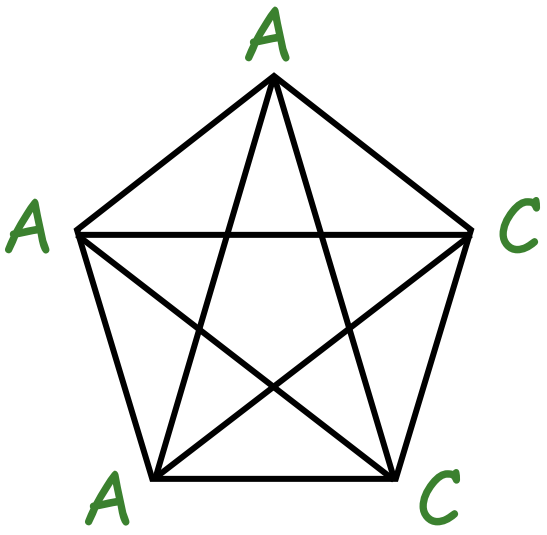
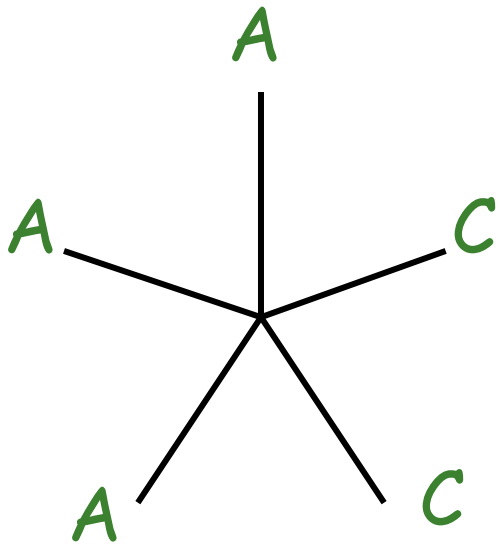
Scoring a multiple alignment

