

Bioinformatics I

- Miscellaneous
- Midterm review

Variants of PAM, BLOSUM

When you run an alignment, you get to choose the substitution matrix. How do you know which to use?

larger evolutionary timestep
↑

BLOSUM45	distant homologs, %ID < 50	Distance relationships, gap penalty = lower
BLOSUM62	homologs, <%ID>=62	Moderate similarity, optimal for mixed MSA
BLOSUM90	close homologs, few gaps	Close homologs, gap penalty = high

**Gap penalties optimized against gold-standard alignments (with the same %ID range) such as the alignments in BALiBase

U	UUU
	UUC
	UUG
	UUA
C	CUU
	CUC
	CUA
	CUG
A	AUU
	AUC
	AUA
	AUG
G	GUU
	GUC
	GUA
	GUG

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. A way of using any blast.

3

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

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*.

Also more vulnerable to *low-complexity*.

PHI-BLAST -- Patterned Hit Initiated BLAST

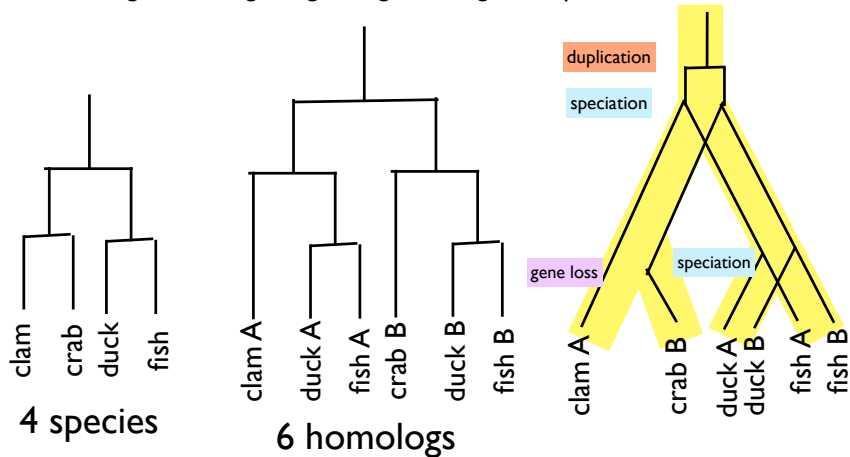
Table 1. Detection of subtle protein sequence relationships using PHI-BLAST

Conserved domain or motif under investigation	Pattern ^a	GenBank (30) accession no. of query	Top non-trivial relevant hit found by PHI-BLAST Accession no. E-value	Top non-trivial relevant hit found by BLAST Accession no. E-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	

5

Orthologs/paralogs

Orthologs: homologs originating from a speciation event
Paralogs: homologs originating from a gene duplication event.



Example parsimony problem

- <http://www.biorecipes.com/CompBioExercises/ParsimonySol.html>

A “regular expression” for an essay question answer.

*/*Thesis statement*

($N > 1$) (**There are** [perhaps, at least, exactly] N [reasons, algorithms, theories, models,...] **that** [explain, justify, allow, model, show,...] **the** [data, observation, goal, mechanism ...] **with regard to** X .) X **can be** [understood explained, ruled out,...] **as** Y **by** [using, applying, realizing,...] Z .

*/*Itemized list*

{First}, ([<explanation>,<observation>,<instruction set>, ...])* n (where $1 \leq n \leq 3$). [Second], ([<explanation>,<observation>,<instruction set>, ...])* n (where $1 \leq n \leq 3$) [Third, To a lesser extent, In the rare case that, ...], ... [Nth]

*/*Thesis restatement*

[Taken together, The most likely conclusion given the aggregate evidence, With the exception of X , Given current theoretical understanding, Given the observed data,...] **we can conclude that** X , [when, if and only if, because] **is** Y **because** Z .

{optional}

[options]

(conditions)

X , Y , Z : clauses.

N , n : integers

alignment affine gap aligned scrambled sequences algorithms additivity BLAST
Bioinformatics class cladogram clade ClustalW complexity DDBJ database searching
data structures databases dynamic programming evolution e-value expected
Erdos-Renyi equation Extreme value distribution evolutionary models genetic drift
FASTA format Fitch-Margoliash 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 NJ normal distribution optimal
p-distance parsimony position-specific gap penalty probability
p-value pairwise sequence alignment phylogram profiles phylogenetic trees
progressive phylogram root software star selective/sensitive syteny
sequence weights semi-global significance SSEARCH
substitution matrix terminology transitions/transversions triangle inequality
ultrametric tree weights

Items for mind map