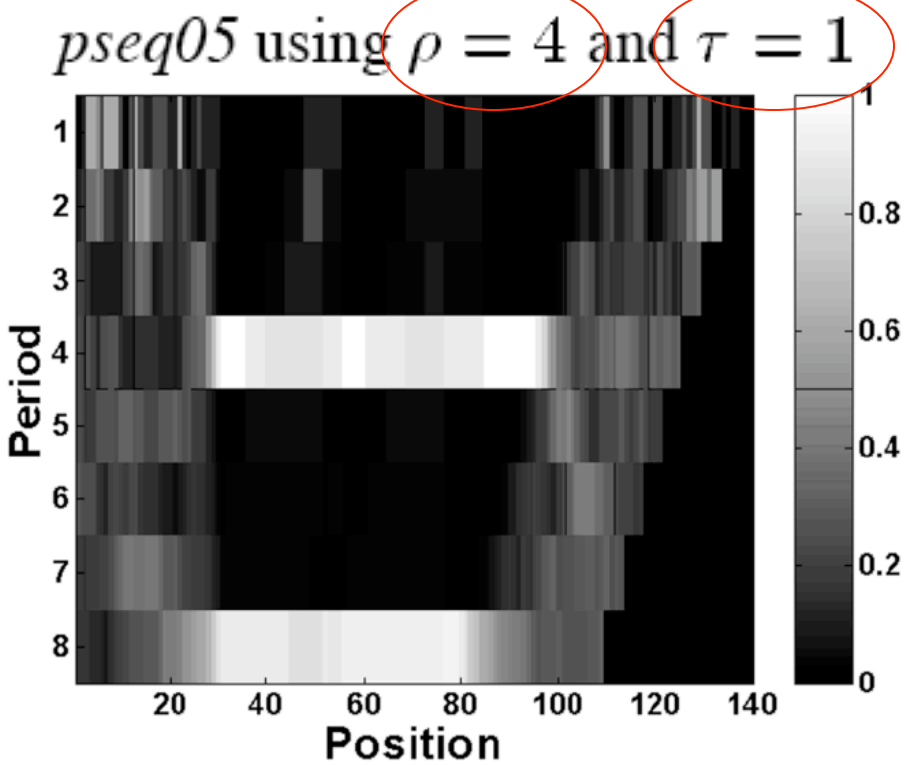
TRD: Periodogram (Buchner)

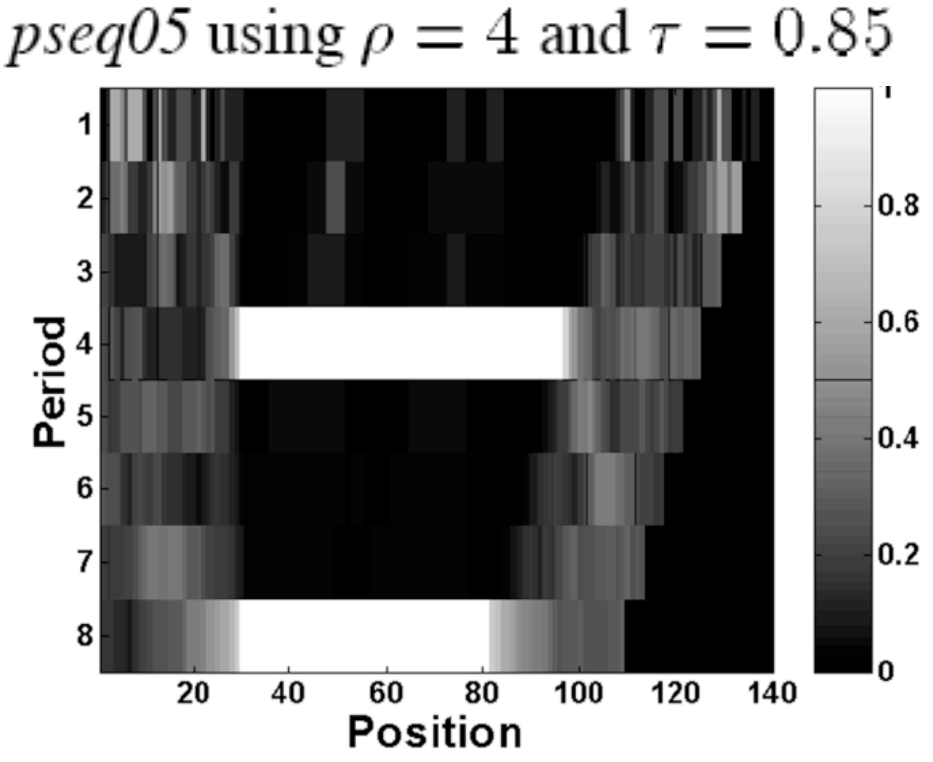
Periodogram coeff: Measure of localized sequence to closest periodic sequence:

$$\Pi[k,p] = \frac{\left\| \pi(x'_{w,k}, P_p) \right\|^2}{\left\| x_{w,k} \right\|^2}$$