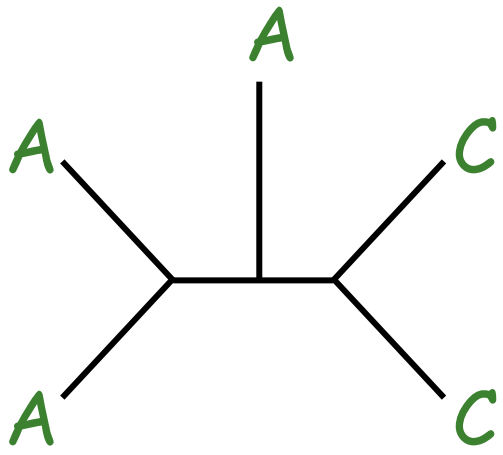
Scoring a multiple alignment



Sum of pairs



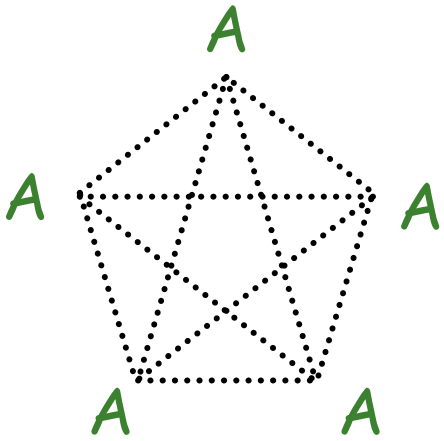
Star

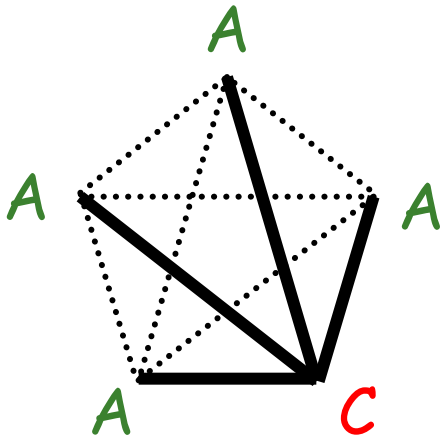


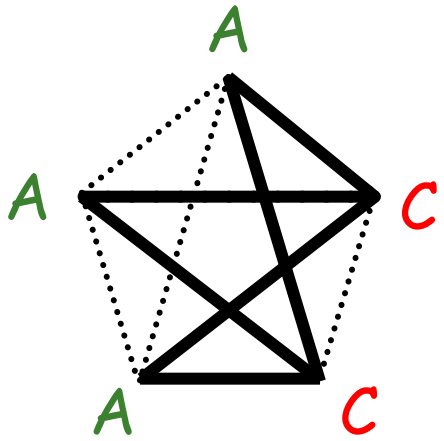
Tree

Sum of Pairs

AAA
AAA
AAA
AAC
ACC





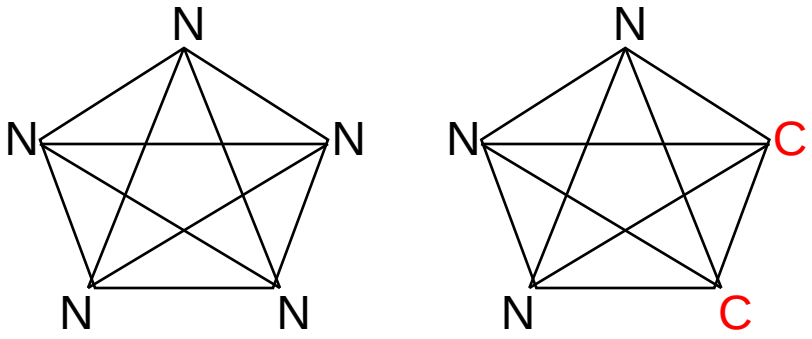


10α + (6α + 4β) + (4α + 6β)

= 20α + 10β

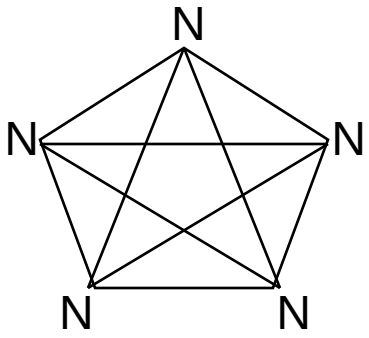
Sum of pairs score (SP score)

Seq	Column-A -B
1	……N……N……
2	……N……N……
3	……N……N……
4	……N……C……
5	……N……C……



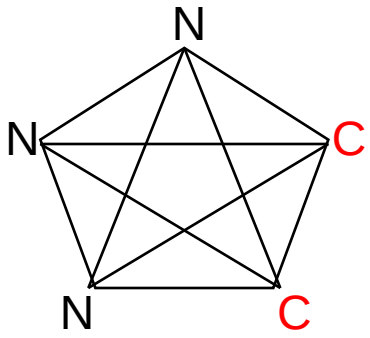
Sum of pairs score (SP score)

Seq	Column-A -B
1	……N……N……
2	……N……N……
3	……N……N……
4	……N……C……
5	……N……C……



Score= 10 * S(N,N)

= 10 * 6 = 60



Score= 3 * S(N,N) + 6 * S(N,C) + S(C,C)

= 3 * 6 + 6 * (-3) + 9 = 9

(BLOSUM62)

Problem: over-estimation of the mutation costs (assuming each sequence is the ancestor of itself; requires a weighting scheme

Sum-of-Pairs Scoring Function

Score of multiple alignment

$$= \sum_{i<j} \text{score}(S_i, S_j)$$

where

$\text{score}(S_i, S_j)$ = score of induced pairwise alignment

Induced Pairwise Alignment

```
S1   S - T I S C T G - S - N I
S2   L - T I - C N G S S - N I
S3   L R T I S C S G F S Q N I
```

Induced pairwise alignment of S_1, S_2:

```
S1   S T I S C T G - S N I
S2   L T I - C N G S S N I
```

