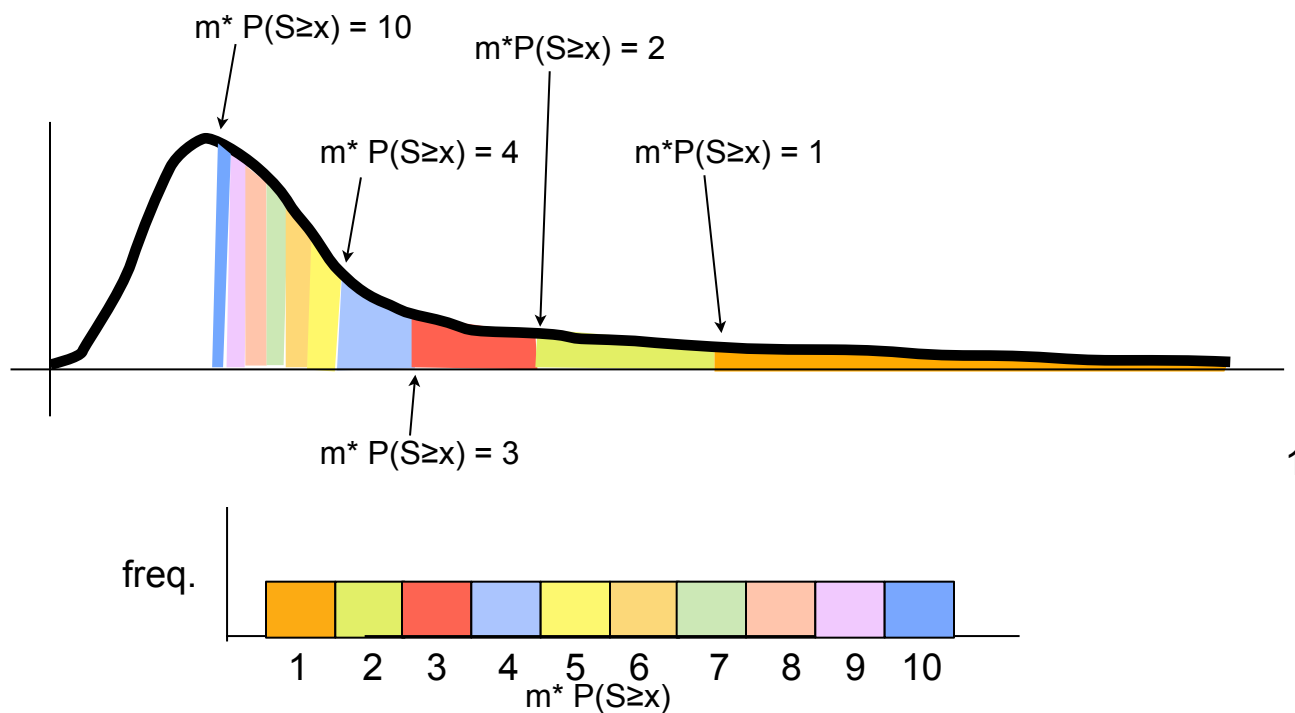


pop-quiz explanation

- At the end of lecture 7, I presented a pop quiz. How many random scores have e-values less than or equal to 10? (Answer? 10) Why?
- Consider this random score distribution. Each zone marked has the same area, right?
- If you randomly pick from this distribution (think about darts), what does the distribution zones (e-values) look like? (it's flat!)