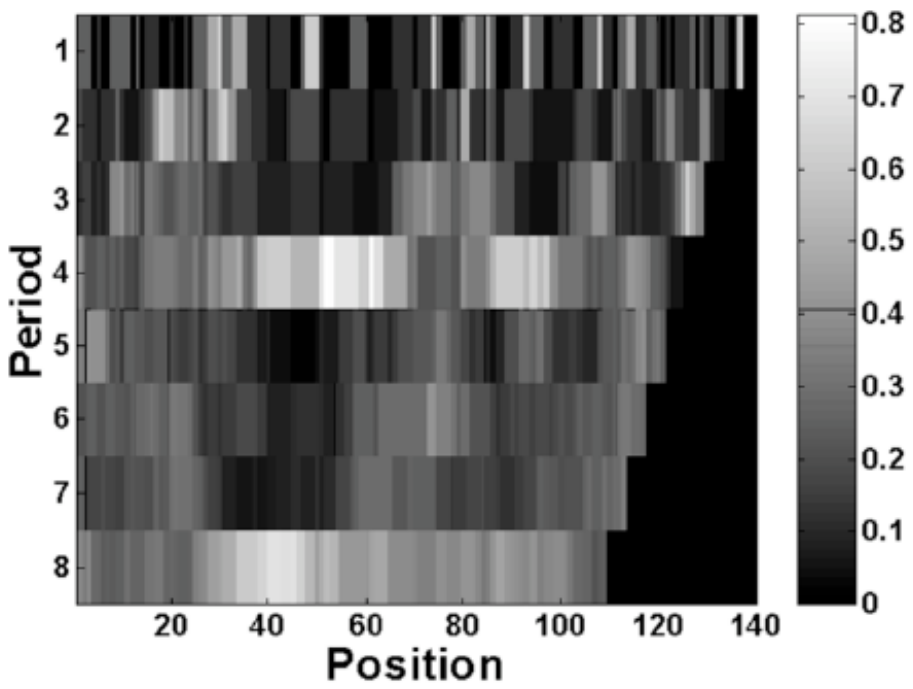
P-Transform order

Periodogram coeff threshold

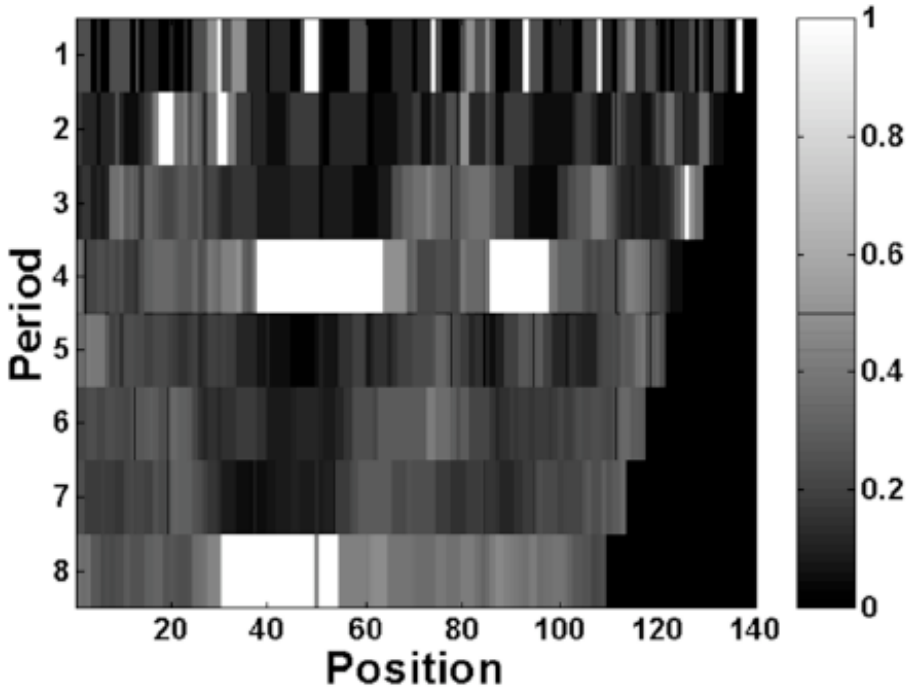




Periodogram continued

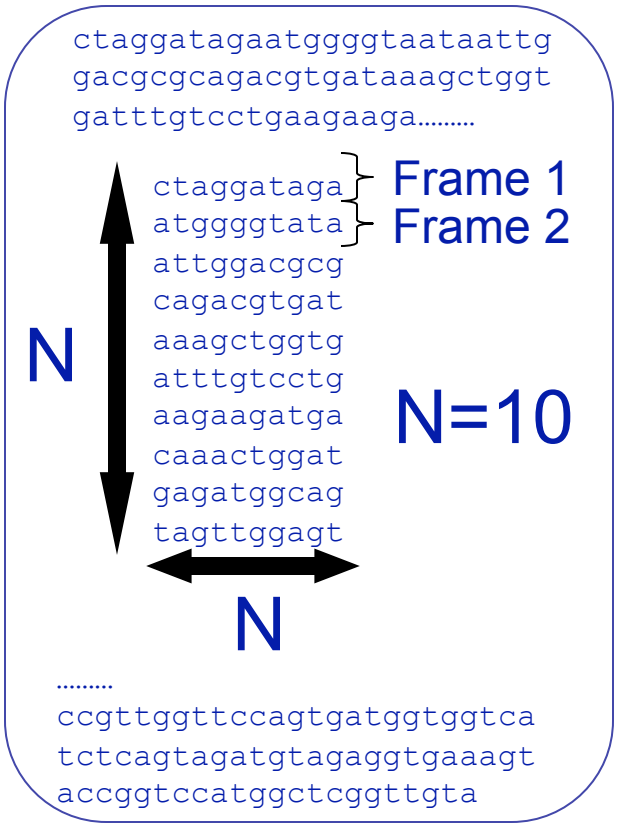


pseq25 using $\rho = 4$ and $\tau = 1.0$



pseq25 using $\rho = 4$ and $\tau = 0.50$

Rank-Deficiency Test for TRDs (Rosen)



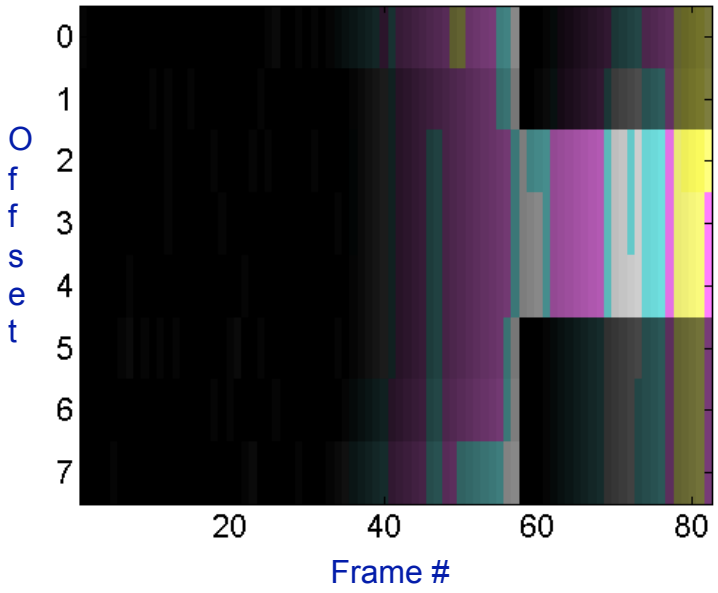
- For analysis frame length, N, collect N consecutive vectors to form NxN window.
- Perform a rank computation of the NxN matrix.
  - Based on Gaussian-elimination, modified for GF(4) arithmetic.
- Increment by one frame for each iteration.
- Note consistent rank-deficiency.

Advantages: Will find localized redundancy within a sequence.

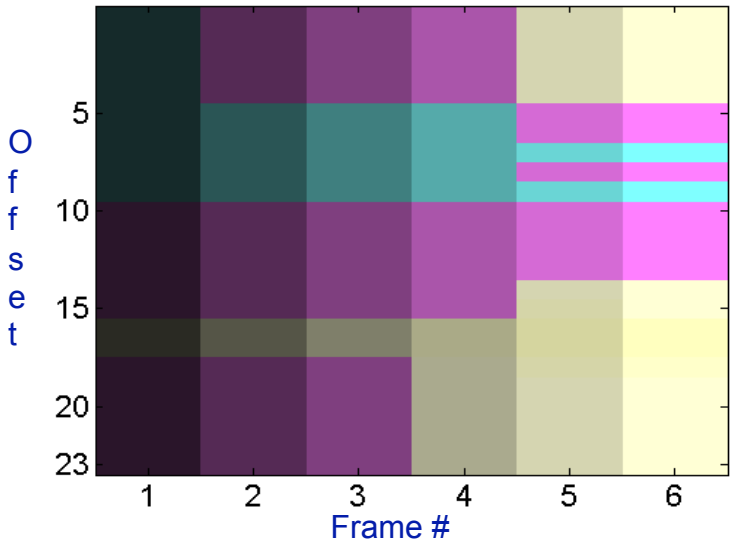
Disadvantages: Does not indicate a consistent linear subspace.

