

CS 364  
COMPUTATIONAL  
BIOLOGY

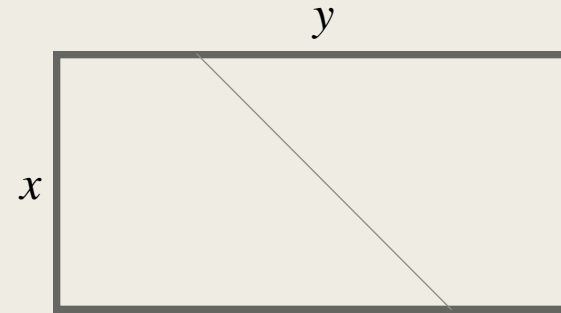
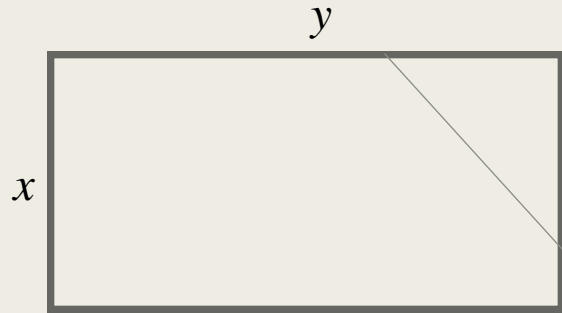
Sara Mathieson  
Haverford College

# Outline

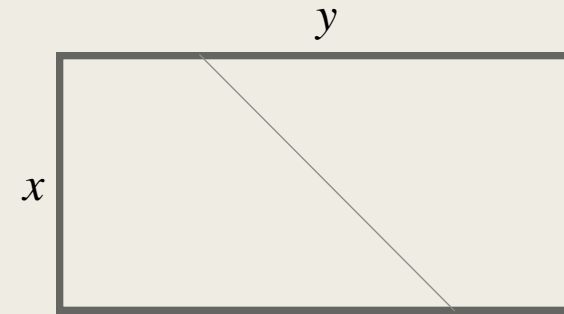
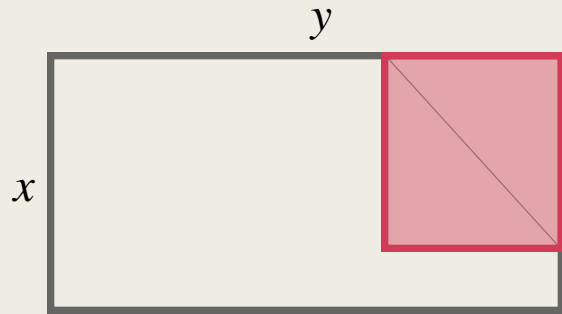
- Pairwise alignment variations
- Multiple sequence alignment
- Introduction to phylogenetics

# Local alignment variations

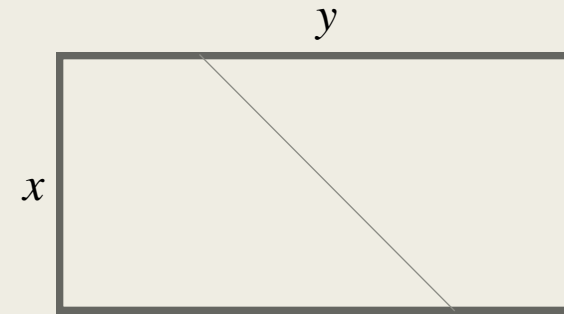
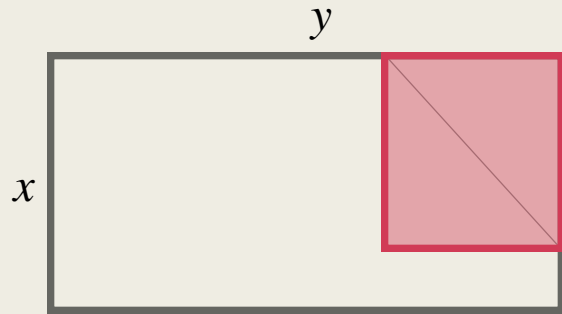
Handout 7: what portion of  $x$  aligns to what portion of  $y$ ?



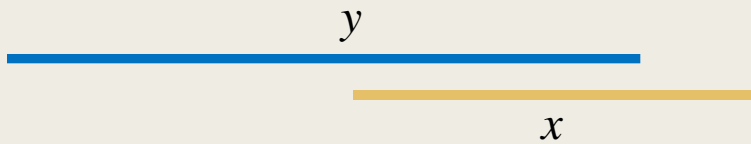
Handout 7: what portion of  $x$  aligns to what portion of  $y$ ?



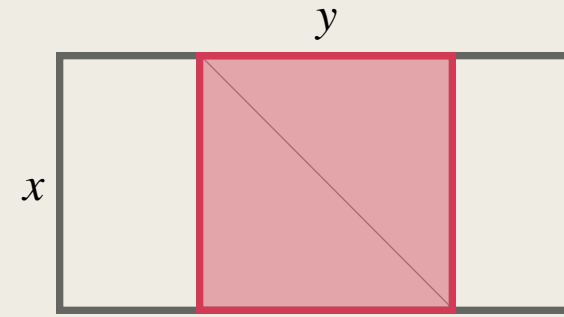
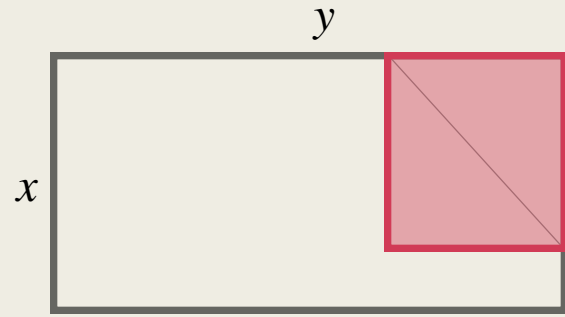
# Handout 7: what portion of $x$ aligns to what portion of $y$ ?



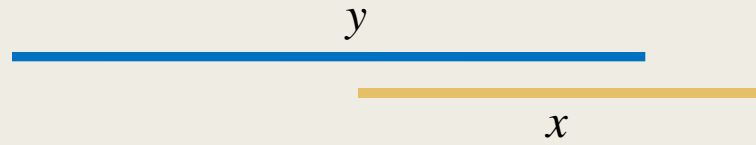
Beginning of  $x$  with end of  $y$



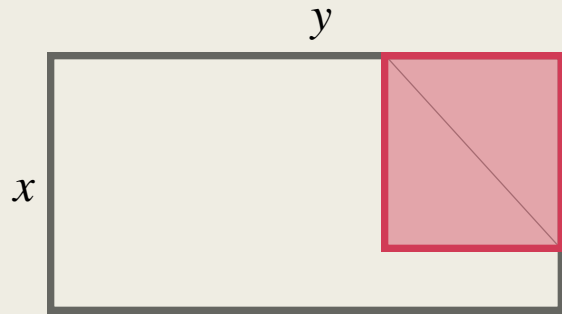
# Handout 7: what portion of x aligns to what portion of y?



