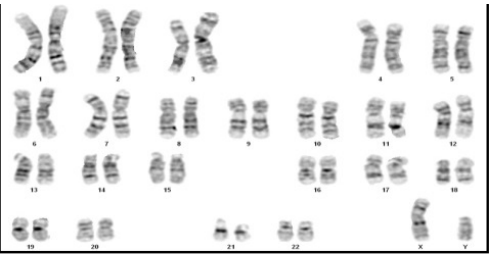
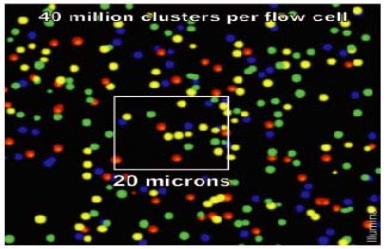
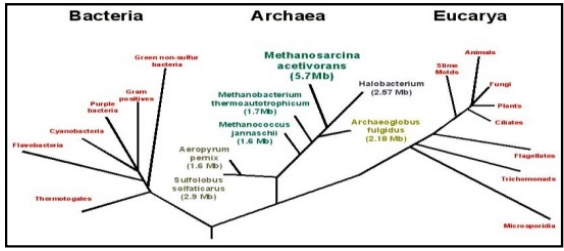


TAACCCTAACCCTAACCCTAACCCTAACCCTA
CCTAACCCTAACCCTAACCCTAACCCTAACCC
CCCTAACCCCTAACCCTAACCCTAACCCTAAC
AACCCTAACCCTAACCCTAACCCCTAACCCTA
ACCCTAACCCCAACCCCAACCCCAACCCCAAC
CTACCCTAACCCTAACCCTAACCCTAACCCTA
ACCCTAACCCTAACCCTAACCCCTAACCCTAA



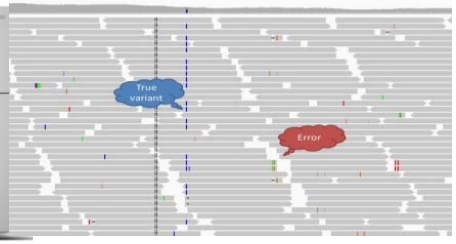


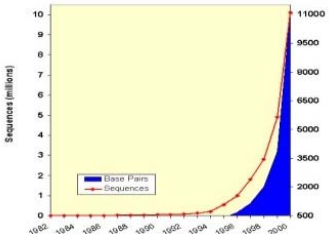
Supplementary Learning Materials

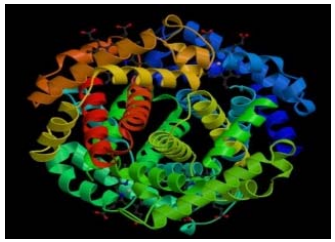
TA: 叶永鑫(Adam Y. Ye)