Rank-Deficiency Results (BOVTGN)

N=8



N=24

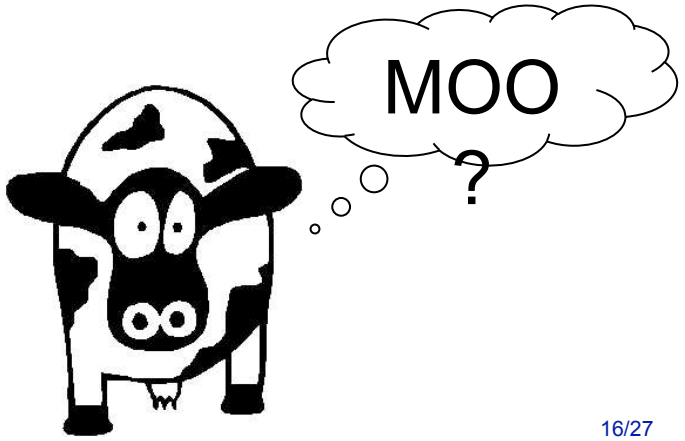


White: N-1 subspace

Blue: N-2 subspace

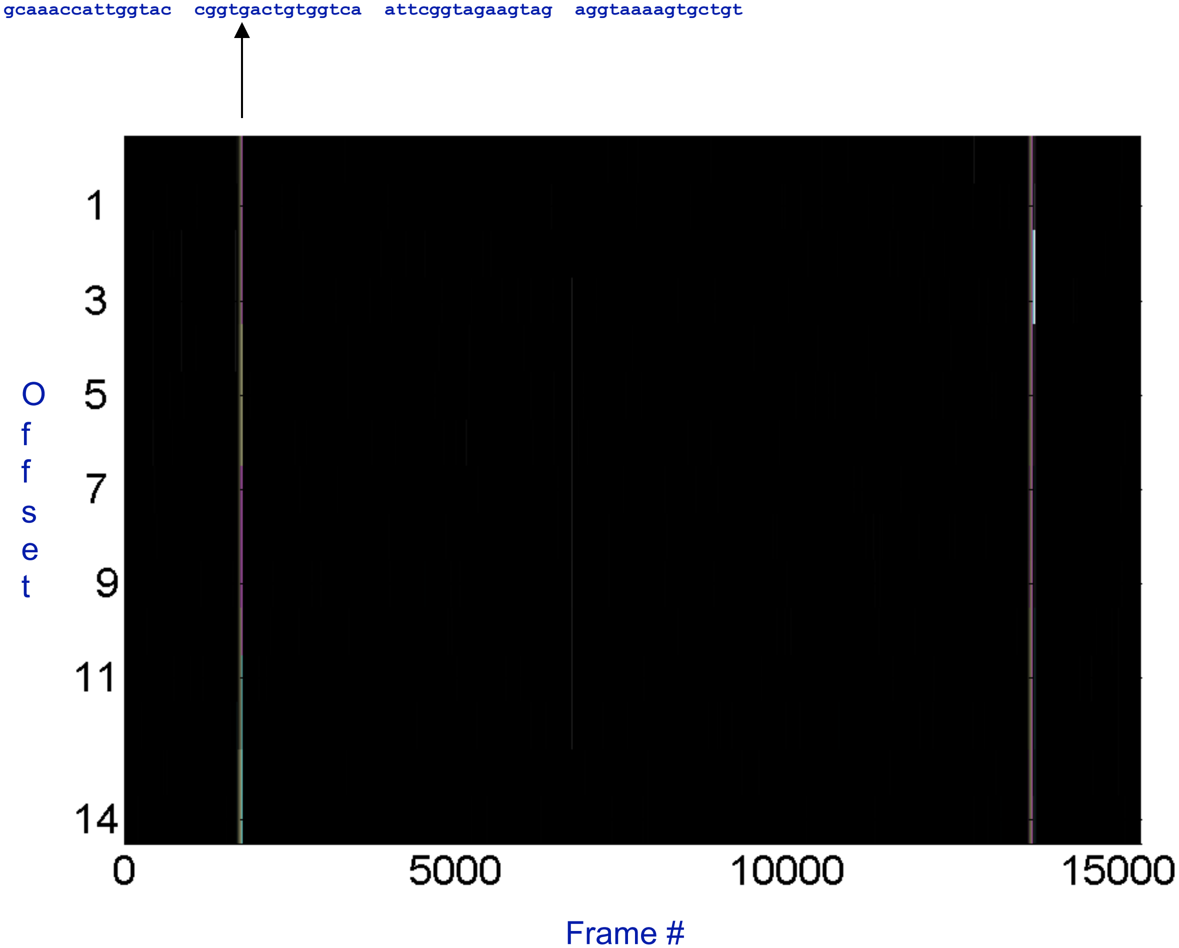
Magenta: N-3 subspace

Yellow:>=N-4 subspace



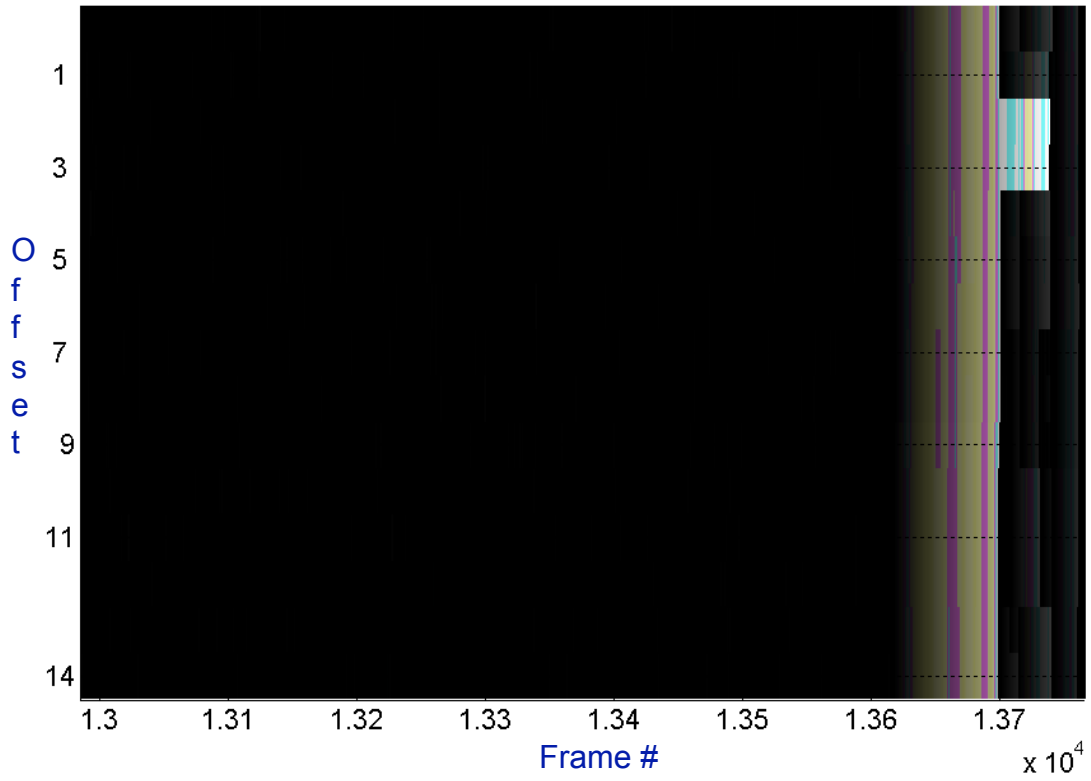
FLO9 Gene in Yeast N=15

cagtttcatcagttg	gttgaccgttggttc	cagtgatggtggtca	tctcagtagatgtag	aggtgaaagtaccgg	tccatggctcggttg	tagttgtaaccaaac	cttcactggttggag	ttctgataacaatca
cggtttcgtcagttg	gttgaccgttagtac	cggtgacggtggtca	tctcagtggatgtag	aggtgaaagtaccag	tccatggttcagtgg	tggtgctgattagac	cttcactagttggag	ttctgatgacaatga
cggtttcgtcagttg	gaacgccgttggtac	cggtgacggtggtca	tttcagtggatgtag	aggtaaaagtgtcgt	tccatggctgagttg	tagtcatggcagtag	tggctgttgttggtg	ttctgatgacaatga
tggtctcatcagttg	gcaaaccgttggtac	cggtgacggtggtga	tttcagtggatgtag	aggtaaaagtgtcgt	tccatggctgagttg	tagtcatggcagtag	tggctgttgttggtg	ttctgatgacaatga
tggtttcatcagttg	gcaaaccgttggtac	cggtgacggtggtca	tttcagtggatgtag	aggtaaaagtgtcgt	tccatggctgagttg	tagttatggcagtag	tggctgttgttggtg	ttctgatgacaatga
tggtttcatcagttg	gcaaaccgttggtac	cggtgacggtggtca	tttcagtggatgtag	aggtaaaagtgctgt	tccatggctcagttg	tagttatggcagtag	tggctgttgttggtg	ttctgatgacaatga
tggtttcatcagttg	gcaaaccgttggtac	cggtgacggtggtca	tttcagtggatgtag	aggtaaaagtgtcgt	tccatggctgagttg	tagtcatggcagtag	tggctgttgttggtg	ttctgatgacaatga
tggtctcatcagttg	gcaaaccattggtac	cggtgacggtggtga	tttcagtggatgtag	aggtaaaagtgtcgt	tccatggctgagttg	tagtcatggcagtag	tggctgttgttggtg	ttctgatgacaatga
tggtctcatcagttg	gcaaaccattggtac	cggtgactgtggtca	attcggtagaagtag	aggtaaaagtgtcgt	tccatggctgagttg	tagtcatggcagtag	tggctgttgttggtg	ttctgatgacaatga
tggtctcatcagttg	gcaaaccattggtac	cggtgactgtggtca	attcggtagaagtag	aggtaaaagtgtcgt	tccatggctgagttg	tagtcatggcagtag	tggctgttgttggtg	ttctgatgacaatga
tggtctcatcagttg	gcaaaccattggtac	cggtgactgtggtca	attcggtagaagtag	aggtaaaagtgctgt				



FLO1 Gene in Yeast N=15

```
ttccaactgacgaaa ccgtcattgtcatca gaactccaacaactg ctagcaccatcataa ctacaactgagccat ggaacagcactttta cctctacttctaccg aattgaccacagtca ctggcaccaatggtg
