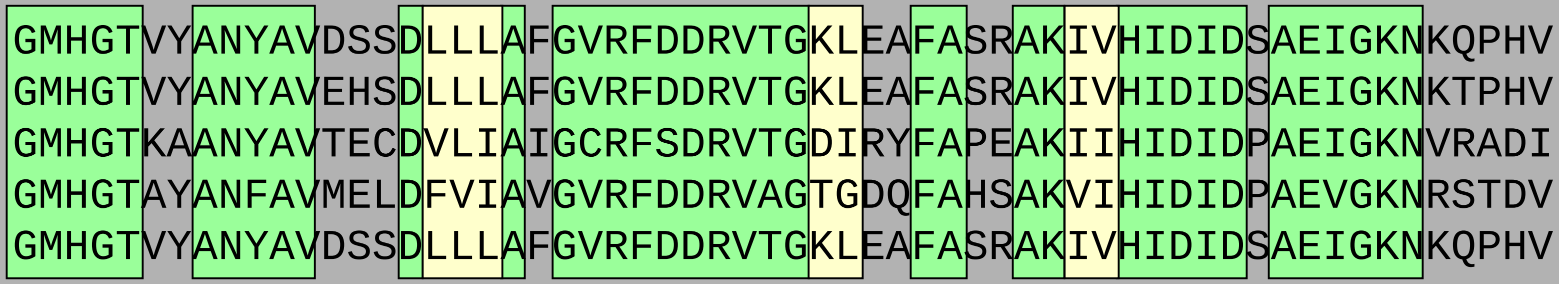
Multiple Sequence Alignment

Mark Whitsitt - NCSA

What is a Multiple Sequence Alignment (MA)?



```
GMHGTVYANYAVDSSDLLLAFGVRFDDRVTGKLEAFASRAKIVHIDIDSAEIGKNKQPHV
GMHGTVYANYAVEHSDLLLAFGVRFDDRVTGKLEAFASRAKIVHIDIDSAEIGKNKTPHV
GMHGTKAANYAVTECDVLIAIGCRFSDRVTGDIRYFAPEAKIIHIDIDPAEIGKNVRADI
GMHGTAYANFAVMELDFVIAVGVRFDDRVAGTGDQFAHSAKVIHIDIDPAEVGKNRSTDV
GMHGTVYANYAVDSSDLLLAFGVRFDDRVTGKLEAFASRAKIVHIDIDSAEIGKNKQPHV
```

A simple textual representation

The Significance of MA

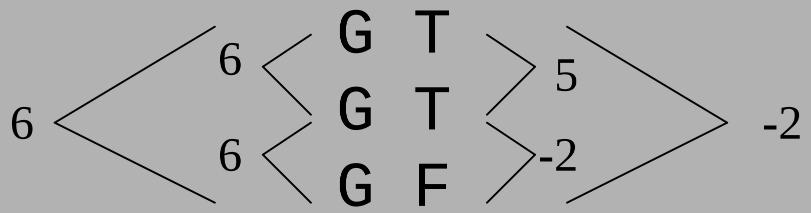
- Valuable for both DNA and Protein
  - Family Identification and Representation
  - Residue Conservation and Acceptable Amino Acid Substitutions
  - Evolutionary Relationships- Phylogenetic Trees
  - Structural Alignment (Modeling)
- Each may use a different graphical representation of the multiple alignment.

Methods - How do we do this?

- Generalized Dynamic Programming
  - Expand to use *N*-dimensions
  - Becomes intractable for more than a few sequences or large sequences.
  - “Optimal” Alignment?
- Progressive Alignments
  - Heuristic - use “guide tree” based on pairwise similarities
  - No “optimal” alignment, but biologically sound

Scoring Multiple Alignments -
Sum of Pairs

6 + 6 + 6 = 18



6

6

6

G T

G T

G F

5

-2

-2

5 - 2 - 2 = 1

18 + 1 = 19

Blosum62
s(G,G) = 6
s(T,T) = 5
s(T,F) = -2

Sum of Pairs (SP) scoring ignores the assumption of a common ancestor for related sequences.