Bioinformatics 1-- lecture 8

Multiple sequence alignment

```
Ra 1  CCCCAGGGTGGTGGCTGGGGCAG
Rb 2  CCTCATGGTGGTGGCTGGGGCAA
Rc 3  CCCCATGGTGGCGGCTGGGGACAG
Rd 4  CCCCATGGTGGCGGATGGGGACAG
Re 5  CCTCATGGTGGCGGCTGGGGTCAA
```

```
Ra 1  CCCCAGGGTGGTGGCTGGGGCAG
Rb 2  CCTCATGGTGGTGGCTGGGGCAA
Rc 3  CCCCATGGTGGCGGCTGGGGACAG
Rd 3  CCCCATGGTGGCGGCTGGGGACAG
Re 4' CCCCATGGTGGTGGCTGGGGACAG
Rf 5  CCTCATGGTGGCGGCTGGGGTCAA
```

In class ~~competition~~ exercise:
Editing a multiple sequence alignment in
UGENE

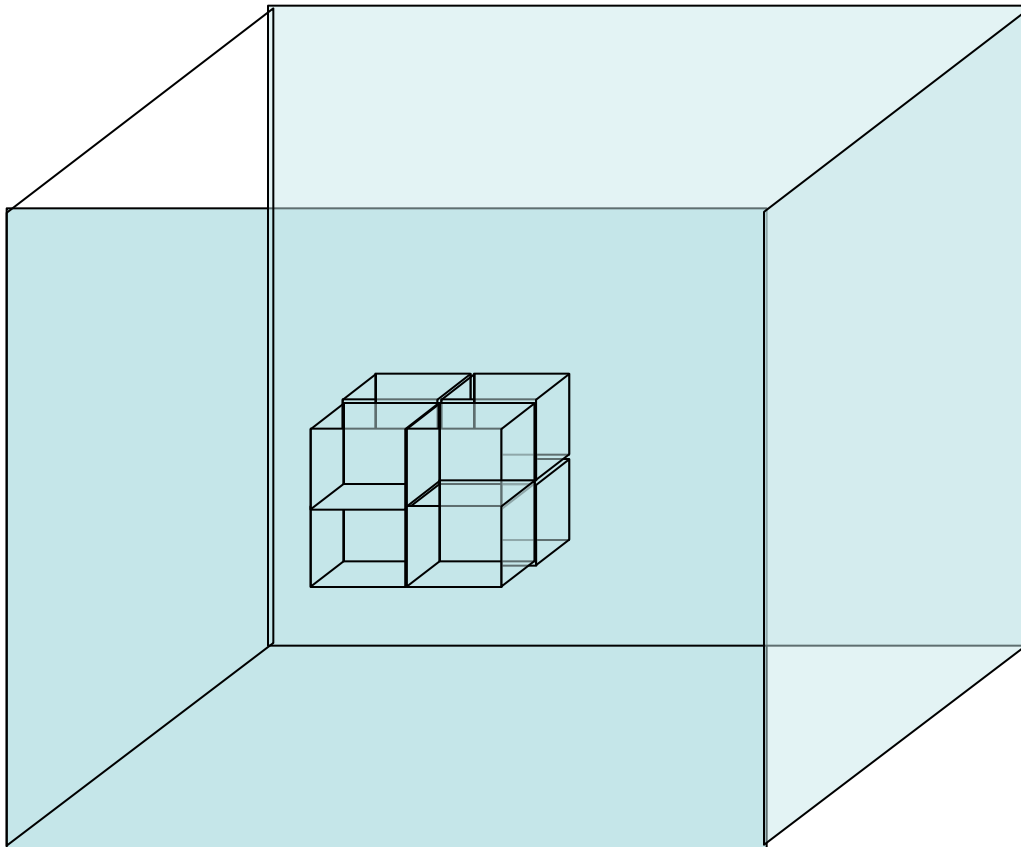
- Download and open “bad alignment” from the course web page
- Export all sequences as alignment.
- Edit the alignment.
- Try to improve the %identity, and consolidate gaps.

Methods for multiple sequence alignment

- Dynamic programming
- Star
- Progressive
- ClustalW
- Muscle

- Is optimality possible? Can we do Dynamic Programming for three or more sequences?