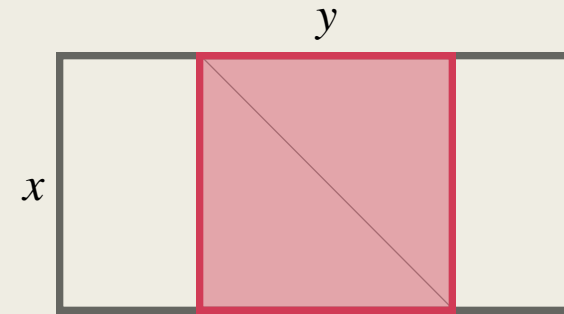
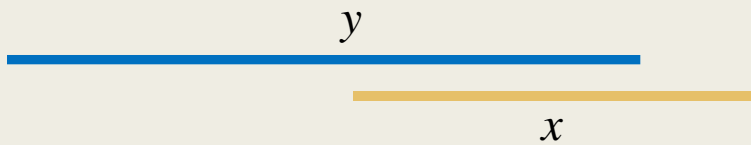
Beginning of x with end of y



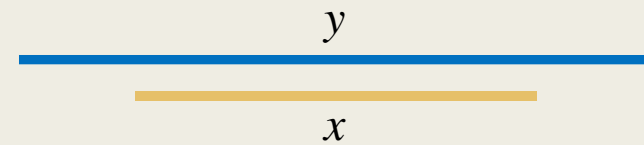
# Handout 7: what portion of $x$ aligns to what portion of $y$ ?



Beginning of  $x$  with end of  $y$

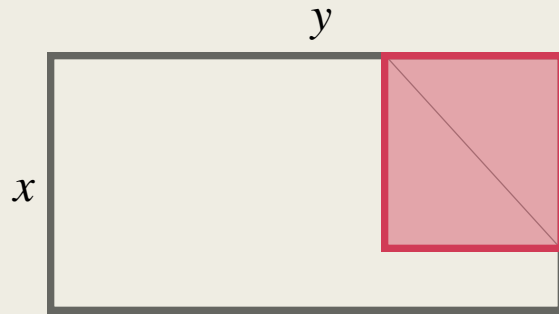


$x$  aligns with the middle of  $y$

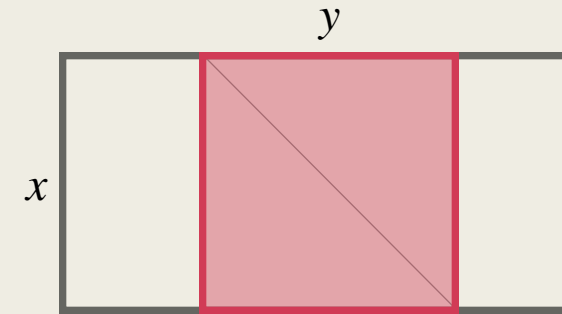
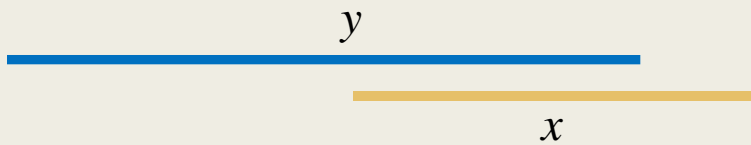




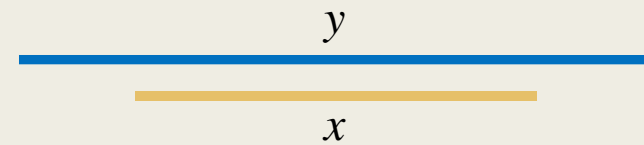
# Handout 7: what portion of x aligns to what portion of y?



Beginning of  $x$  with end of  $y$

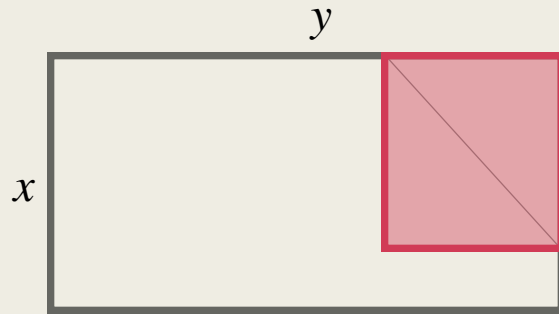


$x$  aligns with the middle of  $y$

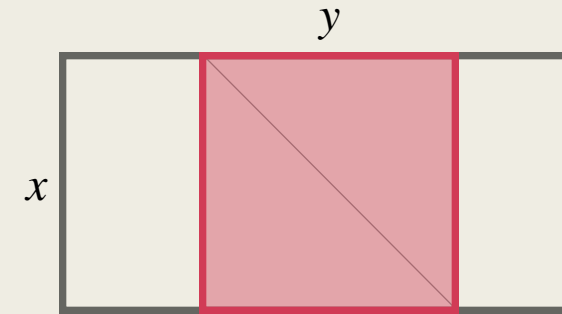
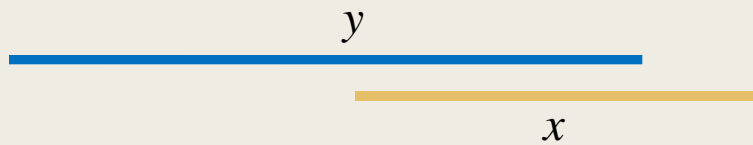


DP algorithm modifications:

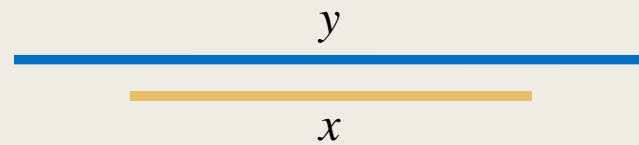
# Handout 7: what portion of x aligns to what portion of y?



Beginning of x with end of y



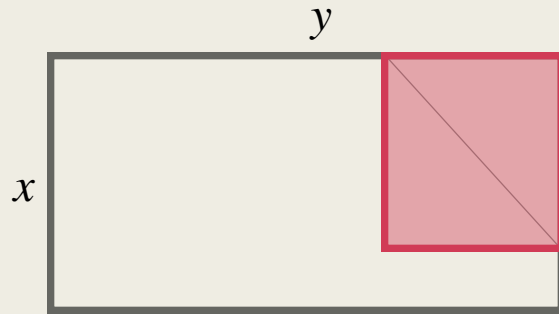
x aligns with the middle of y



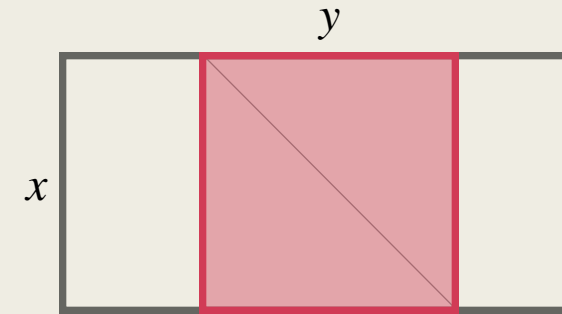
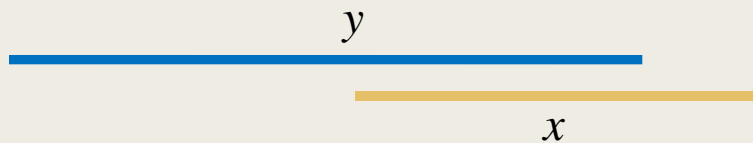
DP algorithm modifications:

- 1) **Initialization:**  $0^{\text{th}}$  row and  $0^{\text{th}}$  column with 0's to not penalize leading/trailing gaps

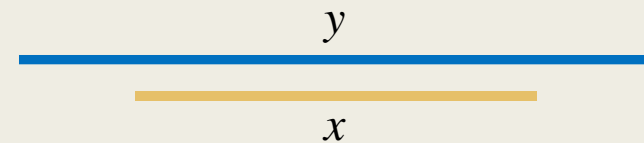
# Handout 7: what portion of x aligns to what portion of y?



