

Homology: Orthologs and Paralogs

Homology: Similarity in characteristics resulting from shared ancestry.

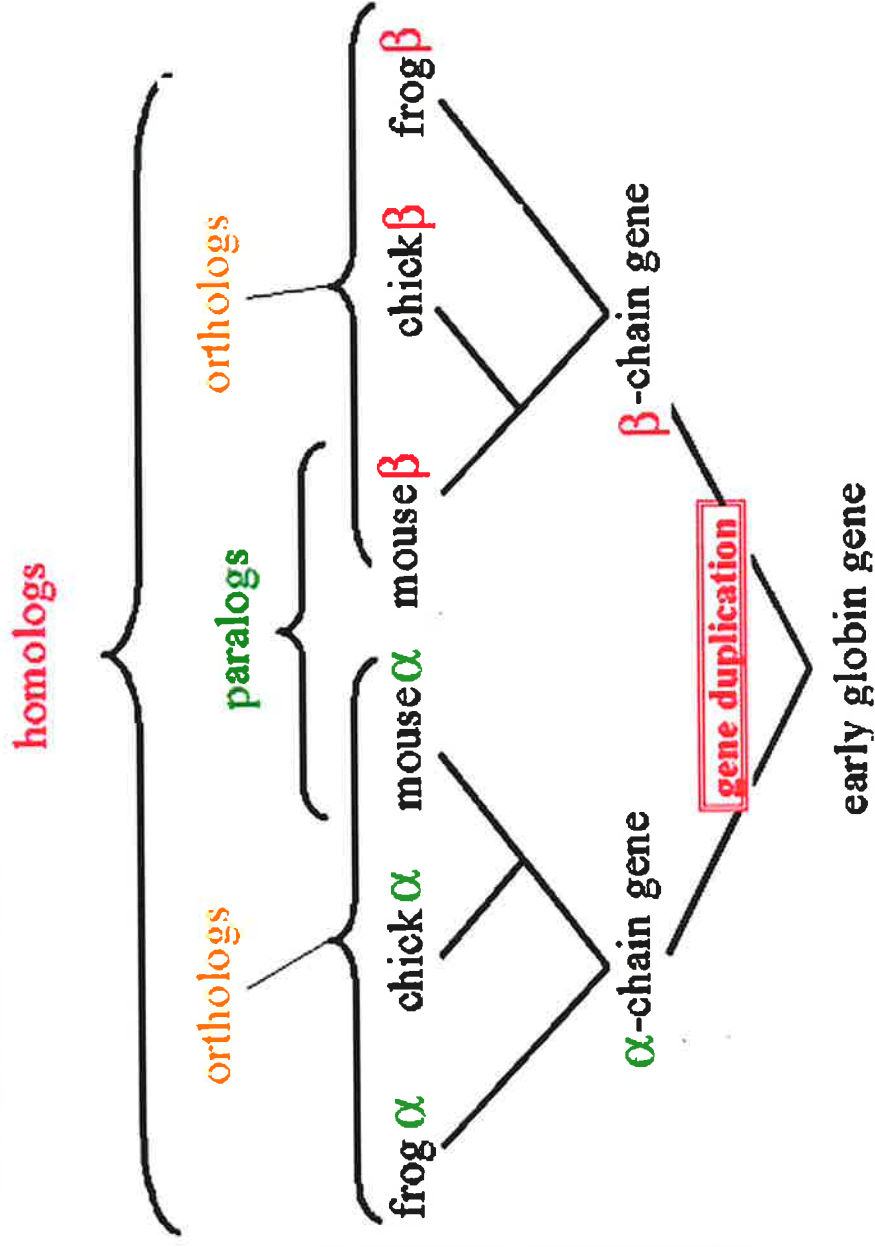
Paralogy: Homologous sequences are paralogous if they were separated by a gene duplication event

Orthology: Homologous sequences are orthologous if they were separated by a speciation event

Further reading:

Koonin EV (2005). "Orthologs, paralogs, and evolutionary genomics".
Annu. Rev. Genet. 39:309-338.

Homology: Orthologs and Paralogs



Sequence Alignment

①

I) Sequence comparison

A Sequence is a string of letters.

DNA sequence A T G G C

Protein level: W C Y L V I

When it comes to sequence comparison, if you have access to both DNA and protein sequence, use the protein sequence.

Approach 1 : Comparing compositions only lead to weak signals.

Sequence comparison 1/28/26

①

I) Comparing "letters"

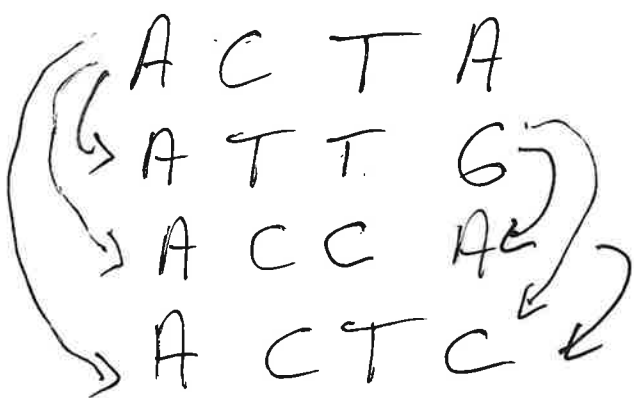
a) Conservative

	A	T	G	C
A	1	0	0	0
T	0	1	0	0
G	0	0	1	0
C	0	0	0	1

b) Chemistry

	A	T	G	C
A	2	0	1	0
T	0	2	0	1
G	1	0	2	0
C	0	1	0	2

c) Machine Learning approach



	A	T	G	C
A			'	''
T				
G	'			'
C		''		