A 3D alignment matrix...



$$S(i,j,k) = \text{MAX} \{ \\ A(i-1,j-1,k-1)+S(i,j,k), \\ A(i-1,j,k)\text{-gap}, \\ A(i,j-1,k)\text{-gap}, \\ A(i,j,k-1)\text{-gap}, \\ A(i-1,j-1,k)\text{-gap}, \\ A(i-1,j,k-1)\text{-gap}, \\ A(i,j-1,k-1)\text{-gap} \}$$

Multiple sequence alignment -- Progressive method

1. align all pairs
2. pairwise align two most similar
3. add the next most similar
4. continue until all sequences are aligned

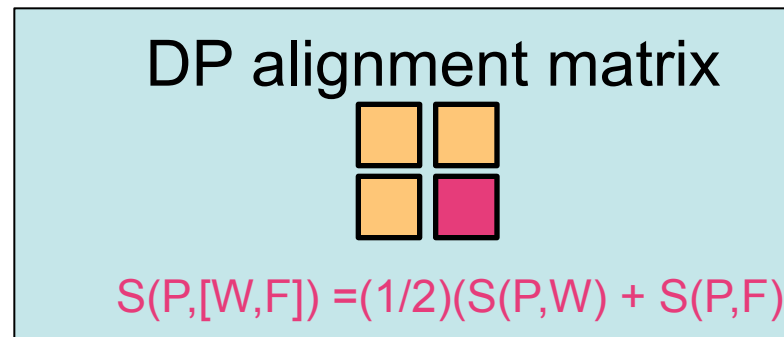


Current alignment {

A	G	H	I	.	W	W	P	F
A	G	H	I	I	F	W	P	Y

sequence to add

A
W
P
Y



distance and similarity are interconvertible

Maximizing similarity and Minimizing distance are equivalent if

- $d(i,j) + s(i,j) = s_{\max}$,

where s_{\max} is the maximum possible similarity, and the minimum distance is $d=0$. For each position in the alignment.

- Distance based on identity score (p-distance)

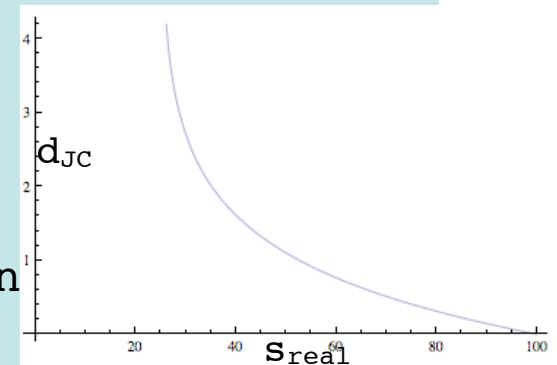
$$d = 100 - \%identity$$

- Distance using empirical J-C correction

$$d_{JC} = -\ln((S_{\text{real}} - S_{\text{rand}}) / (S_{\text{ident}} - S_{\text{rand}}))$$

where S_{ident} = score of an identity alignment, and S_{rand} = mode score of a false alignment.

- For proteins, $S_{\text{rand}} \approx 25\%$. "Twilight zone" (R. Doolittle, 1986)



progressive alignment

Making a guide tree

Neighbor-joining algorithm:

	A	B	C	D	E	F
A		97	81	82	59	32
B			77	80	55	31
C				90	65	40
D					61	42
E						33
F						

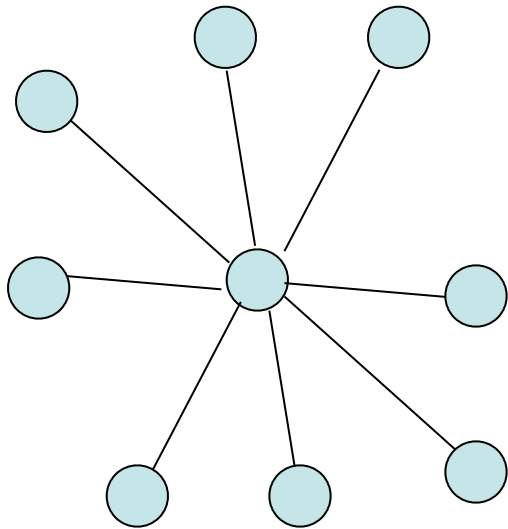
A
B
C
D
E
F

Draw guide tree here

Fill in J-C distances.