Beginning of x with end of y



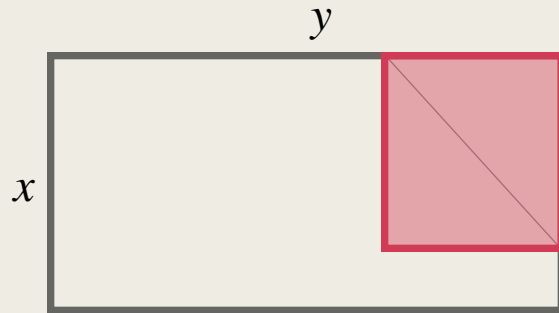
x aligns with the middle of y



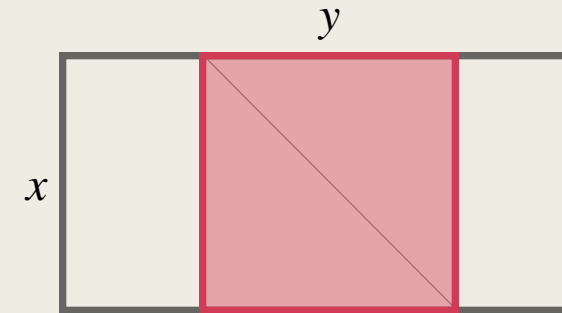
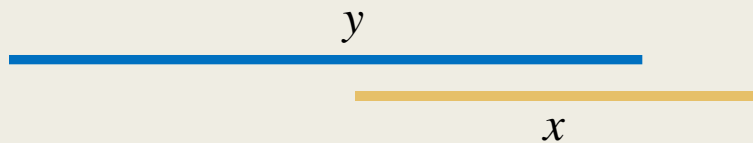
DP algorithm modifications:

- 1) **Initialization:**  $0^{\text{th}}$  row and  $0^{\text{th}}$  column with 0's to not penalize leading/trailing gaps
- 2) **Recursion:** to fill in the rest of the table, use global alignment (i.e. don't restart at 0)

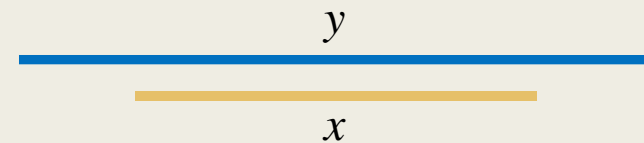
# Handout 7: what portion of x aligns to what portion of y?



Beginning of x with end of y



x aligns with the middle of y



## DP algorithm modifications:

- 1) **Initialization:**  $0^{\text{th}}$  row and  $0^{\text{th}}$  column with 0's to not penalize leading/trailing gaps
- 2) **Recursion:** to fill in the rest of the table, use global alignment (i.e. don't restart at 0)
- 3) **Traceback:** start at the maximum value along the last row or last column

# Approximate local alignment: BLAST

I sequenced something – what is it?

Want to compare with all the sequences in published databases

Smith-Waterman far too slow ( $O(n^2)$  time and space)

So in practice , often use heuristic (approximate) methods

BLAST (Basic Local Alignment Search Tool)

<https://blast.ncbi.nlm.nih.gov/Blast.cgi>

“Basic local alignment search tool.” Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. [\*J. Mol. Biol.\* 215, 403–410 \(1990\)](#). 12<sup>th</sup> most cited scientific paper of all time (as of 2014).

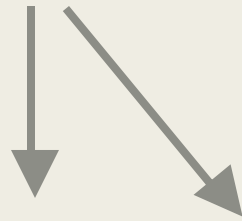
# Approximate local alignment: BLAST

## 1 Find potential matches (“seeds”)

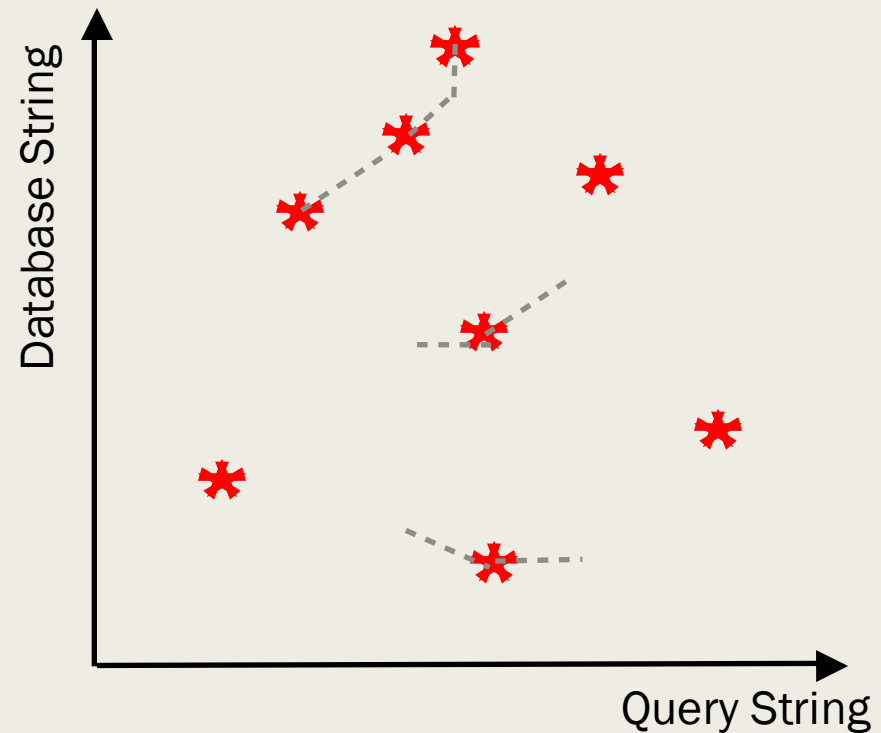
Scoring matrix aware k-mer search

Query string: ...RP**PEG**IAAXXPAPX...

Database string: ...G**PIG**ISID**PEG**PX...



## 2 Extend matches around seed and join up high-scoring pairs



Not guaranteed to find the optimal alignment  
But in practice it is very good!