Star alignment

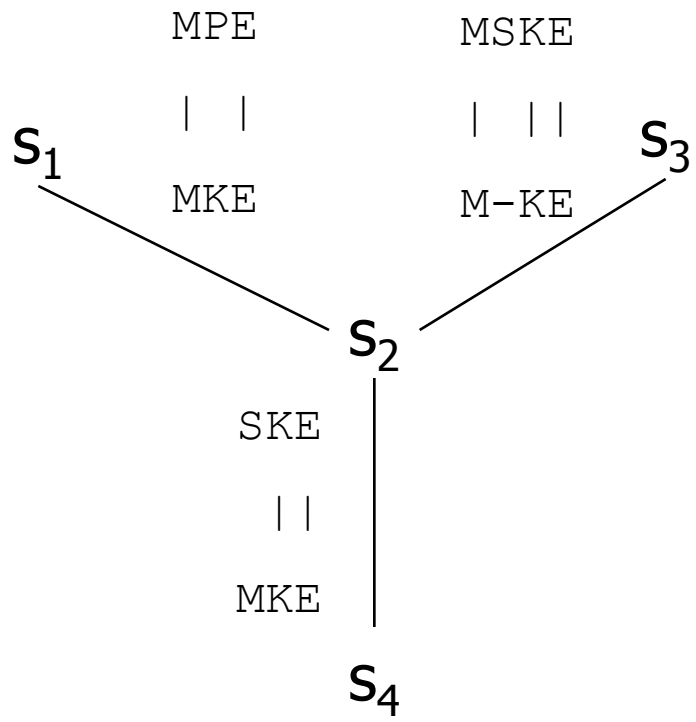
- Heuristic method for multiple sequence alignments
- Select a sequence *c* as the center of the star
- For each sequence x_1, …, x_k such that index $i \neq c$, perform a Needleman-Wunsch global alignment
- Aggregate alignments with the principle “once a gap, always a gap.”

Choosing a center

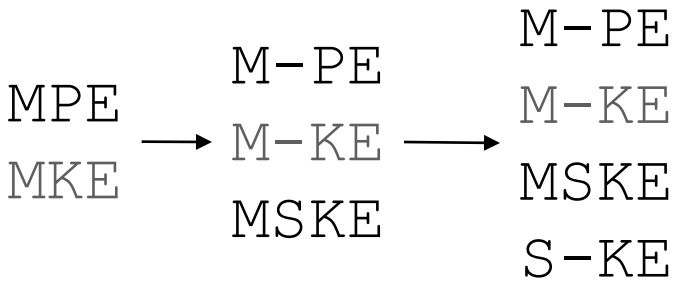
- Try them all and pick the one which is most similar to all of the sequences
- Let $S(x_i,x_j)$ be the optimal score between sequences x_i and x_j.
- Calculate all $O(k^2)$ alignments, and choose as x_c the sequence x_i that maximizes the following

$$\sum_{j \neq i} S(x_i,x_j)$$

Star alignment example



S_1: MPE
S_2: MKE
S_3: MSKE
S_4: SKE



Scoring multiple alignments

- Ideally, a scoring scheme should
  - Penalize variations in conserved positions higher
  - Relate sequences by a phylogenetic tree
    - Tree alignment
  - Quality computation
    - Entropy-based scoring
      - Compute the Shannon entropy of each column
    - Sum-of-pairs (SP) score

Multiple Alignments: Scoring

- Number of matches (multiple longest common subsequence score)
- Entropy score
- Sum of pairs (SP-Score)

Multiple LCS Score

- A column is a “match” if all the letters in the column are the same

AAA
AAA
AAT
ATC

- Only good for very similar sequences

Entropy

- Define frequencies for the occurrence of each letter in each column of multiple alignment
  - $p_A = 1$, $p_T=p_G=p_C=0$ (1st column)
  - $p_A = 0.75$, $p_T = 0.25$, $p_G=p_C=0$ (2nd column)
  - $p_A = 0.50$, $p_T = 0.25$, $p_C=0.25$ $p_G=0$ (3rd column)
- Compute entropy of each column

AAA
AAA
AAT
ATC

$$-\sum_{X=A,T,G,C} p_X \log p_X$$

Entropy: Example

$$S = -\sum_{X=A,T,G,C} p_X \log_{\frac{1}{Nx}} p_X$$

$$entropy\begin{pmatrix} A \\ A \\ A \\ A \end{pmatrix} = 0$$

Best case

Worst case

$$entropy\begin{pmatrix} A \\ T \\ G \\ C \end{pmatrix} = -\sum \frac{1}{4} \log \frac{1}{4} = -4(\frac{1}{4} * -2) = 2$$