Star, using all-to-one distances

- “Star” alignment, all sequences are aligned to one.
- No guide tree. Not progressive.
- BLAST does this.



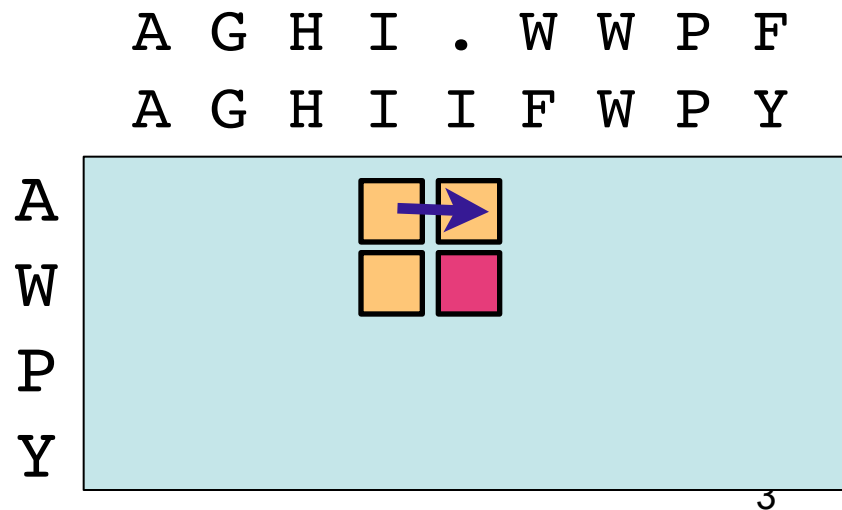
CLUSTALW

JD Thompson, DG Higgins, TJ Gibson - Nucleic acids research, 1994

- Start with unrooted tree, using Neighbor joining.
- choose root to get guide tree
- progressive alignment
 - matches are scored using sequence weights
 - gaps are position dependent
 - GOP lower for polar residues
 - GOP zero where there is already a gap

<http://www.ebi.ac.uk/Tools/msa/clustalw2/>

No gap penalty for aligning a gap to a gap



$$A(i,j) = A(i-1,j) - \text{gap}(i)$$

If i is already a gap position in any sequence, set $\text{gap}_1(i)=0$.