Compensate by using a weighting scheme:
- reduce the weight of score for closely related pairs
- increase the weight of score for more divergent pairs

w(1,2) = 0.3
w(2,3) = 0.8
w(1,3) = 1.2

Generalized Dynamic Programming

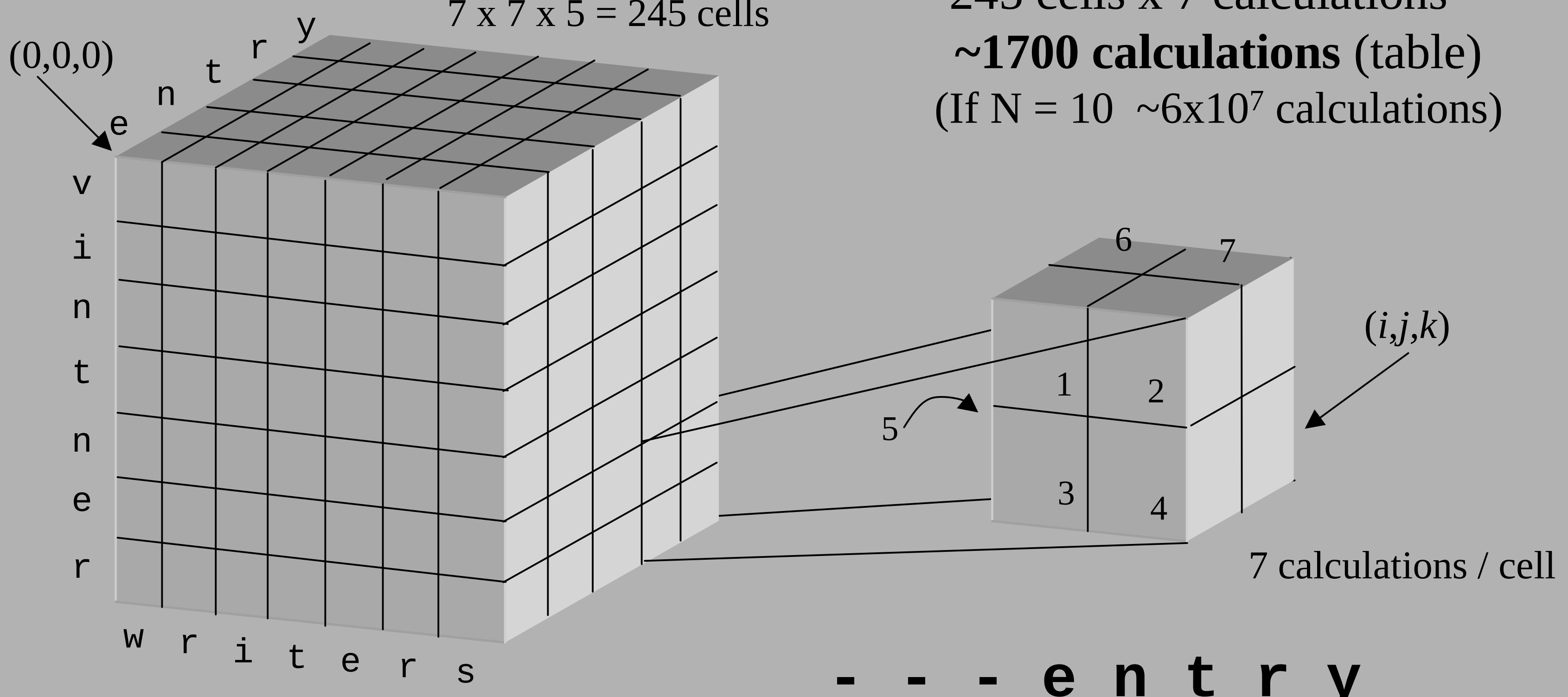
$D(i,j)$	S_2		w	r	i	t	e	r	s
S_1		0	1	2	3	4	5	6	7
	0	0	← 1	← 2	← 3	← 4	← 5	← 6	← 7
v	1	↑ 1	↖ 1	↖ ← 2	↖ ← 3	↖ ← 4	↖ ← 5	↖ ← 6	↖ ← 7
i	2	↑ 2	↖↑ 2	↖ 2	↖ 2	← 3	← 4	← 5	← 6
n	3	↑ 3	↖↑ 3	↖↑ 3	↖↑ 3	↖ 3	↖ ← 4	↖ ← 5	↖ ← 6
t	4	↑ 4	↖↑ 4	↖↑ 4	↖↑ 4	↖ 3	↖ ← 4	↖ ← 5	↖ ← 6
n	5	↑ 5	↖↑ 5	↖↑ 5	↖↑ 5	↑ 4	↖ 4	↖ ← 5	↖ ← 6
e	6	↑ 6	↖↑ 6	↖↑ 6	↖↑ 6	↑ 5	↖ 4	↖ ← 5	↖ ← 6
r	7	↑ 7	↖↑ 7	↖ 6	↖↑ ← 7	↑ 6	↑ 5	↖ 4	← 5

```
w r i t - e r s
v i n t n e r -
```