Multiple Alignment: Entropy Score

Entropy for a multiple alignment is the sum of entropies of its columns:

$$\sum_{\text{over all columns}} -\sum_{X=A,T,G,C} p_X \log p_X$$

Entropy of an Alignment: Example

column entropy:

$$-(p_A \log p_A + p_C \log p_C + p_G \log p_G + p_T \log p_T)$$

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

Entropy of an Alignment: Example

column entropy:

$$-(p_A \log p_A + p_C \log p_C + p_G \log p_G + p_T \log p_T)$$

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

- Column 1 = -[1*log(1) + 0*log0 + 0*log0 +0*log0]
  = 0
- Column 2 = -[($^1/_4$)*log($^1/_4$) + ($^3/_4$)*log($^3/_4$) + 0*log0 + 0*log0]
  = -[($^1/_4$)*(-2) + ($^3/_4$)*(-.415)] = +0.811
- Column 3 = -[($^1/_4$)*log($^1/_4$)+($^1/_4$)*log($^1/_4$)+($^1/_4$)*log($^1/_4$) +($^1/_4$)*log($^1/_4$)]
  = 4* -[($^1/_4$)*(-2)] = +2.0
- Alignment Entropy = 0 + 0.811 + 2.0 = +2.811

Multiple Alignment Induces Pairwise Alignments

Every multiple alignment induces pairwise alignments

```
x:  AC-GCGG-C
y:  AC-GC-GAG
z:  GCCGC-GAG
```

Induces:

```
x: ACGCGG-C;   x: AC-GCGG-C;   y: AC-GCGAG
y: ACGC-GAC;   z: GCCGC-GAG;   z: GCCGCGAG
```

Sum of Pairs (SP) Scoring

- SP scoring is the standard method for scoring multiple sequence alignments.
- Columns are scored by a ‘sum of pairs’ function using a substitution matrix (PAM or BLOSUM)
- Assumes statistical independence for the columns, does not use a phylogenetic tree.

Computing SP-Score

Aligning 4 sequences: 6 pairwise alignments

Given a_1, a_2, a_3, a_4:

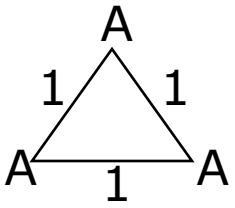
$$s(a_1 \ldots a_4) = \Sigma s^*(a_i, a_j) = s^*(a_1, a_2) + s^*(a_1, a_3) + s^*(a_1, a_4) + s^*(a_2, a_3) + s^*(a_2, a_4) + s^*(a_3, a_4)$$

SP-Score: Example

a_1 ATG–C–AAT
. A–G–CATAT
a_k ATCCCATTT

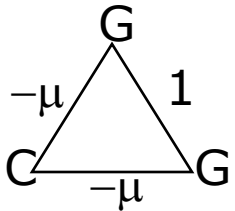
$$S(a_1 \ldots a_k) = \sum_{i,j} S^*(a_i, a_j) \longleftarrow \binom{n}{2} \text{ Pairs of Sequences}$$

May also calculate the scores column by column:



Column 1

Score=3



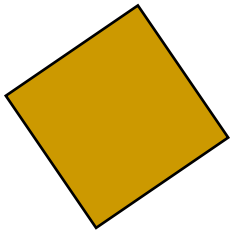
Column 3

Score = $1 - 2\mu$

Example

- Compute Sum of Pairs Score of the following multiple alignment with match = 3, mismatch = -1, S(X,-) = -1, S(-,-) = 0

```
X: G T A C G
Y: T G C C G
Z: C G G C C
W: C G G A C
```

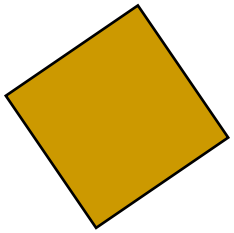


Example

- Compute Sum of Pairs Score of the following multiple alignment with match = 3, mismatch = -1, S(X,-) = -1, S(-,-) = 0

```
X: G T A C G
Y: T G C C G
Z: C G G C C
W: C G G A C
```