CLUSTALW Position specific gap penalty

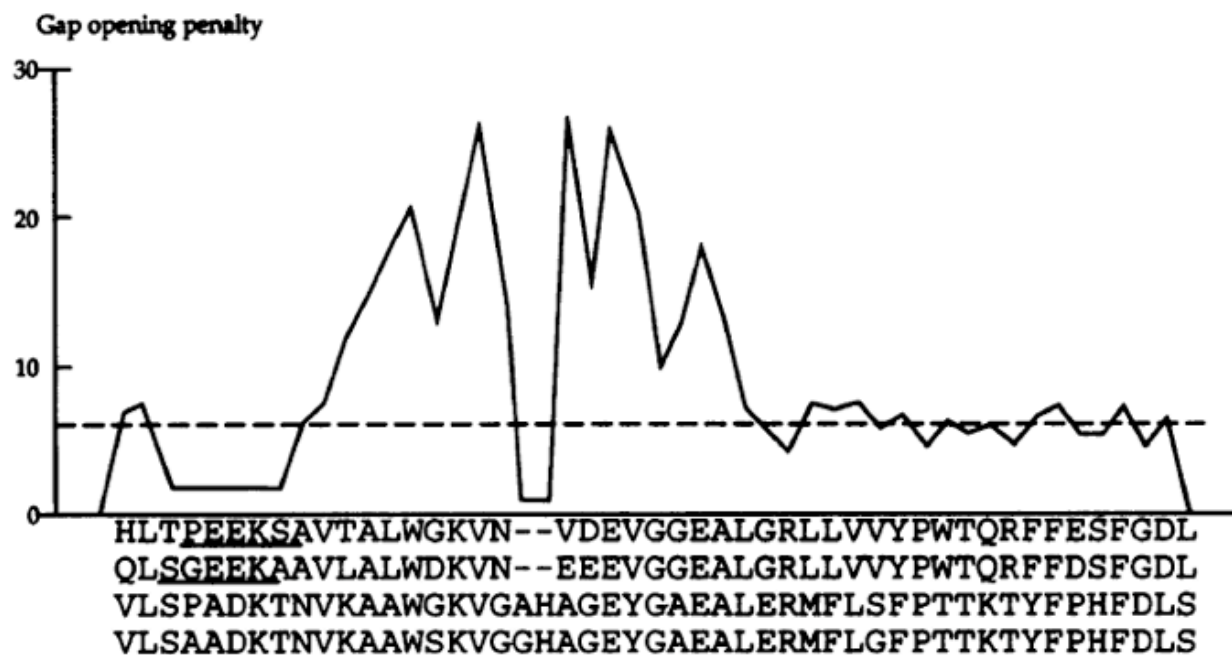


Figure 3. The variation in local gap opening penalty is plotted for a section of alignment. The initial gap opening penalty is indicated by a dotted line. Two hydrophilic stretches are underlined. The lowest penalties correspond to the ends of the alignment, the hydrophilic stretches and the two positions with gaps. The highest values are within 8 residues of the two gap positions. The rest of the variation is caused by the residue specific gap penalties (12).

MUSCLE

RC Edgar - Nucleic acids research, 2004

- Iterative MSA

- k-mer distance matrix
- UPGMA tree
- **progressive alignment**--> MSA1
- Kimura distances from MSA1
- UPGMA tree
- **progressive alignment** -->MSA2
- For all tree branches:
 - split tree into two
 - calculate profiles
 - align profiles
 - accept or reject the alignment.
 - Repeat