Pairwise Dynamic Programming Array

N-Dimensional Dynamic Programming - MSA (Gupta,Kececioglu, Schäffer 1995, J. Comp. Biol. 2:459-472)

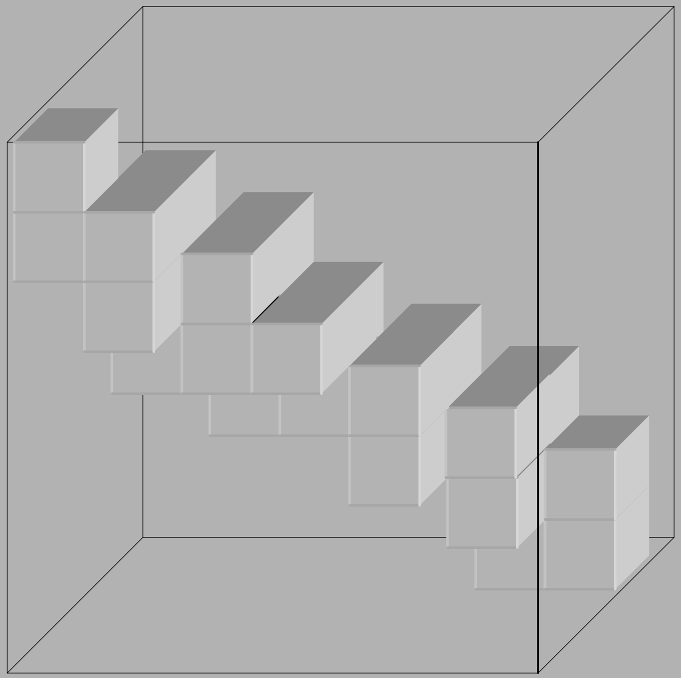
245 cells x 7 calculations =
~1700 calculations (table)
(If N = 10 ~6x10^7 calculations)



```
- - - e n t r y
v i n t n e r -
- w r i t e r s
```

N = 3 Sequences

Carillo-Lipman Lower Bound



Algorithm calculates a lower bound to the SP score of the alignment.

Effectively identifies cells that will not contribute to the proper alignment. (“reduces the search space”)

Speeds up the DP calculations.