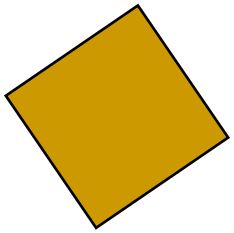
−2



Example

- Compute Sum of Pairs Score of the following multiple alignment with match = 3, mismatch = -1, S(X,-) = -1, S(-,-) = 0

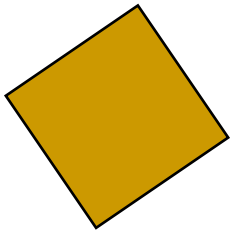
```
X: G T A C G
Y: T G C C G
Z: C G G C C
W: C G G A C
   -2 6
```



Example

- Compute Sum of Pairs Score of the following multiple alignment with match = 3, mismatch = -1, S(X,-) = -1, S(-,-) = 0

```
X: G T A C G
Y: T G C C G
Z: C G G C C
W: C G G A C
   -2 6-2 6 2
```



Sum of pairs = -2+6-2+6+2 = 10

Multiple alignment tools

- Clustal W (Thompson, 1994)
  - Most popular
- PRRP (Gotoh, 1993)
- HMMT (Eddy, 1995)
- DIALIGN (Morgenstern, 1998)
- T-Coffee (Notredame, 2000)
- MUSCLE (Edgar, 2004)
- Align-m (Walle, 2004)
- PROBCONS (Do, 2004)

Table 1. Some recent and less recent available methods for MSAs.

Name	Algorithm	URL
MSA	Exact	http://www.ibc.wustl.edu/ibc/msa.html
DCA	Exact (requires MSA)	http://bibiserv.techfak.uni-biefield.de/dca
OMA	Iterative DCA	http://bibiserv.techfak.uni-biefield.de/oma
ClustalW, ClustalX	Progressive	ftp://ftp-igbmc.u-strasbg.fr/pub/clustalW or clustalX
MultAlin	Progressive	http://www.toulouse.inra.fr/multalin.html
DiAlign	Consistency-based	http://www.gsf.de/biodv/dialign.html
ComAlign	Consistency-based	http://www.daimi.au.df/~ ocaprani
T-Coffee	Consistency-based/progressive	http://igs-server.cnrs-mrs.fr/~ cnotred
Praline	Iterative/progressive	jhering@nimr.mrc.ac.uk
IterAlign	Iterative	http://giotto.Stanford.edu/~ luciano/iteralign.html
Prrp	Iterative/Stochastic	ftp://ftp.genome.ad.jp/pub/genome/saitama-cc/
SAM	Iterative/Stochastic/HMM	rph@cse.ucsc.edu
HMMER	Iterative/Stochastic/HMM	http://hmmer.wustl.edu/
SAGA	Iterative/Stochastic/GA	http://igs-server.cnrs-mrs.fr/~ cnotred
GA	Iterative/Stochastic/GA	czhang@watnow.uwaterloo.ca

from: C. Notredame, “Recent progresses in multiple alignment: a survey”, *Pharmacogenomics* (2002) 3(1)