tacgaactgacgaaa ccatcattgtaatca gaacaccaacaacag ccactactgccataa ctacaactgagccat ggaacagcactttta cctctacttctaccg aattgaccacagtca ccggtaccaatggtt
tgccaactgatgaga ccatcattgtcatca gaacaccaacaacag ccactactgccatga ctacaactcagccat ggaacgacactttta cctctacttctaccg aattgaccacagtca ccggtaccaatggtt
tgccaactgatgaga ccatcattgtcatca gaacaccaacaacag ccactactgccatga ctacaactcagccat ggaacgacactttta cctctacttctaccg aattgaccacagtca ccggtaccaatggtt
tgccaactgatgaga ccatcattgtcatca gaacaccaacaacag ccactactgccatga ctacaactcagccat ggaacgacactttta cctctacatccactg aaatcaccaccgtca ccggtaccaatggtt
tgccaactgatgaga ccatcattgtcatca gaacaccaacaacag ccactactgccatga ctacacctcagccat ggaacgacactttta cctctacatccactg aaatgaccaccgtca ccggtaccaacggtt
tgccaactgatgaaa ccatcattgtcatca gaacaccaacaacag ccactactgccataa ctacaactgagccat ggaacagcactttta cctctacatccactg aaatgaccaccgtca ccggtaccaacggtt
tgccaactgatgaaa ccatcattgtcatca gaacaccaacaacag ccactactgccataa ctacaactcagccat ggaacgacactttta cctctacatccactg aaatgaccaccgtca ccggtaccaacggtt
tgccaactgatgaaa ccatcattgtcatca gaacaccaacaacag ccactactgccatga ctacaactcagccat ggaacgacactttta cctctacatccactg aaatcaccaccgtca ccggtaccaccggtt
tgccaactgatgaga ccatcattgtcatca gaacaccaacaacag ccactactgccatga ctacaactcagccat ggaacgacactttta cctctacatccactg aaatgaccaccgtca ccggtaccaacggcg
ttccaactgacgaaa ccgtcattgtcatca gaactccaactagtg aaggtctaatcagca ccaccactgaaccat ggactggtactttca cctctacatccactg agatgaccaccgtca ccggtactaacggta
Aaccaactgacgaaa ccgtgattgttatca gaactccaaccagtg aaggtttggttacaa ccaccactgaaccat ggactggtactttta cttctacatctactg aaatgaccaccatta ctggaaccaacggcg
ttccaactgacgaaa ccgtcattgtcatca gaactccaaccagtg aaggtctaatcagca ccaccactgaaccat ggactggtactttta cttctacatctactg aaatgaccaccatta ctggaaccaatggtc
aaccaactgacgaaa ccgttattgttatca
```

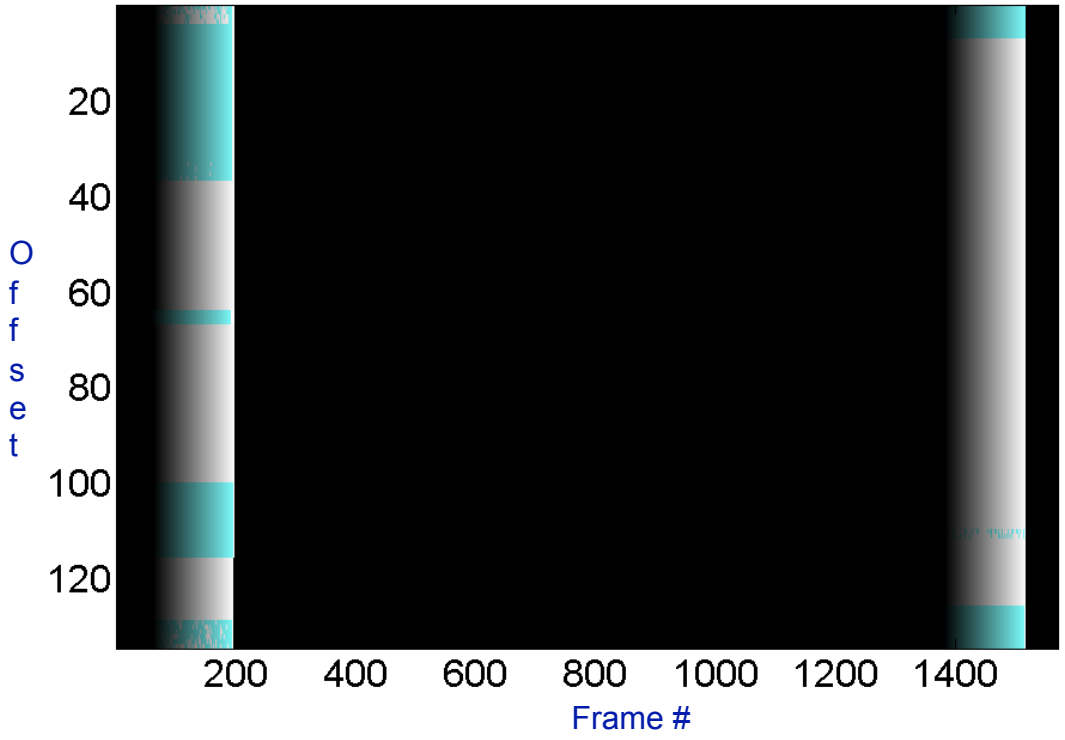


Yeast (Saccharomyces Cerevisiae chromosome I)