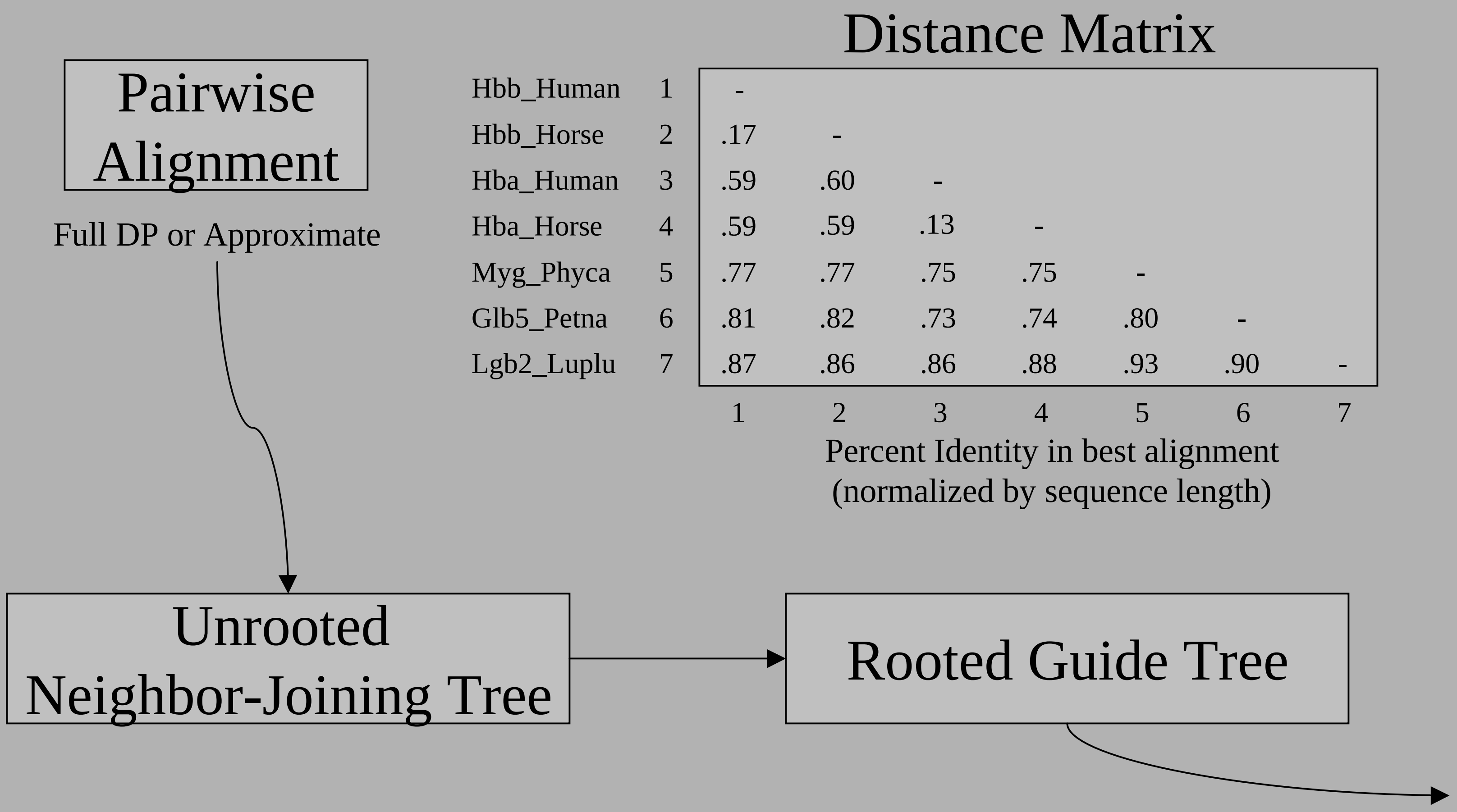
MSA (the program) estimate of the bound is usually high.Cannot guarantee an “optimal” alignment.

Still not practical - Only useful for 5-7 sequences.

Progressive Multiple Sequence Alignment - Clustal W (Thompson, Higgins, and Gibson 1994, Nucl. Acids Res. 22:4673-4680)

- Uses a series of “pairwise” global alignments to complete alignment.
- Useful for almost any number of sequences.
- Alignments guided by a “tree” calculated from pairwise similarity scores for all sequence pairs.
- Other Features:
  - Sequence weighting from initial guide tree.
  - Scoring Matrix selection by relative similarity.
  - Gap penalty modification - residue sensitive and position specific.