Report the “E-value”: Expected number of matches we would see by chance

Affine (non-linear) gaps

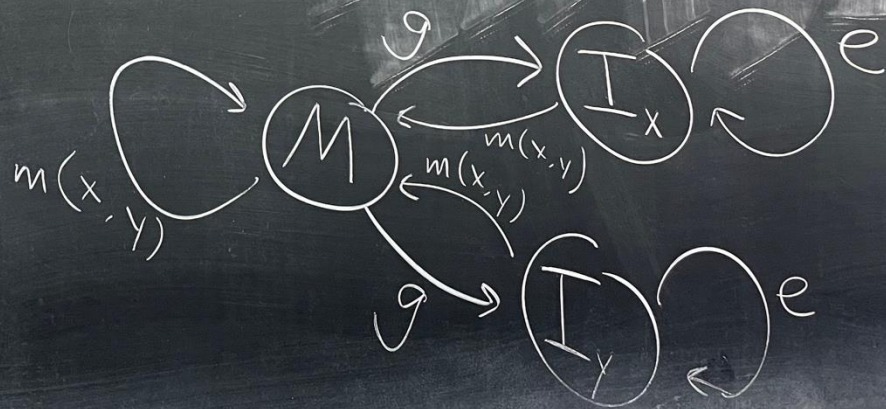
Affine gaps  
(nonlinear)

$$\gamma(l) = g + e \cdot (l-1)$$

$\uparrow$  length of gap  
 $\uparrow$  gap open (i.e. -3)  
 $\uparrow$  gap extension (i.e. -1)

A A C G T G  
 A - - - G  
 $l=4$

$$\gamma(4) = -3 + (-1) \cdot 3 = -6$$





# Multiple Sequence Alignment

# From pairwise to multiple alignment

```
AACTAGAGAG  
AAC--AAGAG
```

K=2 sequences

```
AACTAGA-GAG  
AAC--AA-GAG  
AACTAGA-GAG  
AAATAGA-GAG  
AAATAGA-GAG  
AAGTAGATGAG
```

K=6 sequences

# How to score multiple alignments

A A C T A G A G A G  
A A C - - A A G A G

Score =  $m(A,A) + m(A,A) + \dots$

A A C T A G A - G A G  
A A C - - A A - G A G  
A A C T A G A - G A G  
A A A T A G A - G A G  
A A A T A G A - G A G  
A A G T A G A T G A G

Score =  $S(\text{column 1}) + S(\text{column 2}) + \dots$

Sum of pairs

$$S(\text{column } i) = \sum_{k < l} m(x_i^{(k)}, y_i^{(l)})$$

$$\left. \begin{aligned} m(x_i, -) &= g \\ m(-, -) &= 0 \end{aligned} \right\}$$

i  
G  
G  
C  
|  
|  
G

$$\binom{k}{2} = \frac{6 \cdot 5}{2} = 15$$

match =  $\begin{cases} +1 & \text{if equal} \\ -1 & \text{if not} \end{cases}$   
gap = -1

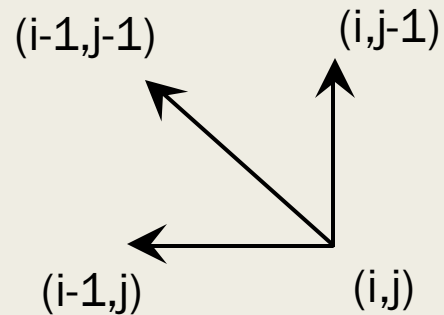
match mismatch gap/gap

$$3 - 3 - 8 + 0 = \boxed{-8}$$

$$K=6$$

# K-dimensional dynamic programming

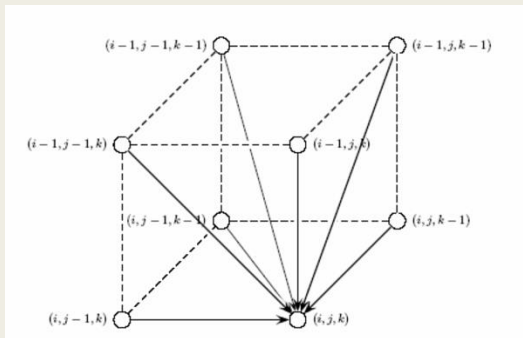
K=2



sequences  $a=a_1a_2a_3\dots$   $b=b_1b_2b_3\dots$

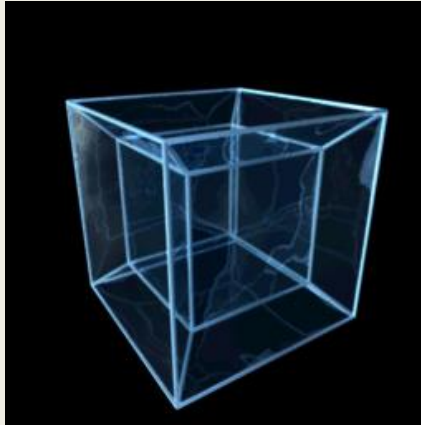
$$S(i, j) = \max \begin{cases} S(i-1, j-1) + S(a_i, b_j) \\ S(i-1, j) + S(a_i, -) \\ S(i, j-1) + S(-, b_j) \end{cases}$$

K=3

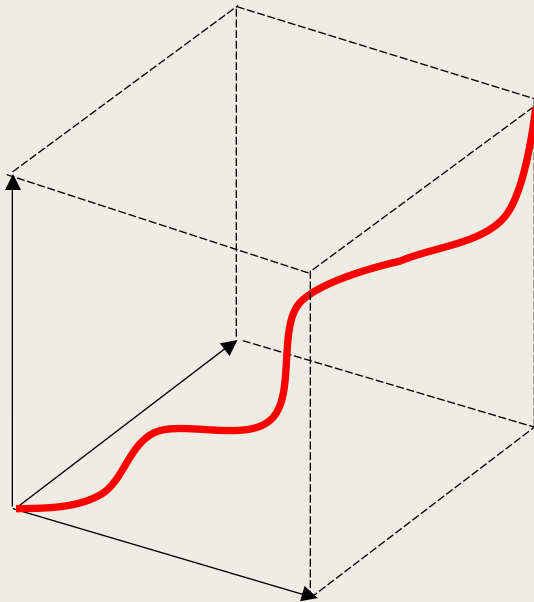


$$S(i, j, k) = \max \begin{cases} S(i-1, j-1, k-1) + S(a_i, b_j, c_k) \\ S(i, j-1, k-1) + S(-, b_j, c_k) \\ S(i-1, j, k-1) + S(a_i, -, c_k) \\ S(i-1, j-1, k) + S(a_i, b_j, -) \\ S(i, j, k-1) + S(-, -, c_k) \\ S(i, j-1, k) + S(-, b_j, -) \\ S(i-1, j, k) + S(a_i, -, -) \end{cases}$$

# K-dimensional dynamic programming



$$S(i, j, k, \dots) = \max \left\{ \begin{array}{l} S(i-1, j-1, k-1, \dots) + S(a_i, b_j, c_k, \dots) \\ S(i, j-1, k-1, \dots) + S(-, b_j, c_k, \dots) \\ \vdots \\ S(i, j, k, \dots) + S(-, -, -, \dots) \\ S(i, j, k, \dots) + S(-, -, -, \dots) \end{array} \right.$$