Useful links

http://cnx.org/content/m11036/latest/

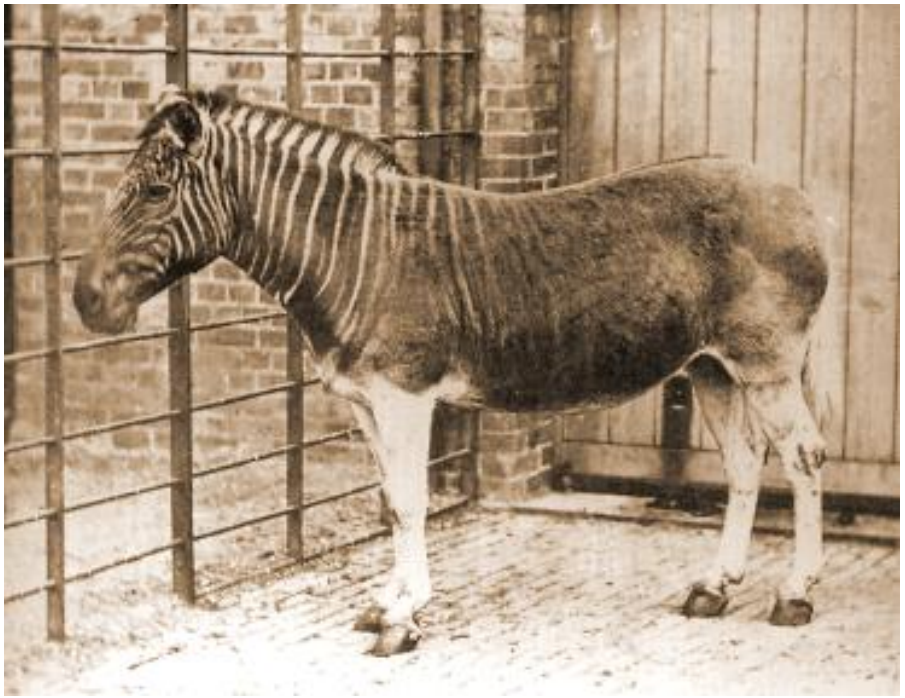
http://www.biokemi.uu.se/Utbildning/Exercises/ClustalX/index.shtm

http://bioinformatics.weizmann.ac.il/~pietro/Making_and_using_protein_MA/

http://homepage.usask.ca/~ctl271/857/paper1_overview.shtml

http://journal-ci.csse.monash.edu.au/ci/vol04/mulali/mulali.html

Was the quagga (now extinct) more like a zebra or a horse?



The quagga was an African animal that is now extinct. It looked partly like a horse, and partly like a zebra. In 1872, the last living quagga was photographed (above). Mitochondrial DNA was obtained from a museum specimen of a quagga and sequenced. Perform a multiple sequence alignment of quagga (*Equus quagga boehmi*), horse (*Equus caballus*), and zebra (*Equus burchelli*) mitochondrial DNA. To which animal was the quagga more closely related?

- For the Entrez search *Equus quagga boehmi*, there is only one mitochondrial DNA sequence (mitochondrial D-loop, AF499309). From Entrez nucleotide the accession numbers are AY246194 (horse) and AF499309 (zebra). The sequences are:
- >gi|29650801|gb|AY246194.1| Equus caballus haplotype be1 mitochondrial D-loop, complete sequence
GATTTCTTCCCCTAAACGACAACAATTTACCCTCATGTGCTATGTCAGTATCAGATTATACCCCCACATA
ACACCATACCCACCTGACATGCAATATCTTATGAATGGCCTATGTACGTCGTGCATTAAATTGTCTGCCC
CATGAATAATAAGCATGTACATAATATCATTTATCTTACATAAGTACATTATATTATTGATCGTGCATAC
CCCATCCAAGTCAAATCATTTCCAGTCAACACGCATATCACAGCCCATGTTCCACGAGCTTAATCACCAA
GCCGCGGGAAATCAGCAACCCTCCCAACTACGTGTCCCAATCCTCGCTCCGGGCCCATCCAAACGTGG
GG
GTTTCTACAATGAAACTATACCTGGCATCTGGTTCTTTCTTCAGGGCCATTCCCACCCAACCTCGCCCAT
TCTTTCCCCTTAAATAAGACATCTCGATGGACTAATGACTAATCAGCCCATGCTCACACATAACTGTGAT
TTCATGCATTTGGTATCTTTTTATATTTGGGGATGCTATGACTCAGCTATGGCCGTCAAAGGCCTCGACG
CAGTCAATTAAATTGAAGCTGGACTTAAATTGAACGTTATTCCTCCGCATCAGCAACCATAAGGTGTTAT
TCAGTCCATGGTAGCGGGACATAGGAAACAAGTGCACCTGTGCACCTACCCGCGCAGTAAGCAAGTAAT
A
TAGCTTTCTTAATCAAACCCCCCCTACCCCCCATTAAACTCCACATATGTACATTCAACACAATCTTGCC
AAACCCCAAAAACAAGACTAAACAATGCACAATACTTCATGAAGCTTAACCCTCGCATGCCAACCATAAT
AACTCAACACACCTAACAATCTTAACAGAACTTTCCCCCCGCCATTAATACCAACATGCTACTTTAATCA
ATAAAATTTCCATAGACAGGCATCCCCCTAGATCTAATTTTCTAAATCTGTCAACCCTTCTTCCCC