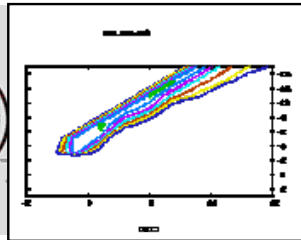
北京大学生物信息学中心

Center for Bioinformatics, Peking University









Outline

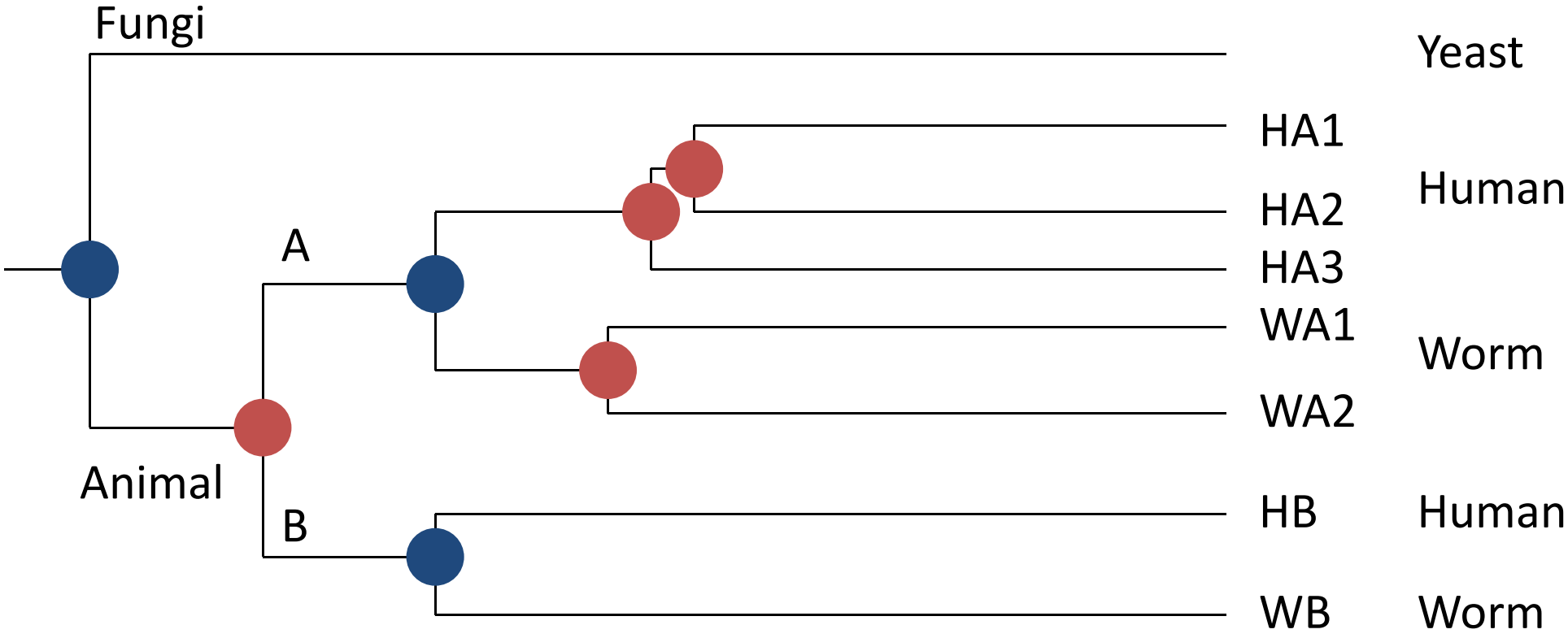
- Homology & Similarity
- Similarity Matrix
- Dot Matrix