$2^K - 1$  choices for each cell

Number of cells =  $\prod_i \text{length}(\text{sequence}_i)$   
 $= n^K$  if all length  $n$

So algorithm is  $O(2^K n^K)$

Options

$$K=2, \quad 3$$

$$K=3, \quad 7$$

$K$

$2^k$

gap

all in  
seqs

$K = \#$  sequences,  $x^{(1)}, x^{(2)} \dots x^{(K)}$   
 $n \approx$  len of each seq  
# entries in DP table  
 $= \text{len}(x^{(1)}) \cdot \text{len}(x^{(2)}) \dots \text{len}(x^{(K)})$   
 $= n^K$

Overall  $\Rightarrow O(n^K 2^K)$

# K-dimensional dynamic programming

Assume you have some sequences that are  $n=50$  characters long, and that pairwise alignment of  $K=2$  such sequences takes one second on your computer.

Then alignment of  $K=4$  sequences takes  $0.0001 \cdot (2n)^K = 10^4 \text{s} \sim 3 \text{ hours}$

Suppose you had unlimited memory and you were willing to wait until the sun burns out and turns into a red dwarf in 5 billion years, how many sequences could you align?





# Progressive alignment

# Iterative alignment

Idea: align sequences to each other in some order, so we only ever have to solve pairwise problems

X1 : AACTAGA-GAG  
X2 : ATC--AA-GAG

X1 : AACTAGA-GAG  
X2 : ATC--AA-GAG  
X3 : AACTAGA-GAG

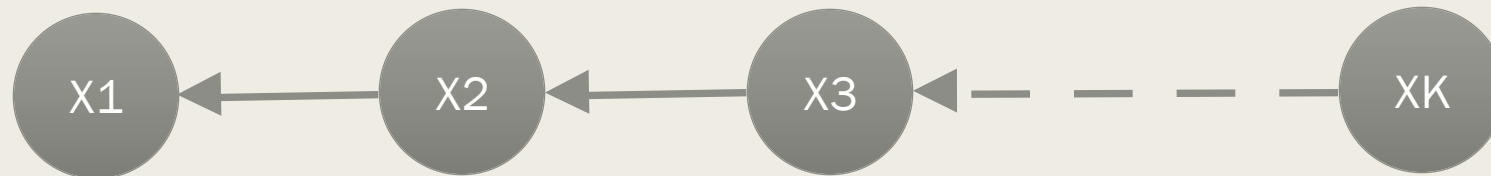
...

AACTAGA-GAG  
ATC--AA-GAG  
AACTAGA-GAG  
AAATAGA-GAG  
AAATAGA-GAG  
AAGTAGATGAG

1) Align X2 to X1

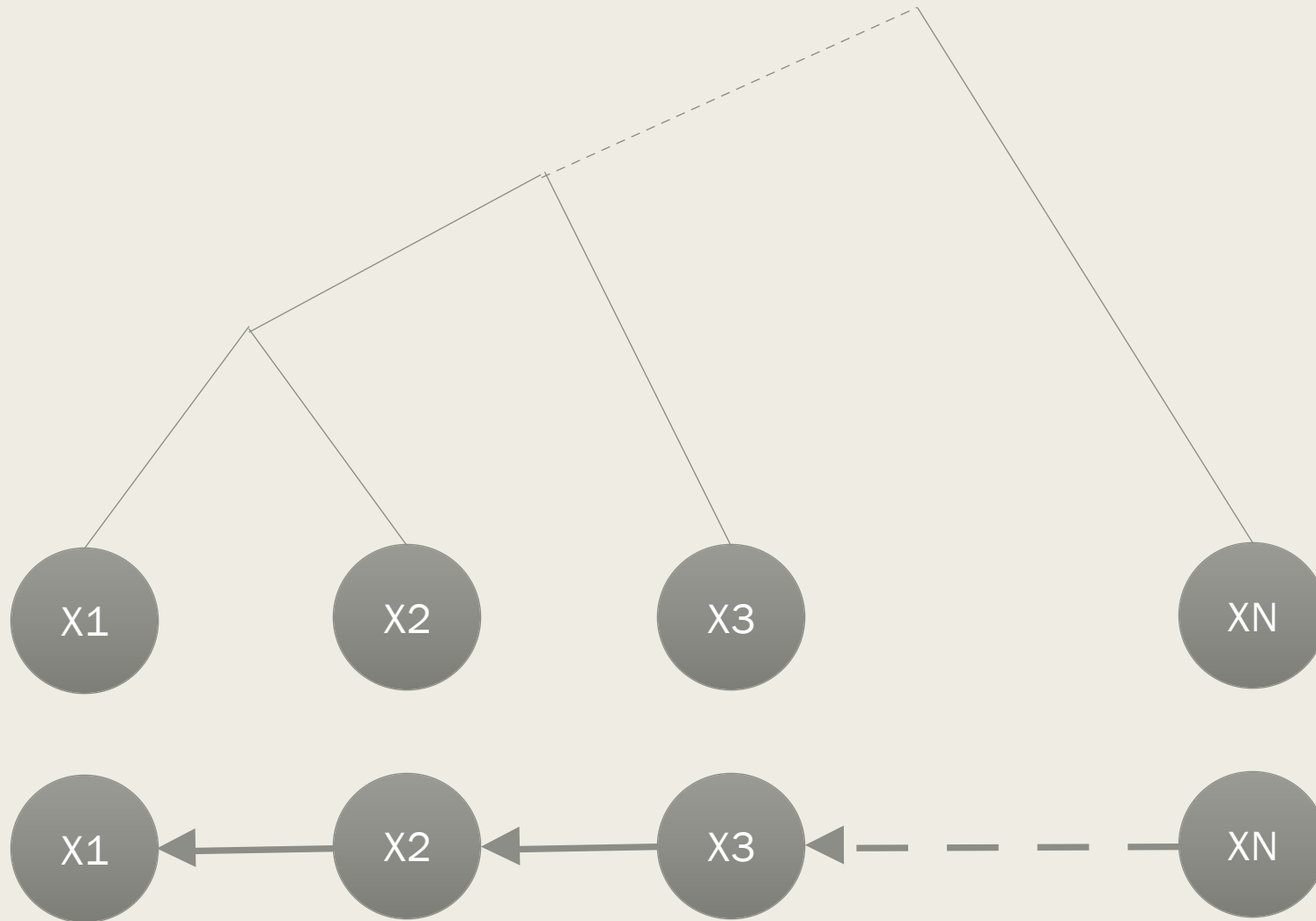
2) Align X3 to  
the alignment  
of X2 and X3\*

K) etc...



\*How to align a sequence to an alignment?