- >gi|20335096|gb|AF499310.1| Equus burchellii isolate Be1 mitochondrial D-loop, partial sequence
GCTCCACCGTCAACACCCAAAGCTGAAATTCTACTTAAACTATTCCTTGATTTCCTCCCCTAAACGACAA
CAATTCACCCTCATGTACTATGTCAGTATTAAAATACATCCTATGTAGCATTATACAGTTCAACATATAA
TACCCTGTTAACATCCTATGTACATCGTGCATTAAATTGTT

- >gi|20335095|gb|AF499309.1| Equus quagga boehmi isolate Bo1 mitochondrial D-loop, partial sequence
GCTCCACCGTCAACACCCAAAGCTGAAATTCTACTTAAACTATTCCTTGATTTCCTCCCCTAAACGACAA
CAGTTCACCCTCATGTACTATGTCAGTATTAAAATACATCCTATGTAGTATTATACAGTTCAACATATAA
TACCCTGTTAACATCCTATGTACGTCGTGCATTAGATTGTT

The portion of the multiple sequence alignment in (excluding additional nonoverlapping horse sequence) is as follows:

```
gi|20335096|gb|AF499310.1| GCTCCACCGTCAACACCCAAAGCTGAAATTCTACTTAAACTATTCCTTGA 50 [zebra]
gi|20335095|gb|AF499309.1| GCTCCACCGTCAACACCCAAAGCTGAAATTCTACTTAAACTATTCCTTGA 50 [quagga]
gi|29650801|gb|AY246194.1| ----------------------------------------------GA 2  [horse]
                                                              **

gi|20335096|gb|AF499310.1| TTTCCTCCCCTAAACGACAACAATTCACCCTCATGTACTATGTCAGTATT 100
gi|20335095|gb|AF499309.1| TTTCCTCCCCTAAACGACAACAGTTCACCCTCATGTACTATGTCAGTATT 100
gi|29650801|gb|AY246194.1| TTTCTTCCCCTAAACGACAACAATTTACCCTCATGTGCTATGTCAGTATC 52
                           **** ***************** ** ********** ************

gi|20335096|gb|AF499310.1| AAAATACATCCT-ATGTAGCATTATACA-GTTCAACATATAATACCCTGT 148
gi|20335095|gb|AF499309.1| AAAATACATCCT-ATGTAGTATTATACA-GTTCAACATATAATACCCTGT 148
gi|29650801|gb|AY246194.1| AGATTATACCCCCACATAACACCATACCCACCTGACATGCAATATCTTAT 102
                           * * ** * **  *  **  *  ****       ****  **** * * *

gi|20335096|gb|AF499310.1| TAACATCCTATGTACATCGTGCATTAAATTGTT----------------- 181
gi|20335095|gb|AF499309.1| TAACATCCTATGTACGTCGTGCATTAGATTGTT----------------- 181
gi|29650801|gb|AY246194.1| GAATGGCCTATGTACGTCGTGCATTAAATTGTCTGCCCCATGAATAATAA 152
                            **   ********* ********** *****
```

Visual inspection of the alignment can give you a clue that the quagga DNA is closely related to zebra DNA: [1] the internal gaps in the alignment suggest that horse DNA (on the bottom row) is an outlier, and [2] positions that are not conserved (i.e. positions lacking an asterisk) also consistently show that horse DNA differs while quagga and zebra sequences match each other. The pairwise alignment scores also show clearly that the quagga is closer to a zebra:

Sequence 1: gi|29650801|gb|AY246194.1| 976 bp [horse]
Sequence 2: gi|20335096|gb|AF499310.1| 181 bp [zebra]
Sequence 3: gi|20335095|gb|AF499309.1| 181 bp [quagga]
Start of Pairwise alignments
Aligning...
Sequences (1:2) Aligned. Score: 56
Sequences (1:3) Aligned. Score: 55
Sequences (2:3) Aligned. Score: 97

Finally a PubMed search with the terms quagga, horse and zebra links to an article suggesting that zebra and quagga shared a common ancestor several million years ago.