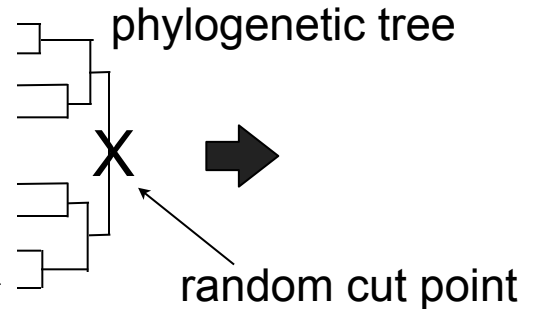
based on short identical matches

Z&B p174

MUSCLE iterative alignment

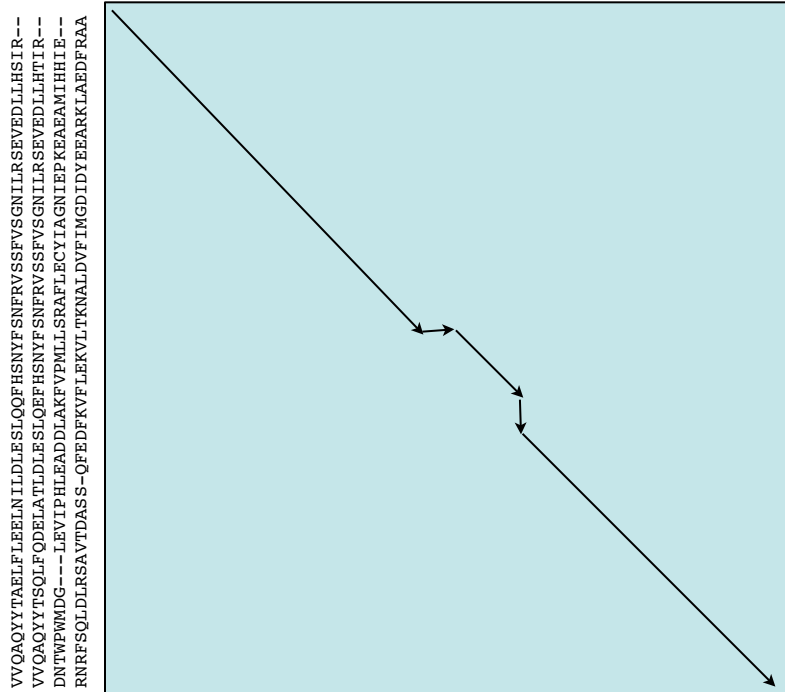
```

XP_001615335 YEPTDKEMDDILSAYFFYPSYKDYTRYVVDIFHRNYVSIFIYGNIAMPTEKEDENATS--
XP_002259219 YDPTDKEMDDLAYSAYFFYPSYKDYTKYVVDFFHRNYVSIFIYGNIAMTTEKENENATS--
XP_001347897 YTPTNKEMYDILNAYFFYPSYNAYRTYVNEYFLRNYVVIFIYGNIIISDLKGEENITKNN
XP_726635      YIPTNKEIYDILNAYLFYPLYSYIKYINNFPHKNYINIFIYGNLSIPNEINIKNETN--
XP_671449      -----
XP_001458064 VVQAQYYTAEFLLEELNILDLESLOQFHSNYFSNFRVSSFFVSGNILRSEVEDLLHSIR--
XP_001347129 VVQAQYYTSQLFQDELATLDLESLOQEFHSNYFSNFRVSSFFVSGNILRSEVEDLLHTIR--
XP_002283970 DNTWPWMDG---LEVIPHLEADDLAKFVPMLLSRAFLECYIAGNIEPKEAEAMIHHE--
XP_002367832 RNRFSQLDLRSVTDASS-QFEDFKVFLKVLTKNALDVFIMGDIYEEARKLAEDFRAA
    
```



```

YEPTDKEMDDILSAYFFYPSYKDYTRYVVDIFHRNYVSIFIYGNIAMPTEKEDENATS--
YDPTDKEMDDLAYSAYFFYPSYKDYTKYVVDFFHRNYVSIFIYGNIAMTTEKENENATS--
YTPTNKEMYDILNAYFFYPSYNAYRTYVNEYFLRNYVVIFIYGNIIISDLKGEENITKNN
YIPTNKEIYDILNAYLFYPLYSYIKYINNFPHKNYINIFIYGNLSIPNEINIKNETN--
    
```



```

YEPTDKEMDDILSAYFFYPSYKDYTRYVVDIFHRNYV .SIFIYGNIAMPTEKEDENATS--
YDPTDKEMDDLAYSAYFFYPSYKDYTKYVVDFFHRNYV .SIFIYGNIAMTTEKENENATS--
YTPTNKEMYDILNAYFFYPSYNAYRTYVNEYFLRNYV .FIYGNIIISDLKGEENITKNN
YIPTNKEIYDILNAYLFYPLYSYIKYINNFPHKNYI .NIFIYGNLSIPNEINIKNETN--
VVQAQYYTAEFLLEELNILDLESLOQFHS .NYFSNFRVSSFFVSGNILRSEVEDLLHSIR--
VVQAQYYTSQLFQDELATLDLESLOQEFHS .NYFSNFRVSSFFVSGNILRSEVEDLLHTIR--
DNTWPWMDG---LEVIPHLEADDLAKFVP .MLLSRAFLECYIAGNIEPKEAEAMIHHE--
RNRFSQLDLRSVTDASS-QFEDFKVFLKVLTKNALDVFIMGDIYEEARKLAEDFRAA
    
```

new MSA

In each iteration:
 The phylogenetic tree is cut at a random branch, the two subtrees are converted to profiles, and aligned. The new alignment is either accepted or rejected