# Progressive alignment



Key idea: Align most similar sequences first

We want to use the multiple alignment to reconstruct the tree



But we need to know the tree to get the right multiple alignment

# Clustal-W: multiple sequence alignment algorithm

- 75,000 citations (2024)

JOURNAL ARTICLE

## **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice**

[Julie D. Thompson](#), [Desmond G. Higgins](#), [Toby J. Gibson](#)  [Author Notes](#)

*Nucleic Acids Research*, Volume 22, Issue 22, 11 November 1994, Pages 4673–4680,

<https://doi.org/10.1093/nar/22.22.4673>

**Published:** 11 November 1994 **Article history** ▼

# Clustal-W: multiple sequence alignment algorithm

- Compute all possible pairwise alignments, use to create a pairwise distance matrix between the strings
- Use the pairwise distance matrix to create a UPGMA or NJ tree graph (next week!)
- Use the tree graph to iteratively merge pairwise alignments of nodes in the tree (which may be leaves or ancestral nodes). Start with most similar nodes first

Pairwise alignment:  
Calculate distance matrix

Hbb_Human	1	-					
Hbb_Horse	2	.17	-				
Hba_Human	3	.59	.60	-			
Hba_Horse	4	.59	.59	.13	-		
Myg_Phyca	5	.77	.77	.75	.75	-	
Glb5_Petma	6	.81	.82	.73	.74	.80	-
Lgb2_Luplu	7	.87	.86	.86	.88	.93	.90
		1	2	3	4	5	6

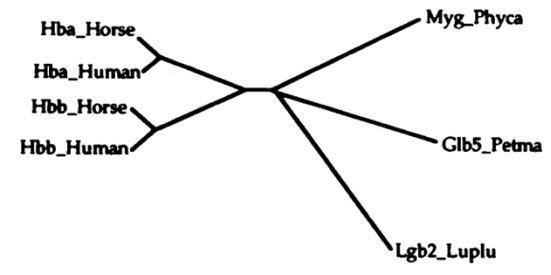
# Clustal-W workflow

# Clustal-W workflow

Pairwise alignment:  
Calculate distance matrix

Hbb_Human	1	-					
Hbb_Horse	2	.17	-				
Hba_Human	3	.59	.60	-			
Hba_Horse	4	.59	.59	.13	-		
Myg_Phyca	5	.77	.77	.75	.75	-	
Glb5_Petma	6	.81	.82	.73	.74	.80	-
Lgb2_Luplu	7	.87	.86	.86	.88	.93	.90
		1	2	3	4	5	6

Unrooted Neighbor-Joining tree



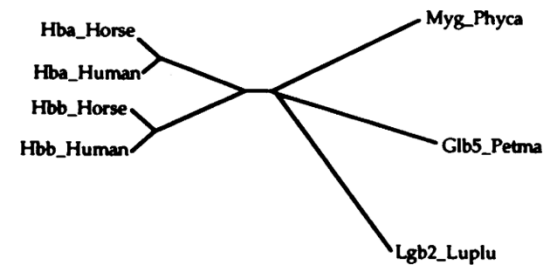


# Clustal-W workflow

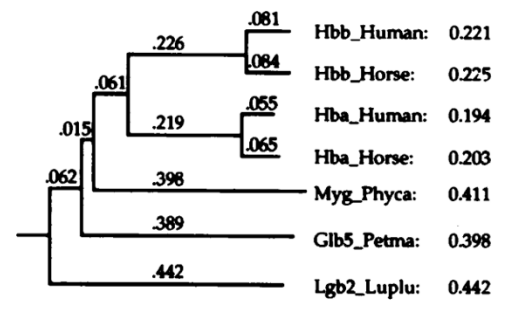
Pairwise alignment:  
Calculate distance matrix

Hbb_Human	1	-					
Hbb_Horse	2	.17	-				
Hba_Human	3	.59	.60	-			
Hba_Horse	4	.59	.59	.13	-		
Myg_Phyca	5	.77	.77	.75	.75	-	
Glb5_Petma	6	.81	.82	.73	.74	.80	-
Lgb2_Luplu	7	.87	.86	.86	.88	.93	.90
		1	2	3	4	5	6

Unrooted Neighbor-Joining tree



Rooted NJ tree (guide tree)  
and sequence weights





X = GTCA  
Y = AGGA  
Z = GTCAA

① pairwise alignment

X: -GTCA  
Y: AGG-A

(-1)

X: GTCA-  
Z: GTCAA

(3)

Y: AGG-A-  
Z: -GTCAA

(-2)

Score?

$$m = \begin{cases} +1 & \text{if same} \\ -1 & \text{if not} \end{cases}$$
$$g = -1$$



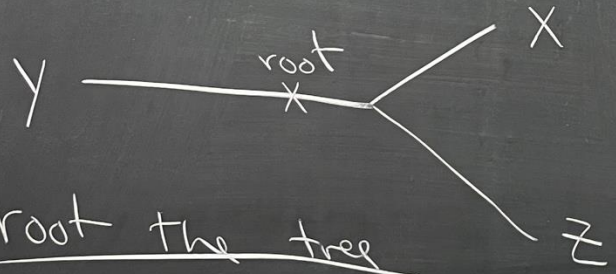
Same  
not

Score  
Matrix

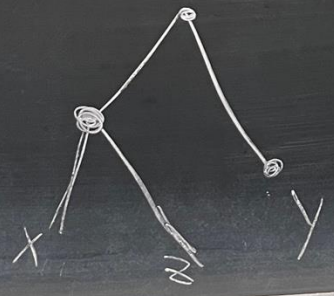
	x	y	z
x	<del>0</del>	-1	3
y	<del>0</del>	<del>0</del>	-2
z	<del>0</del>	<del>0</del>	<del>0</del>

upper  
right  
triangular

② create a tree (unrooted)