17076->18966

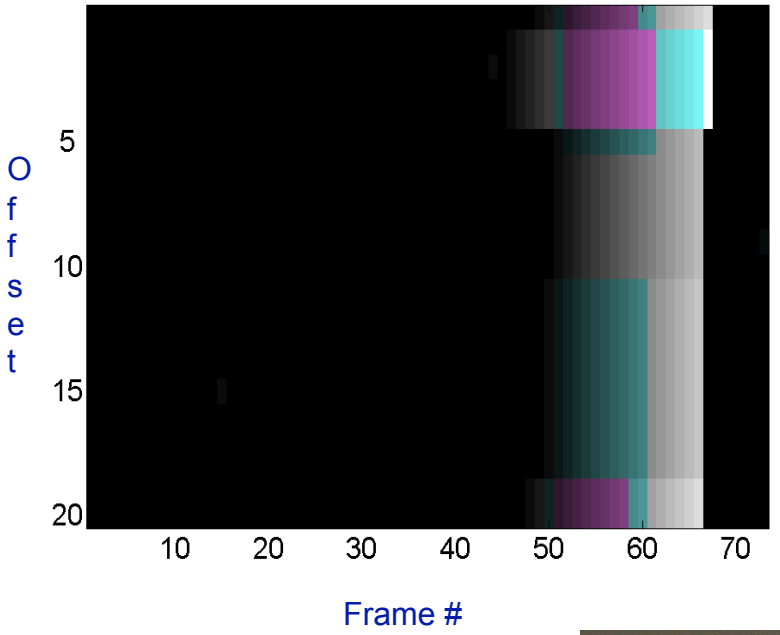
ctaggatagaatggggtaataattggacgcgcagacgtgataaagctggtgatttgtcctgaagaagatgacaaactggatgagatggcagtagttggagttttgacaataatgacagtttcatcagttggttga
ccgttggttccagtgatggtggtcatctcagtagatgtagaggtgaaagtaccggtccatggctcggttgtagttgtaaccaaaccttcactggttggagttctgataacaatcacggtttcgtcagttggttga
ccgttagtaccggtgacggtggtcatctcagtggatgtagaggtgaaagtaccagtccatggttcagtggtggtgctgattagaccttcactagttggagttctgatgacaatgacggtttcgtcagttggaacg
ccgttggtaccggtgacggtggtcatttcagtggatgtagaggtaaaagtgtcgttccatggctgagttgtagtcatggcagtagtggctgttgttggtgttctgatgacaatgatggtctcatcagttggcaaa
ccgttggtaccggtgacggtggtgatttcagtggatgtagaggtaaaagtgtcgttccatggctgagttgtagtcatggcagtagtggctgttgttggtgttctgatgacaatgatggtttcatcagttggcaaa
ccgttggtaccggtgacggtggtcatttcagtggatgtagaggtaaaagtgtcgttccatggctgagttgtagttatggcagtagtggctgttgttggtgttctgatgacaatgatggtttcatcagttggcaaa
ccgttggtaccggtgacggtggtcatttcagtggatgtagaggtaaaagtgctgttccatggctcagttgtagttatggcagtagtggctgttgttggtgttctgatgacaatgatggtttcatcagttggcaaa
ccgttggtaccggtgacggtggtcatttcagtggatgtagaggtaaaagtgtcgttccatggctgagttgtagtcatggcagtagtggctgttgttggtgttctgatgacaatgatggtctcatcagttggcaaa
ccattggtaccggtgacggtggtgatttcagtggatgtagaggtaaaagtgtcgttccatggctgagttgtagtcatggcagtagtggctgttgttggtgttctgatgacaatgatggtctcatcagttggcaaa
ccattggtaccggtgactgtggtcaattcggtagaagtagaggtaaaagtgtcgttccatggctgagttgtagtcatggcagtagtggctgttgttggtgttctgatgacaatgatggtctcatcagttggcaaa
ccattggtaccggtgactgtggtcaattcggtagaagtagaggtaaaagtgtcgttccatggctgagttgtagtcatggcagtagtggctgttgttggtgttctgatgacaatgatggtctcatcagttggcaaa
ccattggtaccggtgactgtggtcaattcggtagaagtagaggtaaaagtgctgttccatggctcagttgtagttatggcagtagtggctgttgttggtgttctgattacaatgatggtttcgtcagttcgtaca
ccattggtgccagtgactgtggtcaattcggtagaagtagaggtaaaagtgctgttccatggctcagttgtagttatgatggtgctagcagttgttggagttctgatgacaatgacggtttcgtcagttggaacg
ccgttggtaccggtgacggtggtcatttcagtagatgtagaagtgaaagtaccggtccatggttccgttgtagttatggtagtactgacagtataatttgaagggtctggaatggtacagtttggctggcttaga
ttgttgtcaaaagtatatacgtacccttcaaagtcatcactaacggtagtgccatctggtagtgtcacactaatttggaagtgtaccccaggcaacggcatttgagtaaacaatcttcattggataatagaaacca

