

Mon, Nov 16

Lecture 10A: Biology: sequence variation



Practical Bioinformatics (Biol 4220)
Instructor: Michael Landis
Email: michael.landis@wustl.edu



Lecture 10A outline

1. Base frequencies
2. GC richness
3. Genetic code
4. Amino acid properties
5. Lab 10A overview

Sequence variation

	* * ***
Felis_cattus_cytB	tccggttattcat t tcaatc
Mus_musculus_cytB	tccggttat c cac a caatc
Homo_sapiens_cytB	tccggttattc t ctcaatc
Bos_taurus_cytB	tc t ggttattcactcaatc

samples across species

Nucleotide composition

The four nucleotides (A, C, G, T) do not occur in equal proportions across every gene and every species. Reasons include:

- Mutational bias
- Repair bias
- Synthesis costs
- Limited resources
- Selection on thermal tolerance
- Selection on amino acids
- Genetic code biases
- Transcription factors motifs
- Transposons
- *etc.*

Nucleotide composition

Do A, C, G, and T occur in equal proportions
across sites? across species?

sp1_gnA A C C T G T

sp2_gnA A C T T G T

sp3_gnA A C C T G A

gene A

sp1_gnB T C G G G C

sp2_gnB G C A G C C

sp3_gnB G C A C C T

gene B

Composition per site (across sequences)

1 2 3 4 5 6
 sp1_gnA A C C T G T
 sp2_gnA A C T T G T
 sp3_gnA A C C T G A

Site	A	C	G	T
1	3/3	-	-	-
2	-	3/3		
3	-	2/3	-	1/3
4	-	-	-	3/3
5	-	-	3/3	-
6	1/3	-	-	2/3
Total	4	5	3	6

Composition per site (across sequences)

1 2 3 4 5 6
sp1_gnB T C G G G C
sp2_gnB G C A G C C
sp3_gnB G C A C C T

Site	A	C	G	T
1	-	-	2/3	1/3
2	-	3/3	-	-
3	2/3	-	1/3	-
4	-	1/3	2/3	-
5	-	2/3	1/3	-
6	-	2/3	-	1/3
Total	2	8	6	2

Composition per sequence (across sites)

	1	2	3	4	5	6
sp1_gnA	A	C	C	T	G	T
sp2_gnA	A	C	T	T	G	T
sp3_gnA	A	C	C	T	G	A

Species	A	C	G	T
sp1	1/6	2/6	1/6	2/6
sp2	1/6	1/6	1/6	3/6
sp3	2/6	2/6	1/6	1/6
Total	4	5	3	6

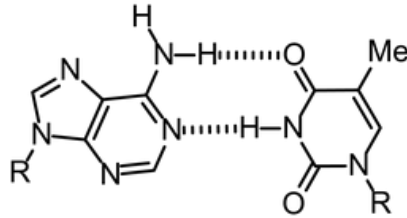
Composition per sequence (across sites)

	1	2	3	4	5	6
sp1_gnB	T	C	G	G	G	C
sp2_gnB	G	C	A	G	C	C
sp3_gnB	G	C	A	C	C	T

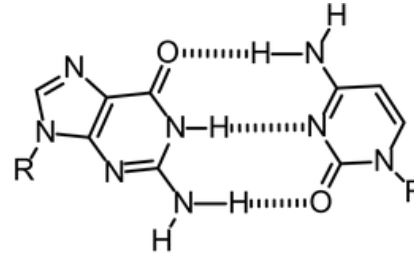
Species	A	C	G	T
sp1	-	2/6	3/6	1/6
sp2	1/6	3/6	2/6	-
sp3	1/6	3/6	1/6	1/6
Total	2	8	6	2

GC-content

What are some causes of GC-content bias?



A·T base pair



G·C base pair

- G-C pairs have greater thermostability
- G-C pairs require more nitrogen than A-T
- G-C pairs have higher synthesis cost
- AT-bias in spontaneous mutation
- GC-biased gene conversion (mismatch repair)
- CpG islands and cytosine-methylation in vertebrates
- High-GC % recombination rates

GC-content

Do G and C occur in equal proportion to A and T?