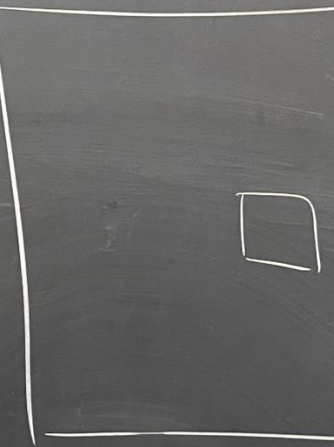
③ root the tree



④ Progressive alignment  
starting from n

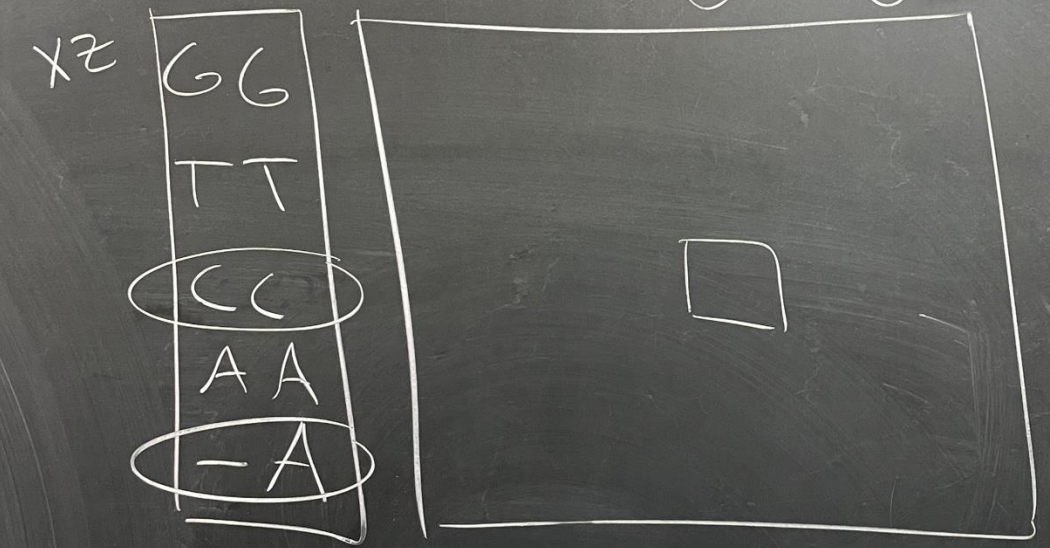
y A G C

xz G G  
T T  
C C  
A A  
- A



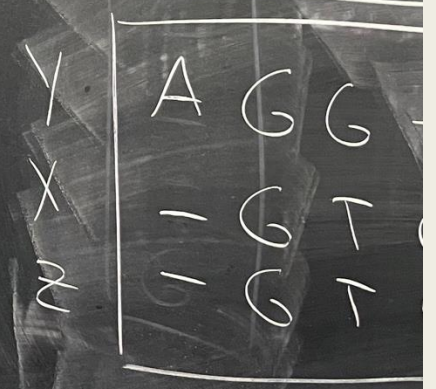
④ progressive alignment  
 starting from most similar seqs

Y A G G A



$$m(C, G) = \frac{m(C, G)}{m(C, C)}$$

$$m(-A, A) = \frac{m(-A, A)}{m(-A, -A)}$$



seqs

pairwise average

$$m(C, G) = \frac{m(C, G) + m(C, G)}{2} = \frac{-2}{2} = -1$$

$$m(-A, A) = \frac{-1 + 1}{2} = 0$$

ancest  
of  
ancestors  
of  
x+z

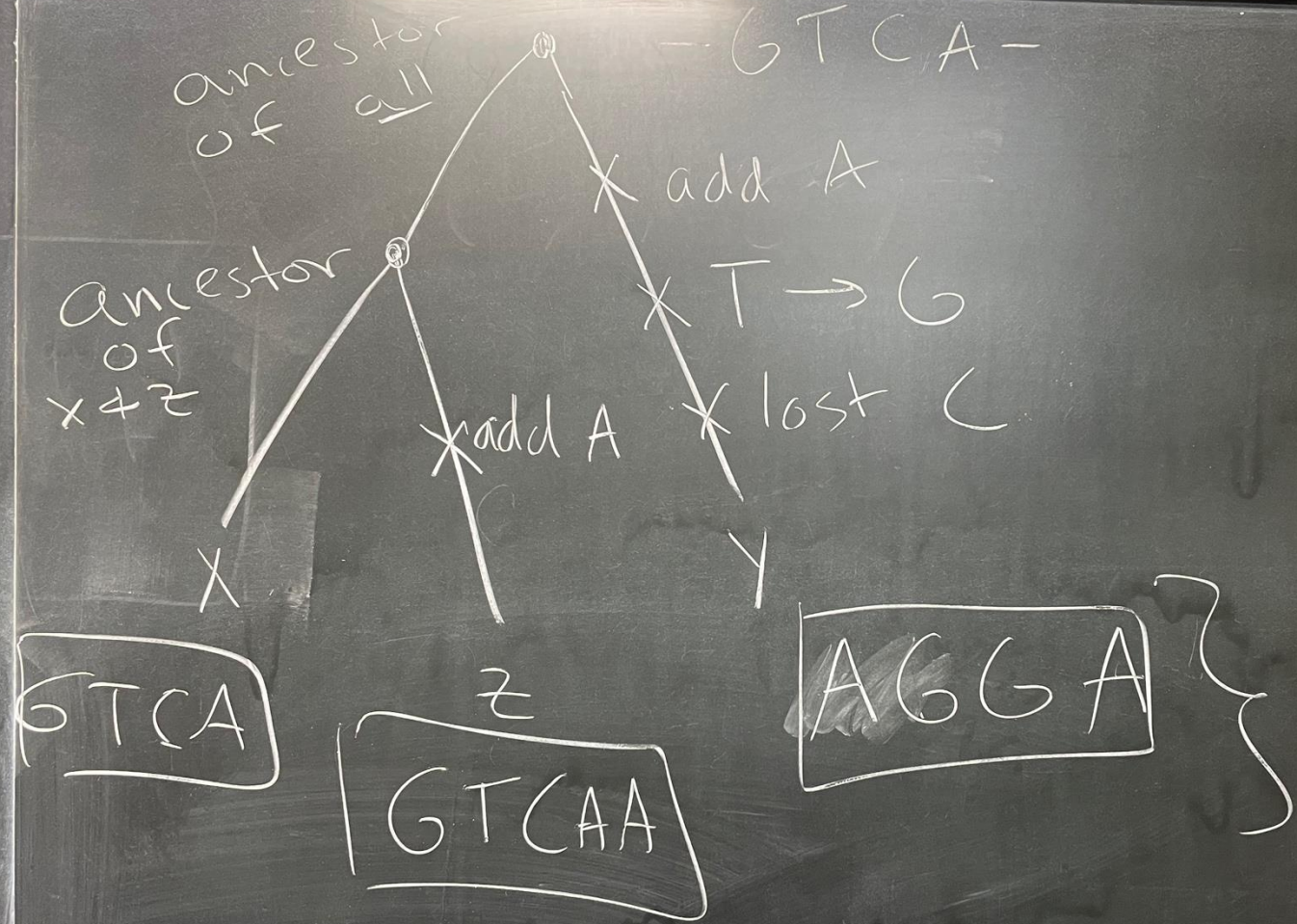
X

GTCA

Y	A	G	G	-	A	-
X	-	G	T	C	A	-
Z	-	G	T	C	A	A

End goal ← MSA

1



present day data

Begin: using variation to reconstruct  
evolutionary events



# VCF file format

**VCF header**

```
##fileformat=VCFv4.0
##fileDate=20100707
##source=VCFtools
##reference=NCBI36
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality (phred score)">
##FORMAT=<ID=GL,Number=3,Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
```

**Mandatory header lines**

**Optional header lines** (meta-data about the annotations in the VCF body)

**Body**

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	SAMPLE1	SAMPLE2
1	1	.	ACG	A,AT	.	PASS	.	GT:DP	1/2:13	0/0:29
1	2	rs1	C	T,CT	.	PASS	H2;AA=T	GT:GQ	0 1:100	2/2:70
1	5	.	A	G	.	PASS	.	GT:GQ	1 0:77	1/1:95
1	100	.	T	<DEL>	.	PASS	SVTYPE=DEL;END=300	GT:GQ:DP	1/1:12:3	0/0:20