Homology

- Homology
  - derived from a common ancestor

  - ortholog: derived from speciation
  - paralog: derived from duplication

Ortholog vs Paralog

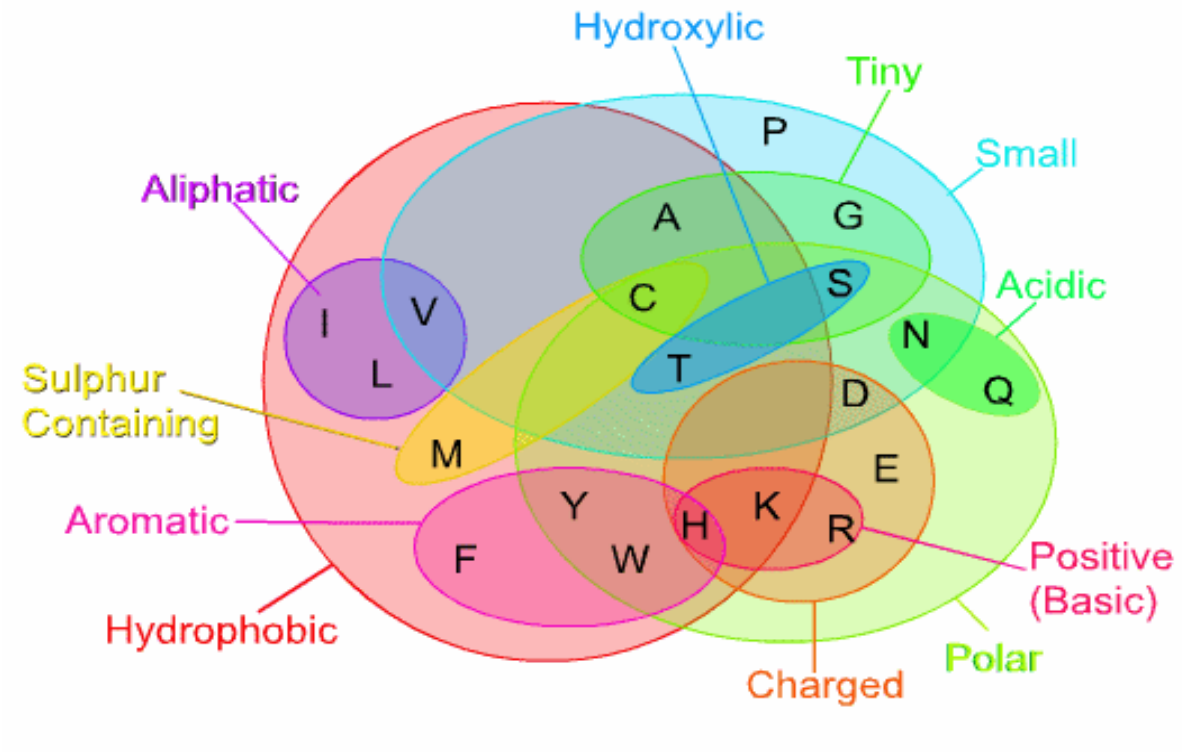


Ortholog comes with speciation Paralog comes with duplication

revised based on Sonnhammer, E.L., and Koonin, E.V. (2002). Orthology, paralogy and proposed classification for paralog subtypes. *TRENDS in Genetics 18*, 619–620.

Similarity vs Identity

- Similarity
- Identity

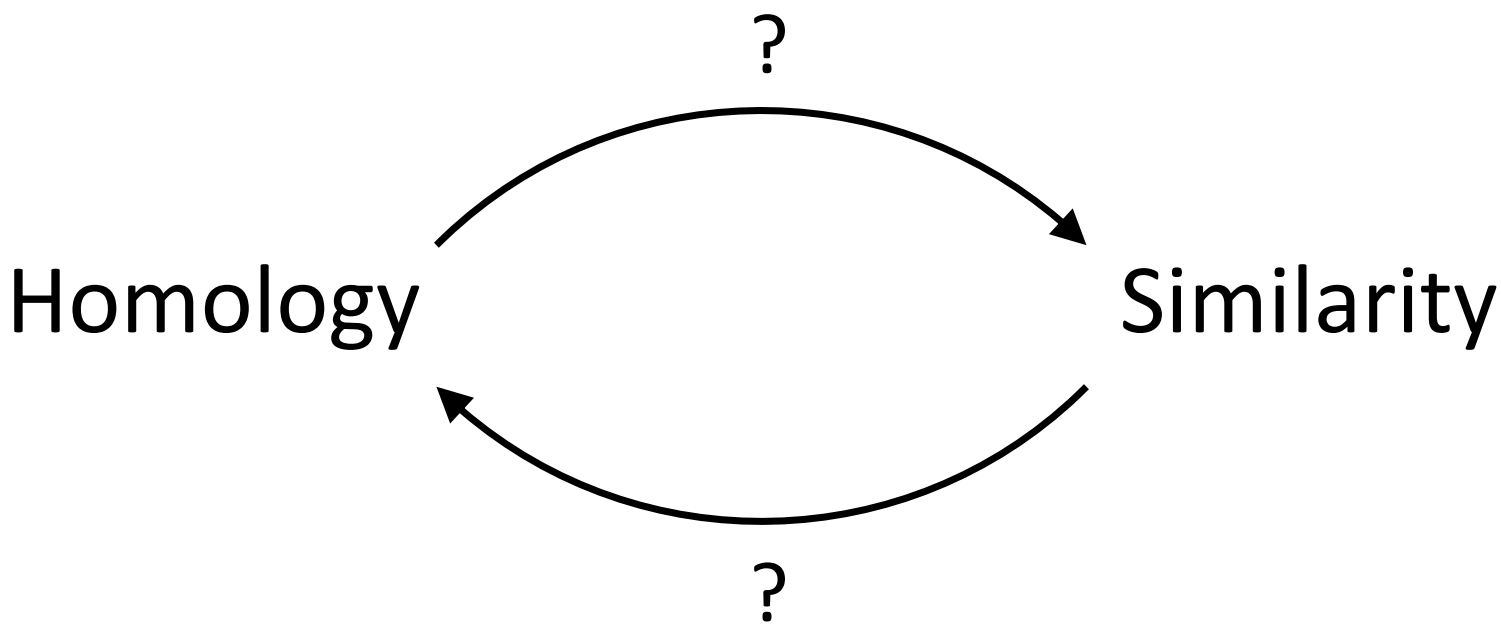


Amino Acids

A alanine (ala)
R arginine (arg)
N asparagine (asn)
D aspartic acid (asp)
C cysteine (cys)
Q glutamine (gln)
E glutamic acid (glu)
G glycine (gly)
H histidine (his)
I isoleucine (ile)
L leucine (leu)
K lysine (lys)
M metioneine (met)
F phenyalanine (phe)
P proline (pro)
S serine (ser)
T threonine (thr)
W trytophan (trp)
Y tyrosine (tyr)

(Adopted from Prof. Jingchu Luo)

Homology vs Similarity



How to let computer do this job?