135 Base Periodicity!

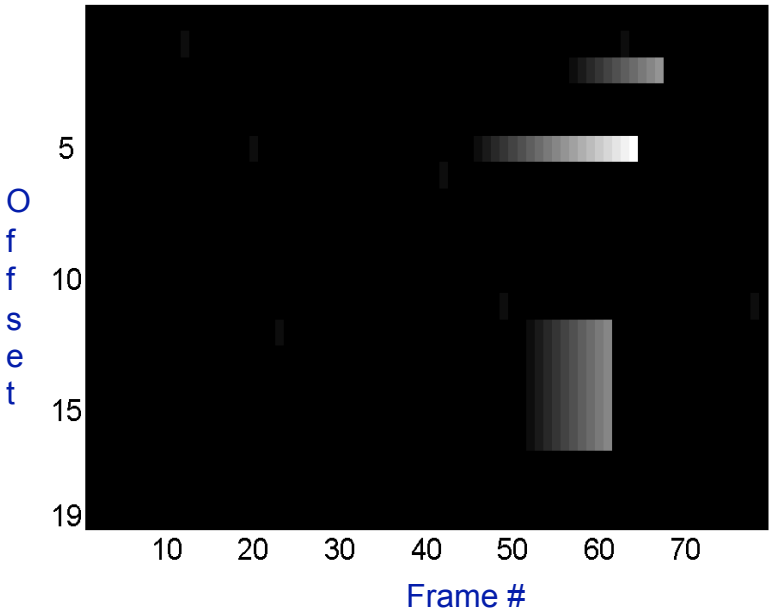


HSVDJSAT Results

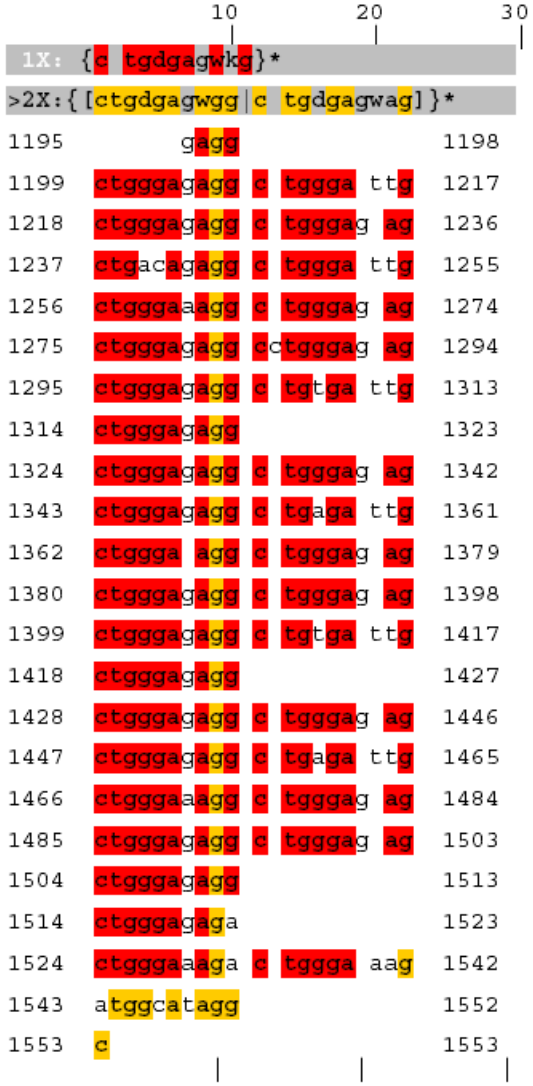
N=21



N=20



HSVDJSAT Comparison



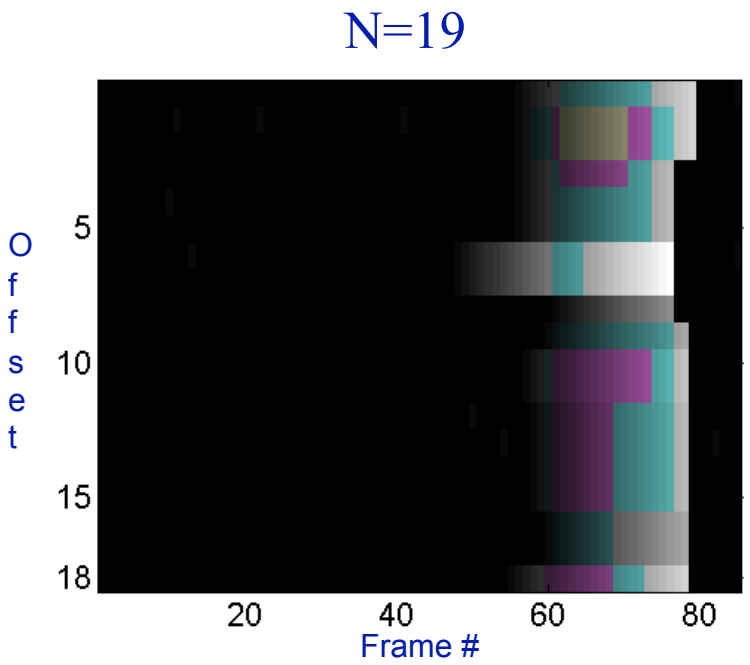


Figure 5.1.10.B: Region characterization for the integer multiple of 2 having a 19 bp pattern associated with the MPTR region in the HSVDJSAT sequence.

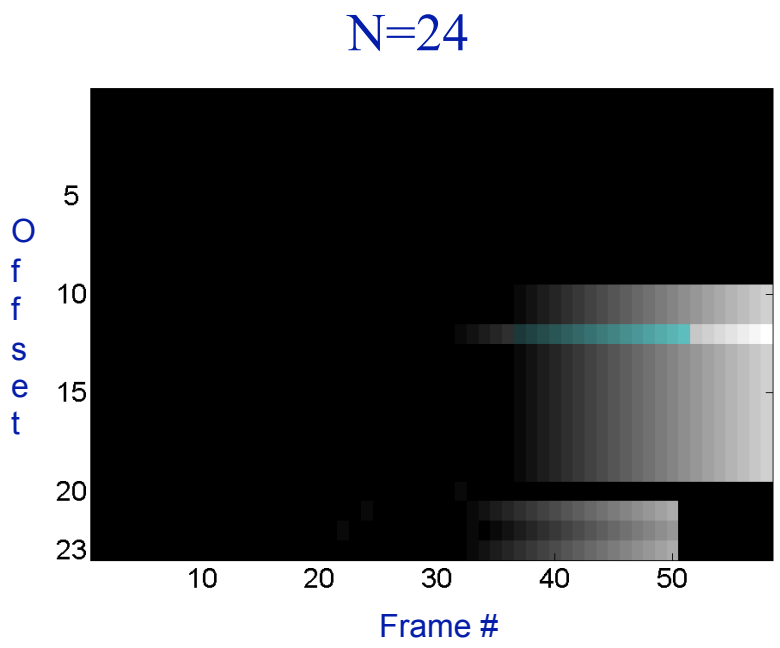
Hauth, Amy. *Identification of Tandem Repeats: Simple and Complex Pattern Structures in DNA.* Ph.D. Thesis, 2002.

HSVDJSAT N=24

1200

```
cactctaggacacccagcagggca
gtgttgagagtgagcatcctggca gggctggaggctgggagaggctgg
gattgctgggagaggctgggagag ctgacagaggctgggattgctggg
aaaggctgggagagctgggagagg cctgggagagctgggagaggctgt
gattgctgggagaggctgggagag gctgggagagctgggagaggctga
gattgctgggaaggctgggagagc tgggagaggctgggagagctggga
gaggctgtgattgctgggagaggc tgggagaggctgggagagctggga
gaggctgagattgctgggaaaggc tgggagagctgggagaggctggga
gagctgggagaggctgggagagac tgggaaagactgggaaagatggca
taggccttgagccaggagtgtgag ttcatgaagataggctgggggagt
gagagatgcgtggggcaaggagga aggcagcagttcaggggtagccca
tggagctgtatctggagcagccac gtgggtcacttctacccacagtgg
aggtggactcttgtagccagagct gtggacaacctctcagaaccagaa
gacccttgctgccctgtatgccaa ggtctcctccggcctgggtctcag
ggatgccagctgcaaactgggagg gccattgtacagacactaggtggc
tgaggtaccagttacagcctggtc ttggtggccacatagaggtccagc
ctcactcagcttgatggccaagct ggtgggttaggatttggagtctgc
agccttgaggccttcccaaggtaa aaccaaattgtcctggcttagaat
```

1141->1980



Satellite 1200..1543 /note="minisatellite (polymorphism)"

The End

Abstract Algebra and Coding

- Finite Fields -- Symbolic Representation

Table 2. Exponential root representation, polynomial representation, numerical label, and nucleotide label for the *GF(4)* representation.

$$\alpha^0 = 1 \Leftrightarrow 1 \Leftrightarrow C$$
$$\alpha^1 = \alpha \Leftrightarrow 2 \Leftrightarrow T$$
$$\alpha^2 = \alpha + 1 \Leftrightarrow 3 \Leftrightarrow G$$
$$0 = 0 \Leftrightarrow 0 \Leftrightarrow A$$

Table 3. Addition and multiplication tables in *GF(4)*.