**Reference alleles (GT=0)**

**Alternate alleles (GT>0 is an index to the ALT column)**

**Deletion**

**SNP**

**Large SV**

**Insertion**

**Other event**

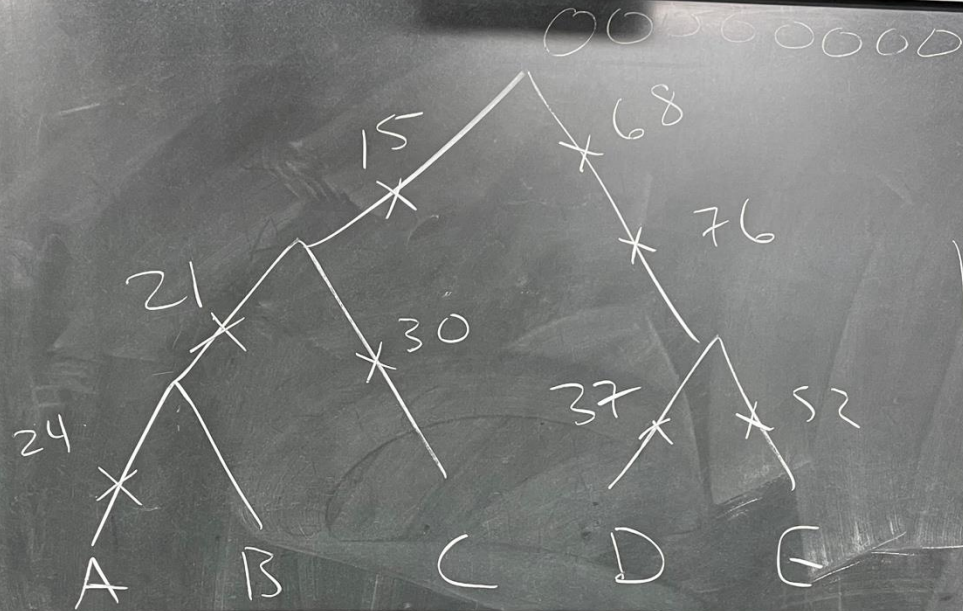
**Phased data** (G and C above are on the same chromosome)

# VCF file format

```

#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT NA18956 NA18999 NA19000 NA19079 NA19074 NA19005 NA1
9085 NA18982 NA18944 NA19055 NA18975 NA19067 NA18949 NA18987 NA18951 NA18994 NA18945 NA18970 NA18964 NA19068 NA18976 NA1894
0 NA19070 NA19012 NA18971 NA18969 NA19063 NA18968 NA19075 NA19082 NA19056 NA19090 NA19083 NA19088 NA18963 NA19081 NA19062 N
A18953 NA18954 NA19084 NA18946 NA18990 NA19057 NA19064 NA19076 NA19002 NA19007 NA19089 NA19091 NA18979 NA19058 NA19060 NA18
947 NA18991 NA18989 NA18984 NA18943 NA18981 NA18986 NA18967 NA19059 NA18998 NA19004 NA19009 NA19066 NA19078 NA19080 NA18974
NA18993 NA18972 NA18977 NA18962 NA18950 NA18948 NA18942 NA18965 NA19054 NA19006 NA19072 NA19001 NA18973 NA18990 NA18966 NA
18910 NA19003 NA18959 NA18952 NA18957 NA18988 NA18978 NA19065 NA18997 NA18939 NA18992 NA18941 NA19077 NA18985 NA18983 NA190
85 NA19087 NA19011 NA18980 NA18961 NA18945
1 10642 rs558604819 G A 100 PASS AC=3;AF=0.00419329;AN=208;NS=2504;DP=1360;EAS_AF=0.003;AMR_
AF=0.0014;AFR_AF=0.0129;EUR_AF=0;SAS_AF=0;AA=.;VT=SNP GT 0|0 0|0 0|0 0|0 0|0 0|0
0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0
0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0
0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0
0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0
0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0
0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0
1 11008 rs575272151 C G 100 PASS AC=3;AF=0.0880591;AN=208;NS=2504;DP=2232;EAS_AF=0.0367;AMR_
AF=0.0965;AFR_AF=0.1346;EUR_AF=0.0885;SAS_AF=0.0716;AA=.;VT=SNP GT 0|0 0|0 0|0 0|0 0|0 0|0
0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0
0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0
0|0 0|0 1|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0
0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0
0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0
0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0
1 11012 rs544419019 C G 100 PASS AC=3;AF=0.0880591;AN=208;NS=2504;DP=2090;EAS_AF=0.0367;AMR_
AF=0.0965;AFR_AF=0.1346;EUR_AF=0.0885;SAS_AF=0.0716;AA=.;VT=SNP GT 0|0 0|0 0|0 0|0 0|0 0|0
0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0
0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0
0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0
0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0
0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0
0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0
0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0
  
```

Handout 8



8 SNPs

	15	20	21	24	30	37	52	68	70
A	1	0	1	1	0	0	0	0	0
B	1	0	1	0	0	0	0	0	0
C	1	0	0	0	1	0	0	0	0
D	0	0	0	0	0	1	0	1	1
E	0	0	0	0	0	0	1	1	1

Haplotype: a sequence from a single chromosome

SNP: single nucleotide polymorphism

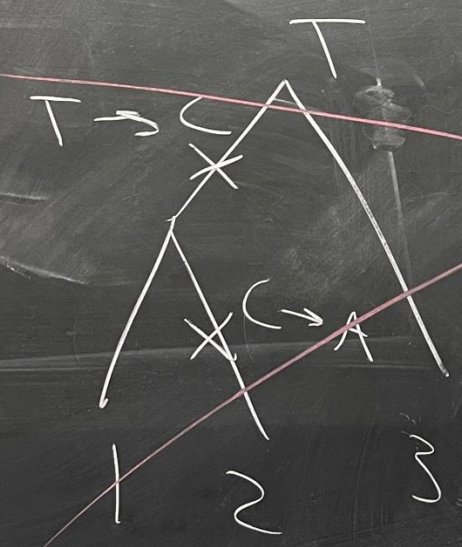
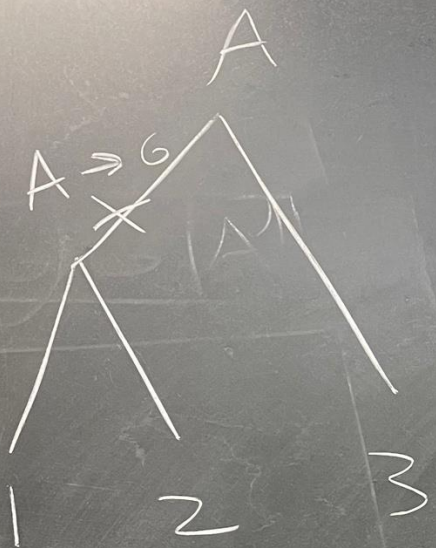
allele: one type of variant at a site

ancestral: allele of the ancestor of all seqs

derived: allele of mutated sequences

reference: "whoever was sequenced first"

alternate: variation discovered later



1	G → 1
2	G → 1
3	A → 0

bi-allelic

1	C → 2
2	A → 1
3	T → 0

ignore  
tri-allelic