- How to measure similarity?
  - Similarity matrix

Similarity Matrix

- For nucleotides,
  - usually only distinguish match / mismatch (identity matrix) for sequence alignment
  - but a more complicated substitution model is used for phylogeny reconstruction

	A	T	C	G
A	1	-2	-2	-2
T	-2	1	-2	-2
C	-2	-2	1	-2
G	-2	-2	-2	1

Similarity Matrix

- For amino acids,
  - PAM (1978, Margaret Dayhoff)
    - Two sequences are 1 PAM apart if they differ in 1 % of the residues.
    - 1 PAM = one step of evolution
  - BLOSUM (1992, Steven Henikoff & Jorja Henikoff)
    - computed by looking at "blocks" of conserved sequences found in multiple protein alignments

PAM

1	A	B	C
A	0.8	0.1	0.1
B	0.05	0.9	0.05
C	0.15	0.05	0.8

- PAM 1

- PAM 2 ?

P(A->?->A) = P(A->A->A) + P(A->B->A) + P(A->C->A)
= P(A->A)P(A->A) + P(A->B)P(B->A) + P(A->C)P(C->A)

P(A->?->B) = P(A->A->B) + P(A->B->B) + P(A->C->B)
= P(A->A)P(A->B) + P(A->B)P(B->B) + P(A->C)P(C->B)

...

PAM

- PAM 1
- PAM 2 = (PAM 1)^2

1	A	B	C
A	0.8	0.1	0.1
B	0.05	0.9	0.05
C	0.15	0.05	0.8

$\times$

1	A	B	C
A	0.8	0.1	0.1
B	0.05	0.9	0.05
C	0.15	0.05	0.8

=

2	A	B	C
A	0.66	0.175	0.165
B	0.093	0.817	0.09
C	0.243	0.1	0.657

PAM

- PAM 1
- PAM 250
  $= (\text{PAM } 1)^{250}$
- Log odds of PAM 250
  log odds = log(p/(1-p))

C	Cys	12																			
S	Ser	0	2																		
T	Thr	-2	1	3																	
P	Pro	-3	1	0	6																
A	Ala	-2	1	1	1	2															
G	Gly	-3	1	0	-1	1	5														
N	Asn	-4	1	0	-1	0	0	2													
D	Asp	-5	0	0	-1	0	1	2	4												
E	Glu	-5	0	0	-1	0	0	1	3	4											
Q	Gln	-5	-1	-1	0	0	-1	1	2	2	4										
H	His	-3	-1	-1	0	-1	-2	2	1	1	3	6									
R	Arg	-4	0	-1	0	-2	-3	0	-1	-1	1	2	6								
K	Lys	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5							
M	Met	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6						
I	Ile	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5					
L	Leu	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	6				
V	Val	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4			
F	Phe	-4	-3	-3	-5	-4	-5	-4	-6	-5	-5	-2	-4	-5	0	1	2	-1	9		
Y	Tyr	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10	
W	Trp	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	17
		C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
		Cys	Ser	Thr	Pro	Ala	Gly	Asn	Asp	Glu	Gln	His	Arg	Lys	Met	Ile	Leu	Val	Phe	Tyr	Trp

BLOSUM

Less divergent ⟷ More divergent

BLOSUM 80	BLOSUM 62	BLOSUM 45
PAM 1	PAM 120	PAM 250

How to let computer do this job?

- How to measure similarity?
  - Similarity matrix
- How to find out alignment?
  - Dot matrix
  - Dynamic programming
  - BLAST