computed as $(C+G) / (A+C+G+T)$

sp1_gnA A C C T G T

sp2_gnA A C T T G T

sp3_gnA A C C T G A

sp1_gnB T C G G G C

sp2_gnB G C A G C C

sp3_gnB G C A C C T

GC content in gene A

$$8/18 = 44\%$$

GC content in gene B

$$14/18 = 78\%$$

Phylogenetically informative sites

	1	2	3	4
spW_gnC	G	C	C	T
spX_gnC	G	C	T	G
spY_gnC	G	A	C	T
spZ_gnC	G	A	C	C

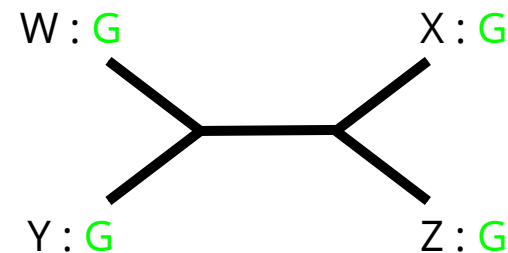
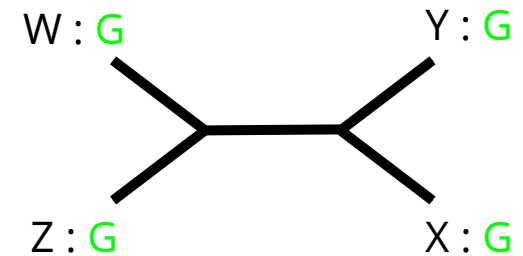
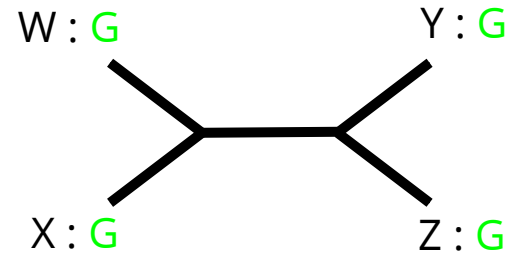
A site is ***phylogenetically informative*** if it can be used to estimate common ancestry

A phylogenetically informative site contains two variants, with at least two samples per variant

Phylogenetically informative sites

	1	2	3	4
spW_gnC	G	C	C	T
spX_gnC	G	C	T	G
spY_gnC	G	A	C	T
spZ_gnC	G	A	C	C

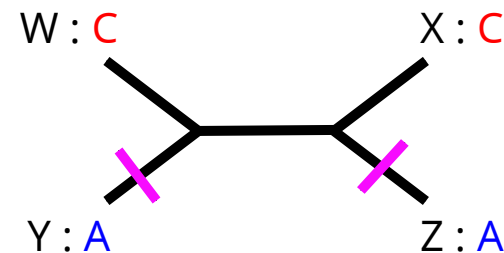
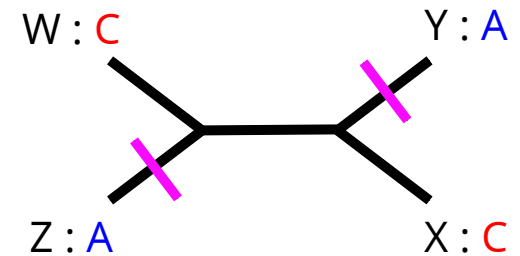
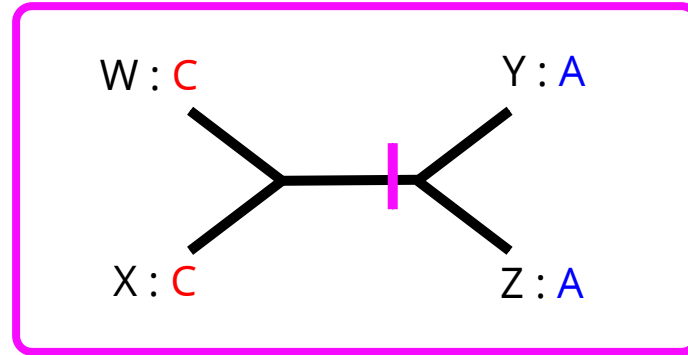
↑
Not PI