②

	A	T	G	C
A	7	0	1	2
T	0	3	0	4
G	1	0	0	1
C	0	2	0	3

↙

	A	T	G	C
A	$\frac{7}{10}$	0	$\frac{1}{10}$	$\frac{2}{10}$
T	0	$\frac{3}{7}$	0	$\frac{4}{7}$
G	$\frac{1}{2}$	0	0	$\frac{1}{2}$
C	0	$\frac{2}{5}$	0	$\frac{3}{5}$

What I defined is P_{ij}

Propensity $A_{ij} = \frac{P_{ij}}{P_i P_j}$

↘ $\log A_{ij}$

③
The concept of taking existing alignments and learn from them to build a "substitution matrix" that quantify the differences between "letters" is implemented in the BLOSUM matrices.

Alignment S_1
 $\vdots$
 S_N } only include sequences that are at least 90% similar $\rightarrow$ BLOSUM 90

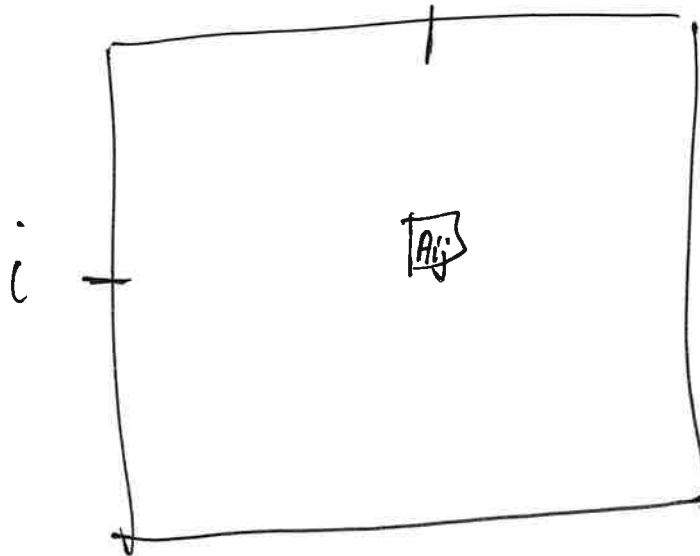
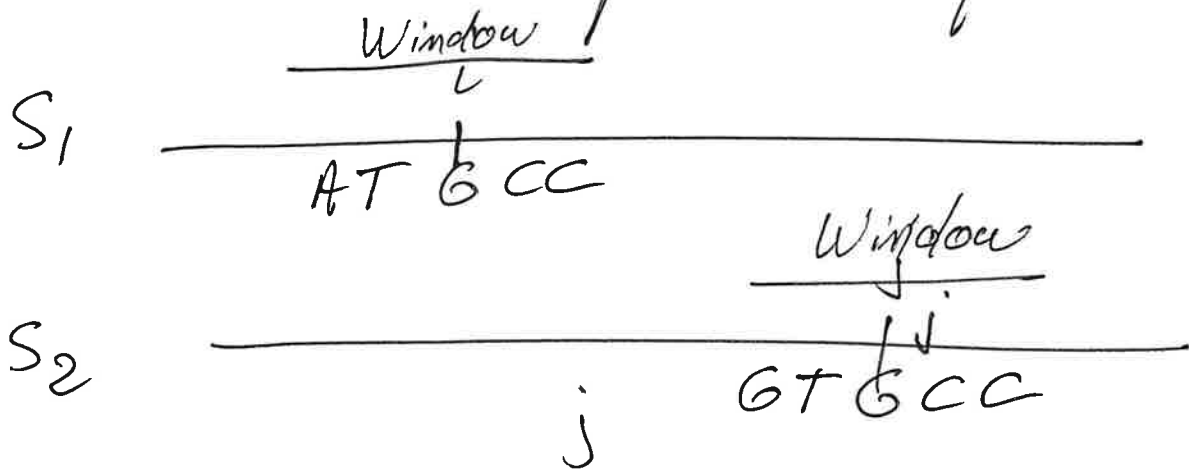
filter at least 30% $\rightarrow$ BLOSUM 30

BLOSUM 100

If you don't know
one size fits all $\rightarrow$ BLOSUM 62

II) How to use the matrices.

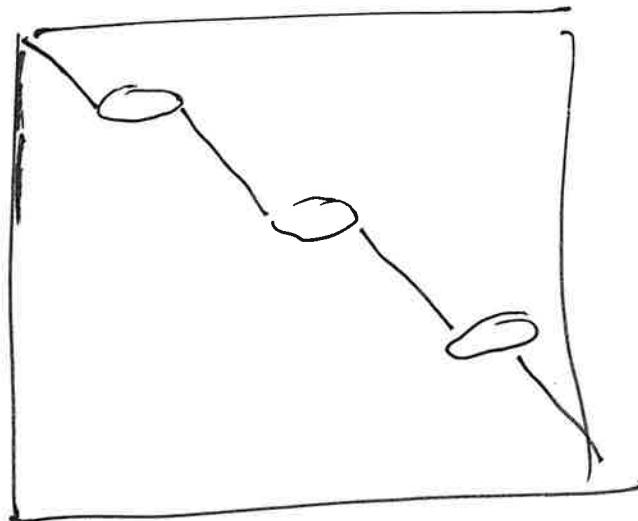
You want to compare 2 sequences



→ Dotplot

Example:

mRNA



DNA