+	0	1	2	3
0	0	1	2	3
1	1	0	3	2
2	2	3	0	1
3	3	2	1	0

x	0	1	2	3
0	0	0	0	0
1	0	1	2	3
2	0	2	3	1
3	0	3	1	2

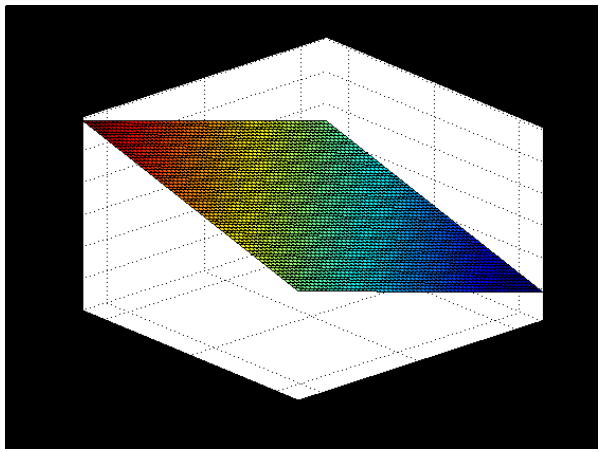
Search Methods for Linear Coding

- Find rank-deficiency in a DNA N^2 length string

  -Uncovers potential redundancy

- Decompose a sequence using Gram-Schmidt

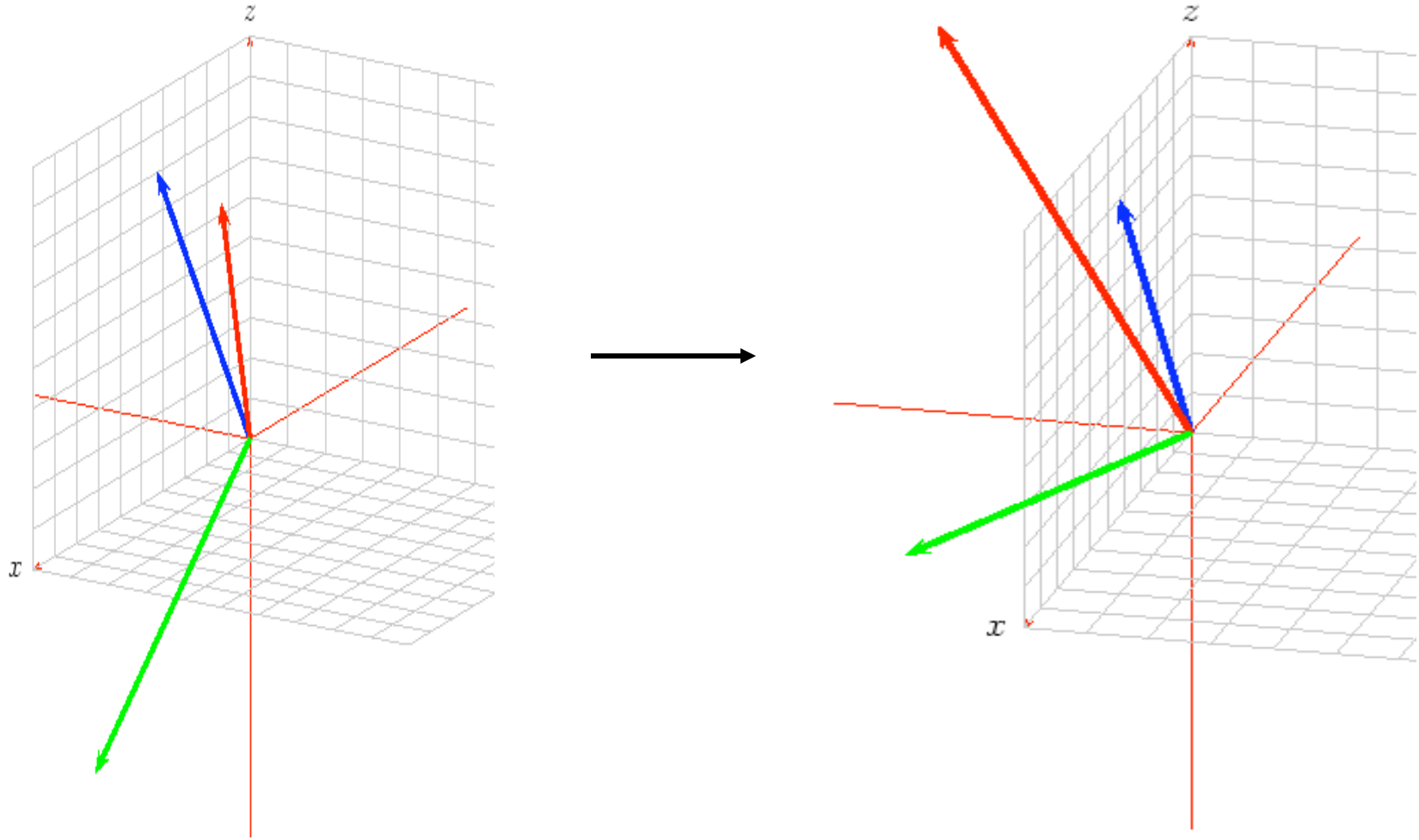
  -If frames of a sequence lie in a subspace, linear coding redundancy exists



If we take n-length frames of a genetic sequence, do they have a linear block code structure as found in error-correcting codes?

(i.e.: for n=3, do most frames lie in a two-dimensional subspace?)

Gram-Schmidt Procedure



Novelty: Represent DNA in Space, Find subspaces, etc.

Gram-Schmidt Orthonormal basis

Given a set of *M* linearly independent vectors, we can construct an orthonormal set which are linear combinations of the original set and which span the same space.

$$\bar{\mathbf{s}}_i = \left[s_{i0}, s_{i1}, s_{i2}, \ldots s_{iN} \right]$$

$0 < i < M$: number of vectors

Gram-Schmidt Procedure

1. Set $\tilde{\mathbf{s}}_0 \stackrel{\triangle}{=} \frac{\bar{\mathbf{s}}_0}{|\,\bar{\mathbf{s}}_0\,|}$

2. $\bar{\mathbf{x}}_1 \stackrel{\triangle}{=} \bar{\mathbf{s}}_1 - \mathbf{P}_{\tilde{\mathbf{s}}_0}(\bar{\mathbf{s}}_1) = \bar{\mathbf{s}}_1 - \langle \bar{\mathbf{s}}_1, \tilde{\mathbf{s}}_0 \rangle \, \tilde{\mathbf{s}}_0$

3. $\tilde{\mathbf{s}}_1 \stackrel{\triangle}{=} \frac{\bar{\mathbf{x}}_1}{\|\,\bar{\mathbf{x}}_1\,\|}$

4. $\bar{\mathbf{x}}_2 \stackrel{\triangle}{=} \bar{\mathbf{s}}_2 - \mathbf{P}_{\tilde{\mathbf{s}}_0}(\bar{\mathbf{s}}_2) - \mathbf{P}_{\tilde{\mathbf{s}}_1}(\bar{\mathbf{s}}_2) = \bar{\mathbf{s}}_2 - \langle \bar{\mathbf{s}}_2, \tilde{\mathbf{s}}_0 \rangle \, \tilde{\mathbf{s}}_0 - \langle \bar{\mathbf{s}}_2, \tilde{\mathbf{s}}_1 \rangle \, \tilde{\mathbf{s}}_1$

5. $\tilde{\mathbf{s}}_2 \stackrel{\triangle}{=} \frac{\bar{\mathbf{x}}_2}{\|\,\bar{\mathbf{x}}_2\,\|}$

6. Continue this process until $\tilde{\mathbf{s}}_{N-1}$ has been defined.

Subspace Partitioning

1. Obtain the orthonormal basis, $\mathbf{\{e_1,e_2,\ldots,e_n\}}$, by Gram-Schmidt (G-S) orthogonalization from initial frames of data. Form the transform matrix, **G**, from this set.
2. Decompose the sequence with **G** into basis components, $\mathbf{\{t_1,t_2,\ldots t_{N-j}\}}$, across all possible framing offsets:

$$t_i = \mathbf{G}\, v_i \qquad \text{where} \quad \mathbf{G} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

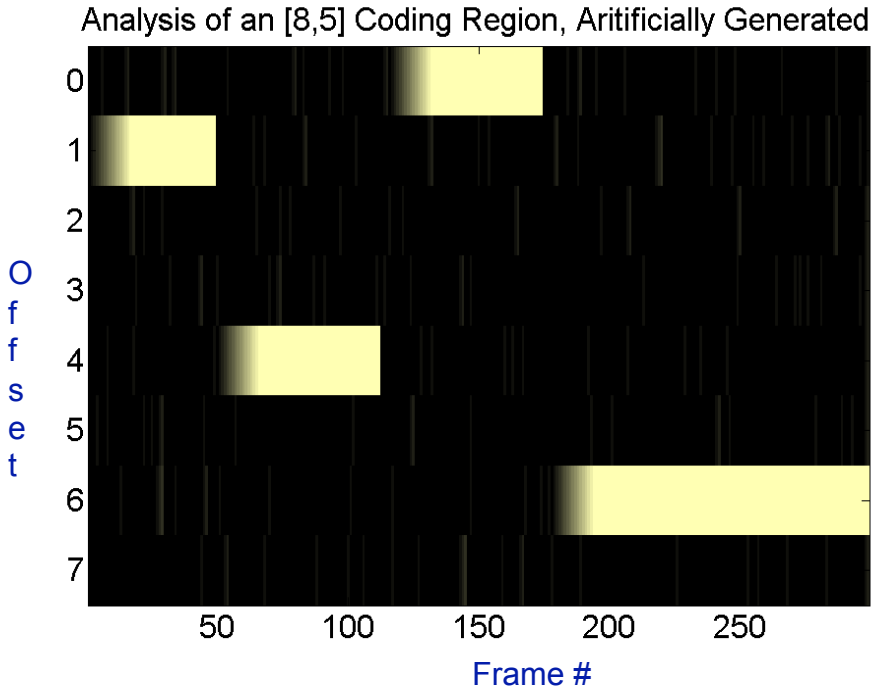
3. Note the persistence of nulls in $\mathbf{t}_i$'s. Calculate confidence over an interval.