Phylogenetically informative sites

	1	2	3	4
spW_gnC	G	C	C	T
spX_gnC	G	C	T	G
spY_gnC	G	A	C	T
spZ_gnC	G	A	C	C

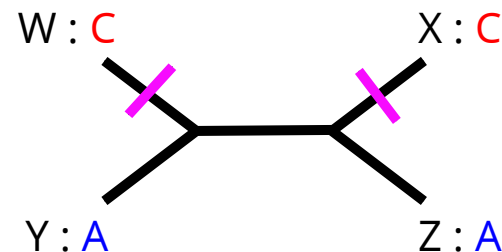
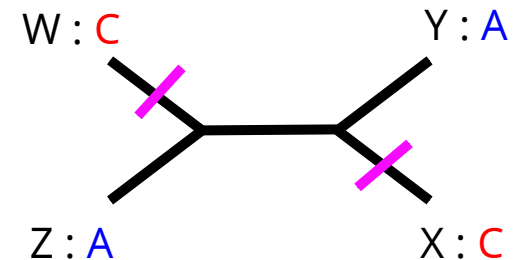
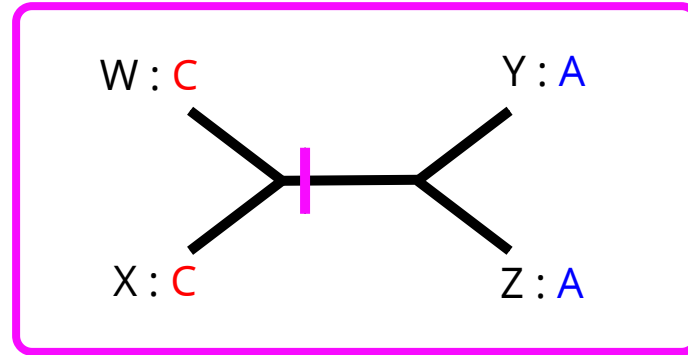
↑
Yes, PI



Phylogenetically informative sites

	1	2	3	4
spW_gnC	G	C	C	T
spX_gnC	G	C	T	G
spY_gnC	G	A	C	T
spZ_gnC	G	A	C	C

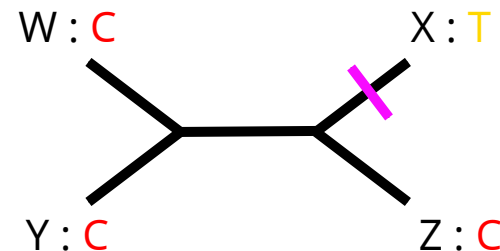
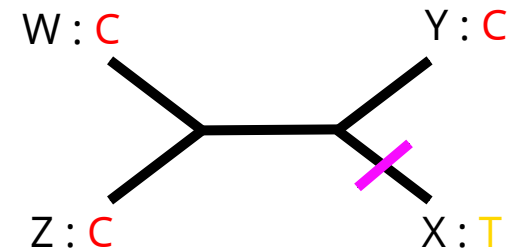
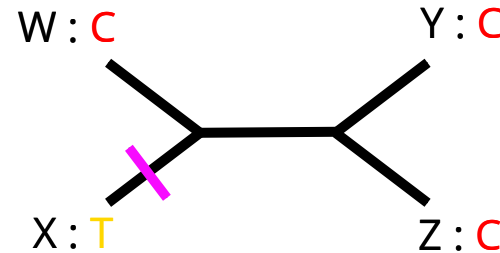
↑
Yes, PI



Phylogenetically informative sites

	1234
spW_gnC	G C C T
spX_gnC	G C T G
spY_gnC	G A C T
spZ_gnC	G A C C

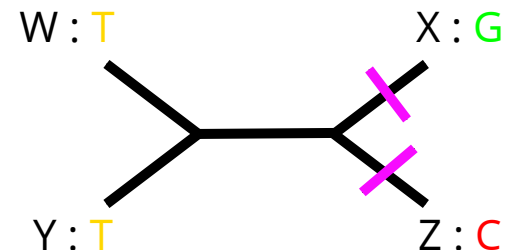
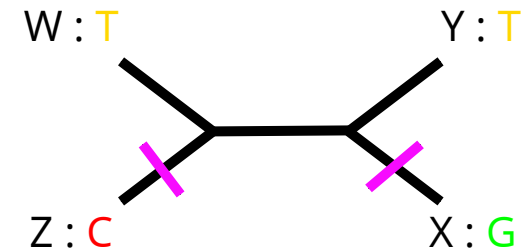
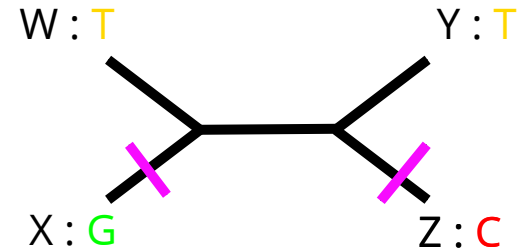
↑
Not PI



Phylogenetically informative sites

	1234
spW_gnC	G C C T
spX_gnC	G C T G
spY_gnC	G A C T
spZ_gnC	G A C C

↑
Not PI



Genetic code

The ***genetic code*** determines how nucleotide triplets (codons) are translated into amino acids