Dot Matrix

ATAGCTA

ATAGCTA

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

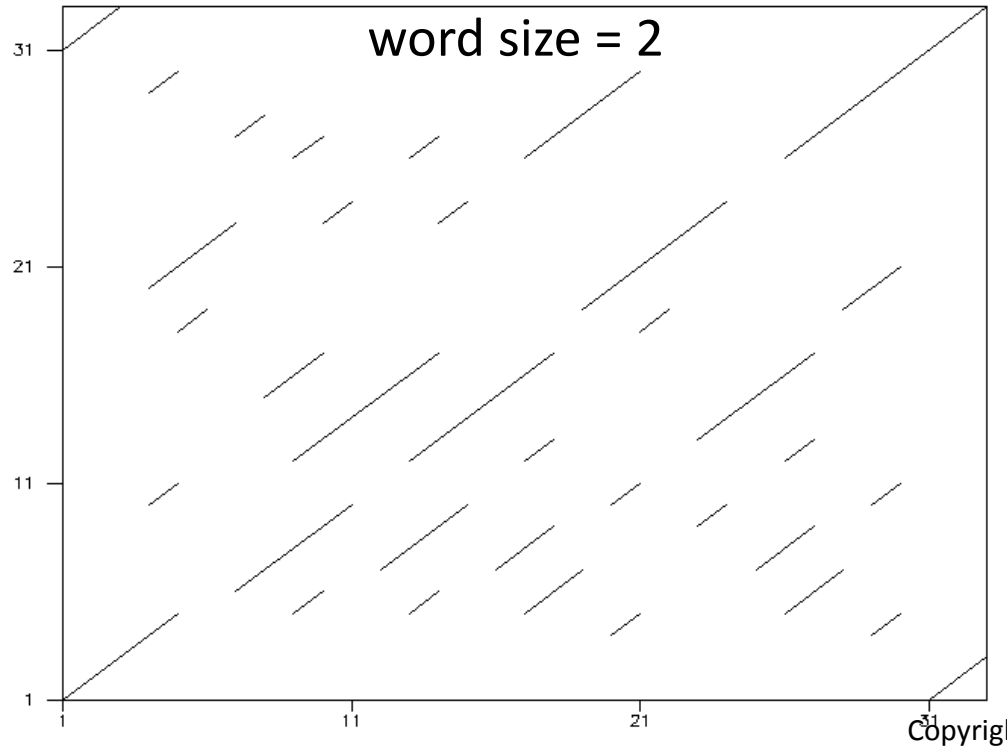
ATAGC-TA

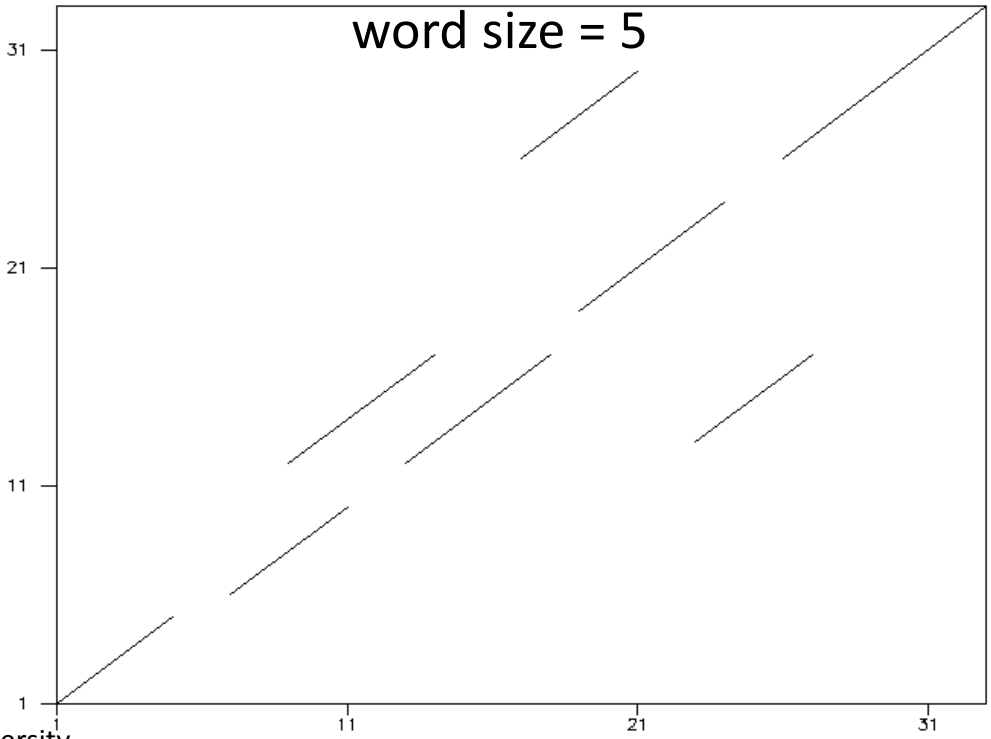
AT-GCCTA

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

Dot Matrix

- PQRACGTGCTAGCTAGCT-GACGTAGCTGACPQR
- PQRAC-TGCTACCTAGCTCGACGTATCTGACPQR





Thank you for your attention