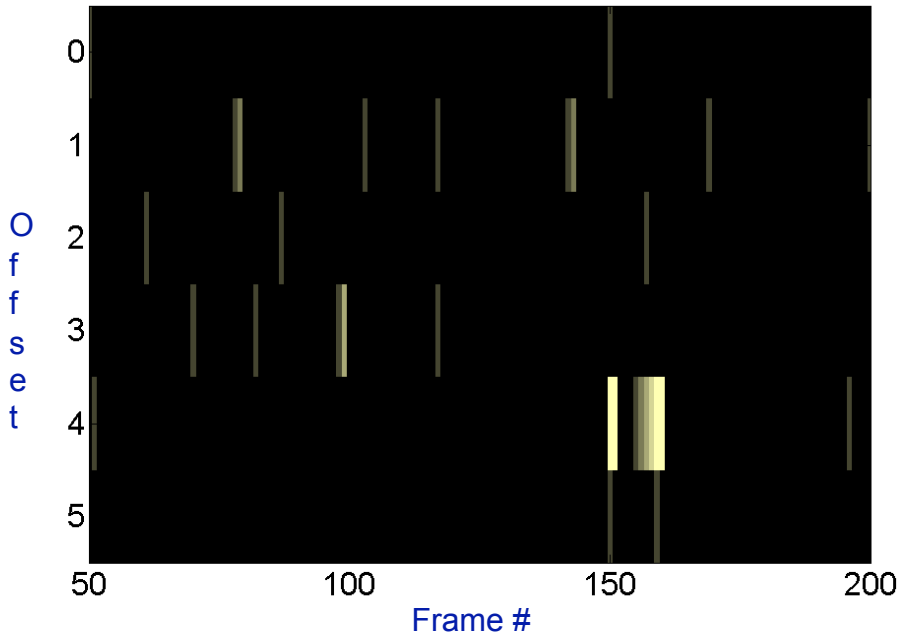
Advantages: Determines if a consistent coding subspace exists.

Disadvantages: When G-S formed, the coding subspace must exist!

Ideal vs. Actual Results

Confidence robust to shift errors



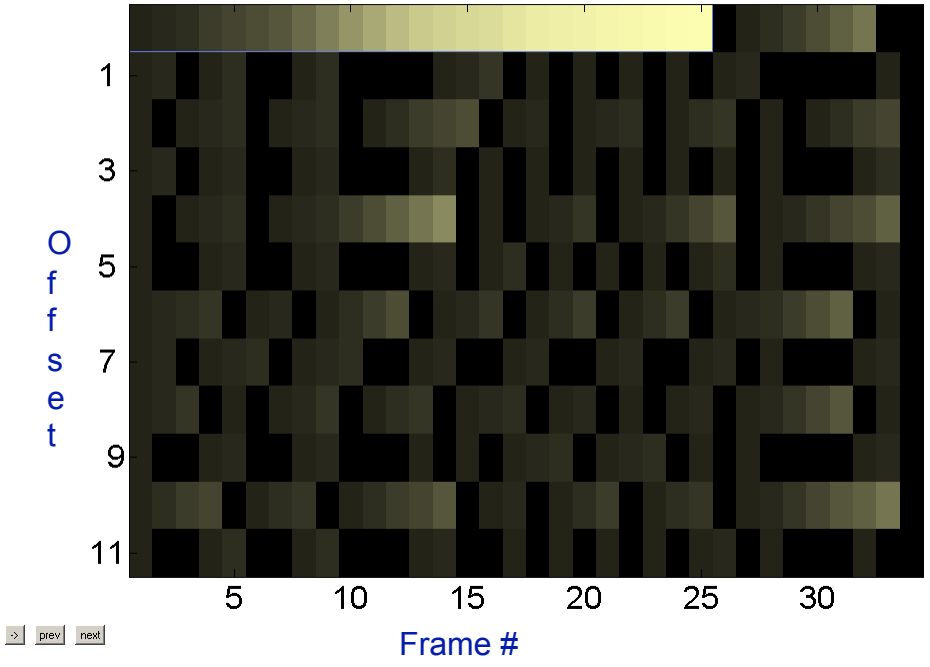


Analysis of an (N=8,K=5) coding region with insertions/deletions, Test data

Analysis of a N=6 E. Coli MG1655 sequence

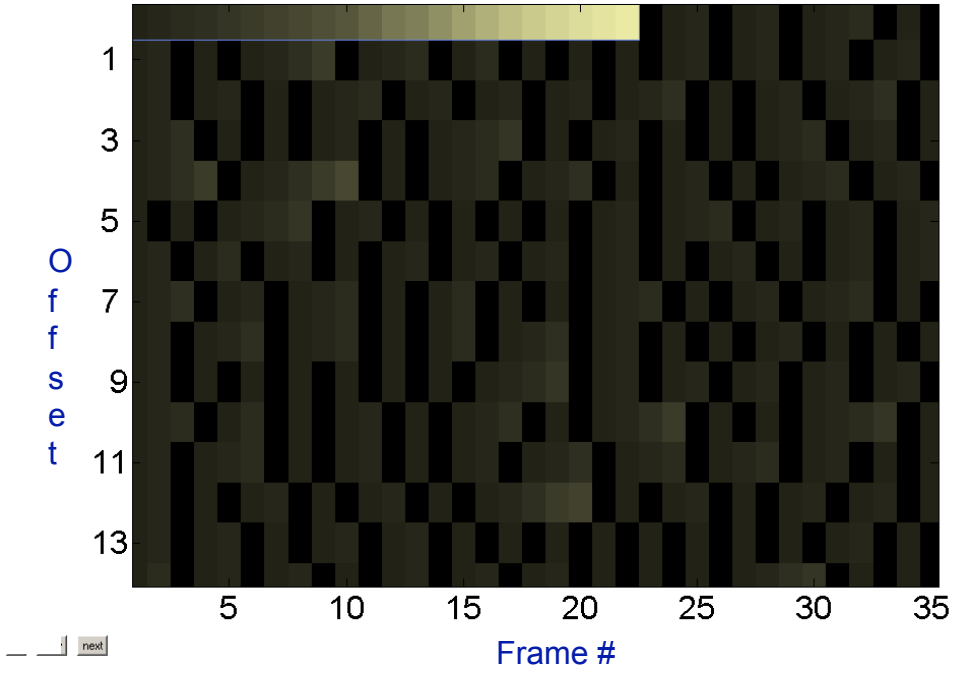
Subspace Partitioning Results

N=12, Shift=9,Offset=296 bp



BOVTGN sequence

N=15, Shift=0,Offset=25424 bp



Yeast

Summary