	1 2 3	1 2 3	1 2 3		
sp1_gnD	A T G	T G T	A T C	G T C	. .
sp2_gnD	A T G	T G T	C T A	G T C	. .
sp3_gnD	A T G	T G C	C T C	G T C	. .
	—	—	—		
	codons				

Genetic code

The ***genetic code*** determines how nucleotide triplets (codons) are translated into amino acids

	1 2 3	1 2 3	1 2 3	
sp1_gnD	Met	Cys	Ile	Val . .
sp2_gnD	Met	Cys	Leu	Val . .
sp3_gnD	Met	Cys	Leu	Val . .
	—	—	—	
	amino acids			

Genetic code

		Second letter					
		U	C	A	G		
First letter	U	UUU] Phenylalanine (Phe) UUC] UUA] Leucine (Leu) UUG]	UCU] Serine (Ser) UCC] UCA] UCG]	UAU] Tyrosine (Tyr) UAC] UAA Stop UAG Stop	UGU] Cysteine (Cys) UGC] UGA Stop UGG Tryptophan (Trp)	U	stop codons
	C	CUU] Leucine (Leu) CUC] CUA] CUG]	CCU] Proline (Pro) CCC] CCA] CCG]	CAU] Histidine (His) CAC] CAA] Glutamine (Gln) CAG]	CGU] Arginine (Arg) CGC] CGA] CGG]	U	
	A	AUU] Isoleucine (Ile) AUC] AUA] AUG Methionine (Met)	ACU] Threonine (Thr) ACC] ACA] ACG]	AAU] Asparagine (Asn) AAC] AAA Lysine (Lys) AAG]	AGU] Serine (Ser) AGC] AGA] Arginine (Arg) AGG]	A	
	G	GUU] Valine (Val) GUC] GUA] GUG]	GCU] Alanine (Ala) GCC] GCA] GCG]	GAU] Aspartic acid (Asp) GAC] GAA] Glutamic acid (Glu) GAG]	GGU] Glycine (Gly) GGC] GGA] GGG]	G	

start codon

stop codons

GUU, GUC, GUA, and GUG all encode Valine

AAA and AAG encode Lysine

© Copyright, 2014. University of Waikato. All rights reserved.
www.biotechlearn.org.nz

Reading frame

The ***reading frame*** determines the identity of each codon, and therefore amino acid identity

	1 2 3	1 2 3	1 2 3		
sp1_gnD	A T G	T G T	A T C	G T C	. .
sp2_gnD	A T G	T G T	C T A	G T C	. .
sp3_gnD	A T G	T G C	C T C	G T C	. .
	—	—	—		
	↑				
Begin reading frame					


Reading frame

The ***reading frame*** determines the identity of each codon, and therefore amino acid identity

		1 2 3	1 2 3	1 2 3		
sp1_gnD	A	TGT	GTA	TCG	TC	.
sp2_gnD	A	TGT	GTC	TAG	TC	.
sp3_gnD	A	TGT	GCC	TCG	TC	.
		_____	_____	_____		
		↑				
		Begin reading frame				

