

Bioinformatics 1 -- lecture 10

Sequence weights from distances

sub-alignment, re-alignment

Finding conserved differences

BLASTing deeper: phi-blast, psi-blast, transitive blast

Psi-BLAST: Blast with profiles

Psi-BLAST searches the database *iteratively*.

(Cycle 1) Normal BLAST (with gaps)

(Cycle 2) (a) Construct a **profile** from the results of **Cycle 1**.

(b) Search the database using the profile.

(Cycle 3) (a) Construct a **profile** from the results of **Cycle 2**.

(b) Search the database using the profile.

And So On... (user sets the number of cycles)

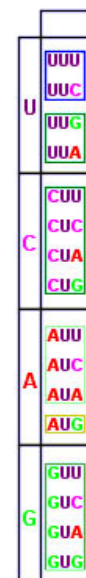
Psi-BLAST is much more *sensitive* than BLAST.

Also more vulnerable to *low-complexity*.

Other forms of BLAST

BLAST	query	database
blastn	nucleotide	nucleotide
blastp	protein	protein
tblastn	protein	translated DNA
blastx	translated DNA	protein
tblastx	translated DNA	translated DNA
psi-blast	protein, profile	protein
phi-blast	pattern	protein
transitive blast*	any	any

*not really a blast. Just a way of using blast.



IUPAC nuc
A
C
G
T (or U)
R
Y
S
W
K
M
B
D
H
V
N
. or -

PHI-BLAST -- Patterned Hit Initiated BLAST

Table I. Detection of subtle protein sequence relationships using PHI-BLAST

Conserved domain or motif under investigation	Pattern ^a	GenBank (30) accession no. of query	Top non-trivial relevant hit found by PHI-BLAST		Top non-trivial relevant hit found by BLAST	
			Accession no.	<i>E</i> -value	Accession no.	<i>E</i> -value
A. P-loop ATPase domain in apoptosis regulators and plant stress response proteins	[GA]xxxxGK[ST]	231729	2213598	0.038	2961373	4.7
B. ATPase domain in mismatch repair protein MutL, type II topoisomerases, histidine kinases, and HS90 molecular chaperones	hxhxDxGxG	127552	488200	0.017	2495364	1.8
C. Nucleotidyltransferase domain in archaeal tRNA nucleotidyltransferases	DhDhhh	2826366	2650333	0.061	2650333	8.6
D. Motif VI of superfamily II helicases in archaeal homologs of bacterial DNA primases	QxxGRx[GA]R	2128723	2499099	0.54		

Distance-based weights

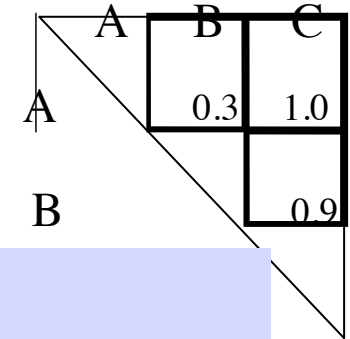
	A	B	C
A		0.3	1.0
B			0.9
C			

- (1) Sum the weighted distances to get new weights.
- (2) Normalize the new weights
- (3) Repeat (1) and (2) until no change.

Pseudocode :

```
all  $w_i$  initialized to 1.  
while ( $w_i \neq w'_i$ ) do  
  for  $i$  from A to C do  
     $w'_i = \sum_j w_j D_{ij}$   
  end do  
  for  $i$  from A to C do  
     $w_i = w'_i / \sum_j w'_j$   
  end do  
end do
```

Distance-based weights



Running the pseudocode :

(1) Sum the weighted distances to get new weights.

$$w'_A = 0.3 + 1.0 = 1.3$$

$$w'_B = 0.3 + 0.9 = 1.2$$

$$w'_C = 1.0 + 0.9 = 1.9$$

(2) Normalize the new weights

$$w_A = 1.3 / (1.3 + 1.2 + 1.9) = 0.30$$

$$w_B = 1.2 / 4.4 = 0.27$$

$$w_C = 1.9 / 4.4 = 0.43$$

(1) Sum the weighted distances to get new weights.

$$w'_A = 0.3 * 0.27 + 1.0 * 0.43 = 0.51$$

$$w'_B = 0.3 * 0.3 + 0.9 * 0.43 = 0.48$$

$$w'_C = 1.0 * 0.3 + 0.9 * 0.27 = 0.54$$

...