Establishing a Guide Tree



Pairwise Alignment

Full DP or Approximate

Distance Matrix

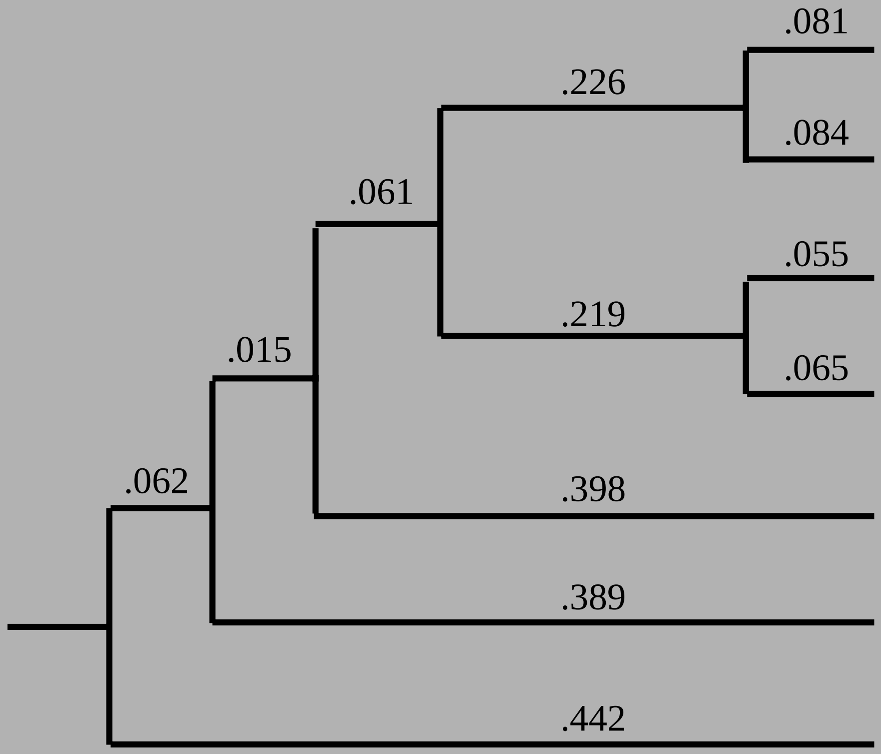
		1	2	3	4	5	6	7
Hbb_Human	1	-						
Hbb_Horse	2	.17	-					
Hba_Human	3	.59	.60	-				
Hba_Horse	4	.59	.59	.13	-			
Myg_Phyca	5	.77	.77	.75	.75	-		
Glb5_Petna	6	.81	.82	.73	.74	.80	-	
Lgb2_Luplu	7	.87	.86	.86	.88	.93	.90	-

Percent Identity in best alignment
(normalized by sequence length)

Unrooted Neighbor-Joining Tree

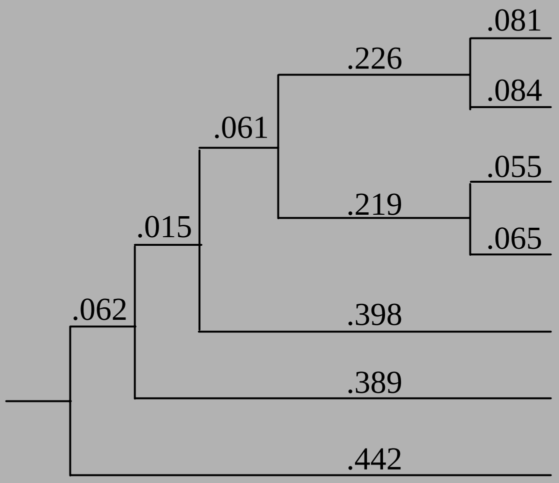
Rooted Guide Tree

The Guide Tree



Sequence	Weight
Hbb_Human:	0.221
Hbb_Horse:	0.225
Hba_Human:	0.194
Hba_Horse:	0.203
Myg_Phyca:	0.411
Glb5_Petna:	0.398
Lgb2_Luplu:	0.442

Using the Tree



Sequence	Weight
Hbb_Human:	0.221
Hbb_Horse:	0.225
Hba_Human:	0.194
Hba_Horse:	0.203
Myg_Phyca:	0.411
Glb5_Petna:	0.398
Lgb2_Luplu:	0.442