Methods of Sequence Alignment

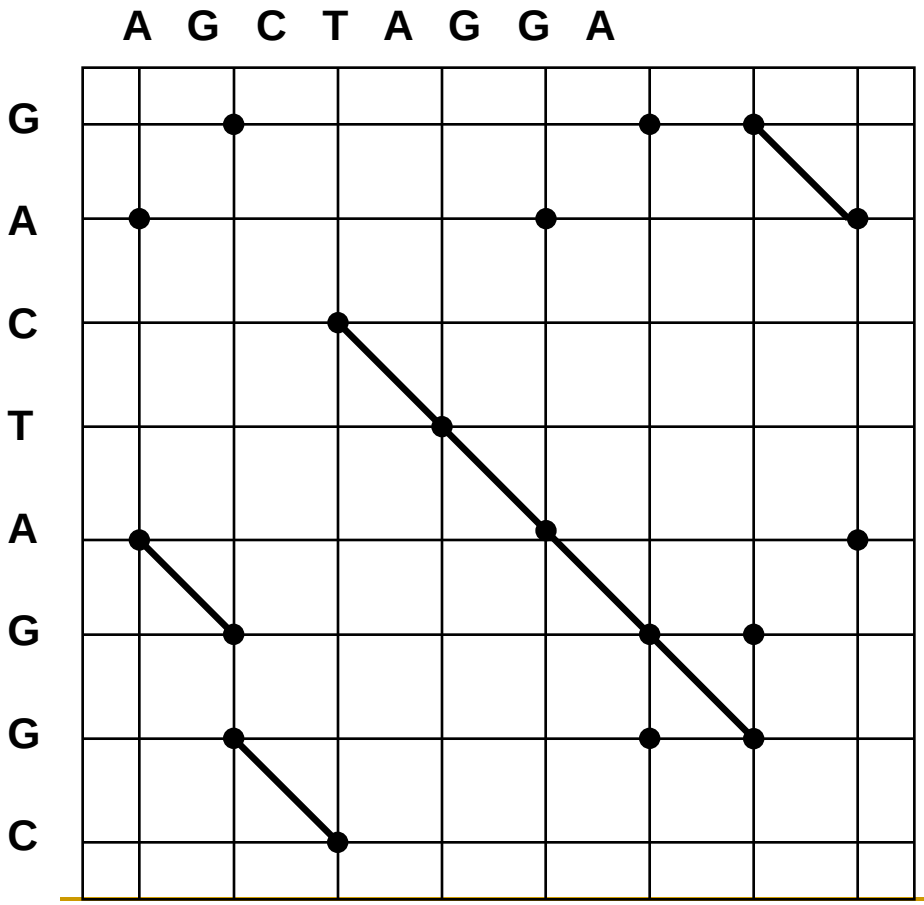
(1) Dot Matrix analysis

(2) Dynamic Programming (or DP)

(3) Multiple Sequence Alignment

The Dot Matrix

- established in 1970 by A.J. Gibbs and G.A.McIntyre
- method for comparing two amino acid or nucleotide sequences



- each sequence builds one axis of the grid one puts a dot, at the intersection of same letters appearing in both sequences
- scan the graph for a series of dots
  - reveals similarity
  - or a string of same characters
- longer sequences can also be compared on a single page, by using smaller dots

The Dot Matrix

- When to use the Dot Matrix method?
  - unless the sequences are known to be very much alike
- limits of the Dot Matrix
  - doesn't readily resolve similarity that is interrupted by insertion or deletions
  - Difficult to find the best possible alignment (optimal alignment)
  - most computer programs don't show an actual alignment

summary

Sequence Alignment

An alignment is a mutual arrangement of two sequences.

- It exhibits where the two sequences are similar, and where they differ.
- An 'optimal' alignment is one that exhibits the most correspondences, and the least differences.
- Sequences that are similar probably have the same function