$$w_{ABC} = 0.33 \quad 0.31 \quad 0.35$$

$$w_{ABC} = 0.30 \quad 0.28 \quad 0.42$$

$$w_{ABC} = 0.31 \quad 0.29 \quad 0.40$$

$$w_{ABC} = 0.30 \quad 0.28 \quad 0.41$$

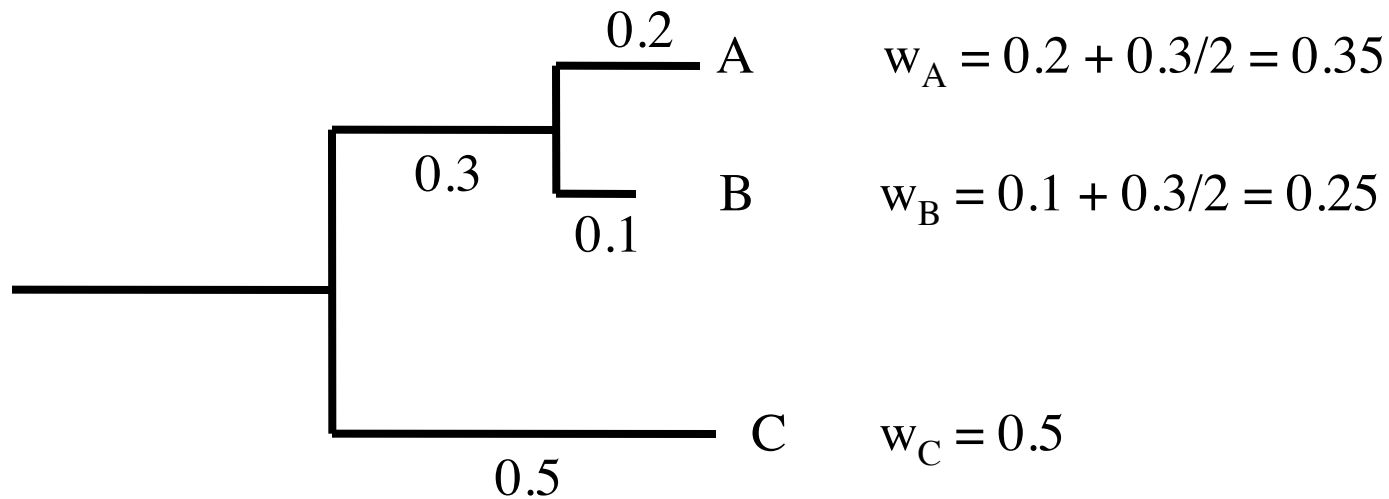
$$w_{ABC} = 0.30 \quad 0.28 \quad 0.41 \text{ converged.}$$

(3) Repeat (1) and (2) until no change.

Exercise:

- Select a MSA or clade with ~10-15 sequences.
- Write pseudocode to ...
 - Calculate the sequence distances using J-C.
 - Calculate the sequence weights (first iteration)
 - Calculate self-consistent weights.

Better weights from a phylogram



The sequence weight is calculated starting from the distance from the taxon to the first ancestor node, adding half of the distance from the first ancestor to the second ancestor, 1/4th of the distance from the second to third ancestor, and so on.

Finally, the weights are normalized to sum to 1.00

Make a MSA

- Search NCBI for “s100” (or your favorite gene). Get the accession number.
- Retrieve the sequence from NCBI protein database in UGENE.
- Do a BLAST search.
- Save hits as alignment. Align using MUSCLE (large alignment node)
- Delete all but one screen worth.
- Align again using MUSCLE (default mode)

Re-aligning using MUSCLE

- Select columns around a gap.
- Align/Align with MUSCLE. Use column range.
- Edit by hand.

Look for Conserved differences

- Select two clusters (clades) within the MSA. (call them clade1, clade2)
- Find positions that are conserved clade1 but not in clade2.
- Find positions that are conserved in clade2 but not in clade1.
- Find positions that are conserved in both clades, but are different.

Review

alignment affine gap aligned scrambled sequences algorithms additivity BLAST
BLOSUM clade ClustalW complexity database searching
data structures databases dynamic programming evolution e-value expected
Erdos-Renyi equation Extreme value distribution evolutionary models genetic drift
FASTA format FASTA algorithm false database hits global
GenBank format heuristic homoplasy iterative refinement Jukes-Cantor
LLR local alignment multiple sequence alignment MUSCLE matrix bias
natural selection null model NCBI normal distribution optimal
p-distance parsimony position-specific gap penalty probability
p-value pairwise sequence alignment phylogram profiles phylogenetic trees
progressive star selective/sensitive synteny
sequence weights semi-global significance SSEARCH
substitution matrix PAM transitions/transversions weights

Things to know