Weighted Sum of Pairs

$$
\begin{aligned}
S(i) = {} & s(a,b) * w1 * w2 \\
& + s(a,b) * w1 * w5 \\
& + s(a,b) * w1 * w3 \\
& + s(a,b) * w1 * w4 \\
& \quad \cdot \\
& \quad \cdot \\
& \quad \cdot \\
& + s(a,b) * w6 * w7 / 21
\end{aligned}
$$

- **Local Minimum Problem**: Early mistakes cannot be corrected.
- Very similar sequences easily and reliably aligned.
- Begin aligning most similar sequences first and progressively add remaining sequences or groups of sequences. (profiles)
- Good guide tree required, but doesn’t have greatest impact on results.

Parameter choices - Scoring Matrix

- Scoring Matrices vs. Sequence Divergence
- Clustal W automatically employs different matrices for different levels of sequence divergence (hard vs. soft matrices).
- Sequence distances measured from guide tree.
  - PAM: 20 (80-100%), 60 (60-80%), 120 (40-60%), 350 (0-40%)
  - BLOSUM: 80 (80-100%), 62 (60-80%), 45 (30-60%), 30 (0-30%)

Parameter Choices - Gap Penalties

- Rules for biologically reasonable gap placement.
- “Once a gap, always a gap” - takes advantage of reliability in alignment of similar sequences first.
- Increase gap opening penalty adjacent to existing gaps. (empirically 8 a.a.)
- Decrease penalties in stretches of hydrophilic residues. (loops)
- Adjust penalties for each a.a. using observed frequencies of adjacent gaps.

Representations of Multiple Alignments

- Sequence Listing
- Consensus Sequence
- Profiles
- Logos

Sequence Listing

```
KVKAHGKKVLGAFSDGLAHLD-----NLKGTFATLSELHCDKLHVDPENFRL
KVKAHGKKVLHSFGEGVHHLD-----NLKGTFAALSELHCDKLHVDPENFRL
QVKGHGKKVADALTNAVAHVD-----DMPNALSALSDLHAHKLRVDPVNFKL
QVKAHGKKVGDALTLAVGHLD-----DLPGALSNLSDLHAHKLRVDPVNFKL
DLKKHGVTVLTALGAILKKKG-----HHEAELKPLAQSHATKHKIPIKYLEF
DVRWHAERIINAVNDAVASMDDT--EKMSMKLRDLSGKHAKSFQVDPQYFKV
ELQAHAGKVFKLVYEAAIQLQVTGVVVTDATLKNLGSVHVSKGVAD-AHFPV
```

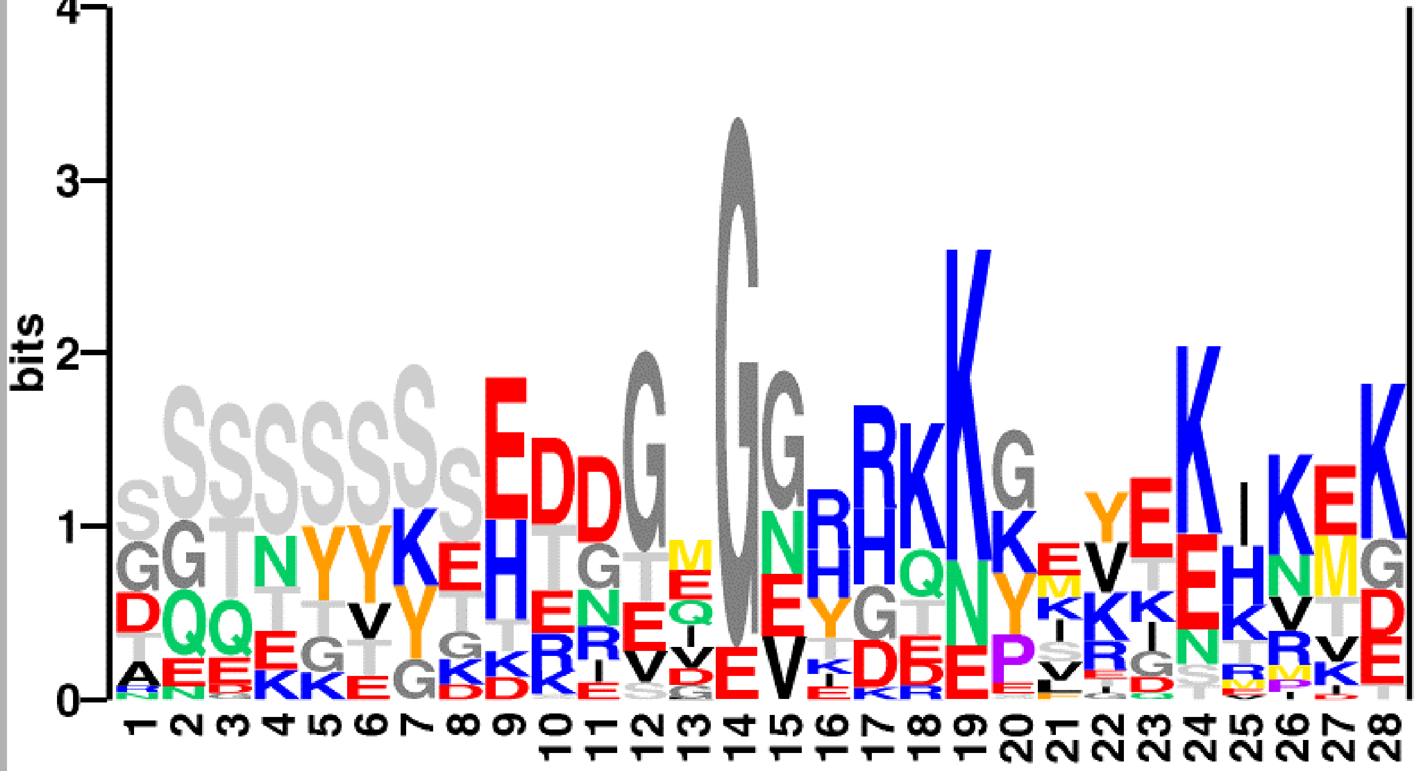
Consensus Sequences

```
KVKAHGKKVLGAFSDGLAHLD-----NLKGTFATLSELHCDKLHVDPENFRL
KVKAHGKKVLHSFGEGVHHLD-----NLKGTFAALSELHCDKLHVDPENFRL
QVKGHGKKVADALTNAVAHVD-----DMPNALSALSDLHAHKLRVDPVNFKL
QVKAHGKKVGDALTLAVGHLD-----DLPGALSNLSDLHAHKLRVDPVNFKL
DLKKHGVTVLTALGAILKKKG-----HHEAELKPLAQSHATKHKIPIKYLEF
DVRWHAERIINAVNDAVASMDDT--EKMSMKLRDLSGKHAKSFQVDPQYFKV
ELQAHAGKVFKLVYEAAIQLQVTGVVVTDATLKNLGSVHVSKGVAD-AHFPV

KVKAHGKKVLDALTDAVAHLDDTGVENLKGTLAALSDLHADKLRVDPENFKL
```

- “Majority rules” or “Winner takes all”
- Often difficult to identify a clear winner.
- Information about observed or acceptable substitutions is lost.
- Other more complex rules, but none comprehensive.

Sequence Logos



PSSM of BL00315A (DEHYDRIN_1;) 26 sequences.

Profiles

	a	**b**	**c**	**-**	**a**
	a	**b**	**a**	**b**	**a**
	a	**c**	**c**	**b**	**-**
	c	**b**	**-**	**b**	**c**
a	.75		.25		.50
b		.75		.75	
c	.25	.25	.50		.25
-			.25	.25	.25

- Column specific residue frequencies at their simplest.

- Used by Clustal W for combining groups of previously aligned sequences.

- Position specific scoring matrix. May be used to search for pattern blocks or in database searches (eg. Blast)

Hidden Markov Models - HMM

- Probabalistic models.
  - Probability of residue *x* occurring at a position.
- Begin with random alignment and iteratively improve using “learned information”
- Can provide information about confidence in specific residue alignments.