Reading frame

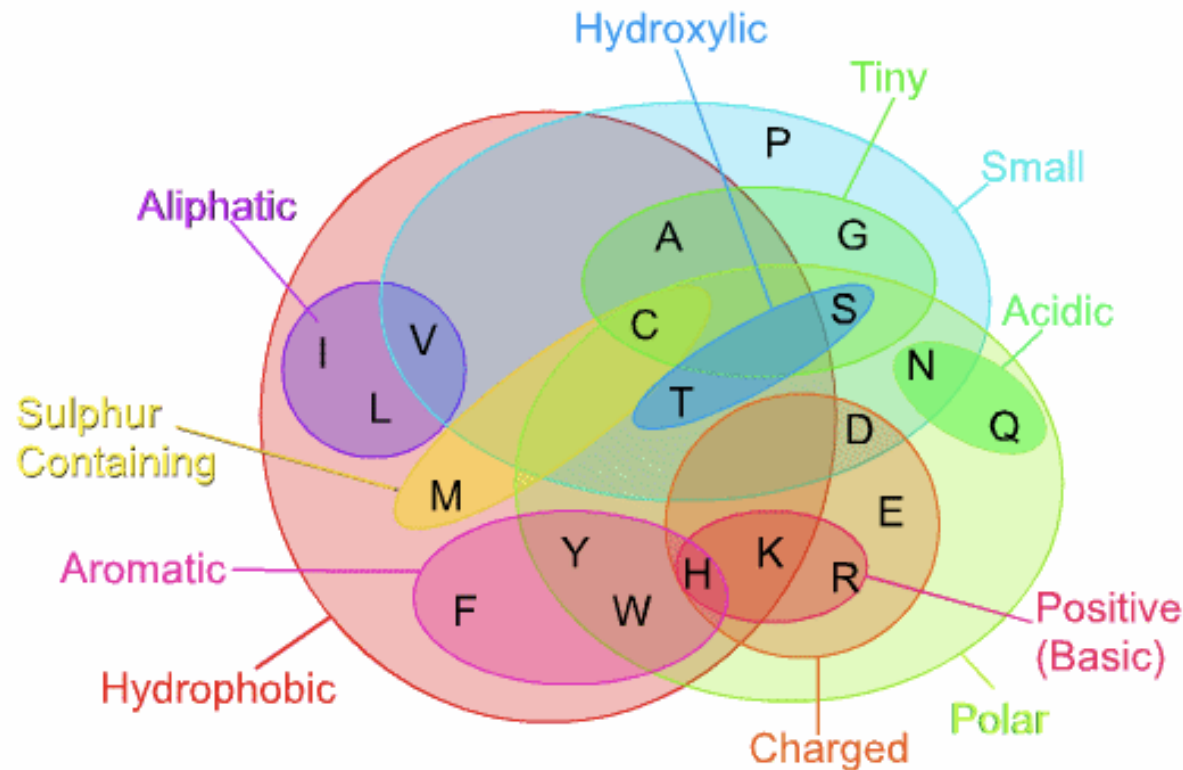
The ***reading frame*** determines the identity of each codon, and therefore amino acid identity

		1 2 3	1 2 3	1 2 3	
sp1_gnD	AT	GTG	TAT	CGT	C . .
sp2_gnD	AT	GTG	TCT	AGT	C . .
sp3_gnD	AT	GTG	CCT	CGT	C . .
		<hr/>	<hr/>	<hr/>	
					
		Begin reading frame			

	123	123	123	
sp1_gnD	Met	Cys	Ile	Val . .
sp2_gnD	Met	Cys	Leu	Val . .
sp3_gnD	Met	Cys	Leu	Val . .
	<u> </u>	<u> </u>	<u> </u>	
	amino acids			

		123	123	123	
sp1_gnD	. .	Val	Tyr	Arg	. . .
sp2_gnD	. .	Val	Ser	Ser	. . .
sp3_gnD	. .	Val	Pro	Arg	. . .
		<u> </u>	<u> </u>	<u> </u>	
		amino acids			

Amino acids



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 methionine (met)
F phenylalanine (phe)
P proline (pro)
S serine (ser)
T threonine (thr)
W tryptophan (trp)
Y tyrosine (tyr)

Physicochemical properties of amino acids influence protein structure and function

Amino acid frequencies

sp1_gnD	..	Val	Tyr	Arg	...
sp2_gnD	..	Val	Ser	Ser	...
sp3_gnD	..	Val	Pro	Arg	...

How common are different properties?

Arg: positive, polar, charged

Ser: small, tiny, hydroxyl, polar

Tyr: aromatic, hydrophobic, polar

Pro: small

Val: small, hydrophobic, aliphatic,

Amino acid frequencies

sp1_gnD	..	<u>Val</u>	Tyr	Arg	...
sp2_gnD	..	<u>Val</u>	<u>Ser</u>	<u>Ser</u>	...
sp3_gnD	..	<u>Val</u>	<u>Pro</u>	Arg	...

~66% of AA in this window are **small**

Arg: positive, polar, charged

Ser: **small**, tiny, hydroxyl, polar

Tyr: aromatic, hydrophobic, polar

Pro: **small**

Val: **small**, hydrophobic, aliphatic,

Amino acid frequencies

sp1_gnD	..	<u>Val</u>	<u>Tyr</u>	Arg	...
sp2_gnD	..	<u>Val</u>	Ser	Ser	...
sp3_gnD	..	<u>Val</u>	Pro	Arg	...

~44% of AA in this window are **hydrophobic**

Arg: positive, polar, charged

Ser: small, tiny, hydroxyl, polar

Tyr: aromatic, **hydrophobic**, polar

Pro: small

Val: small, **hydrophobic**, aliphatic,

Codon usage bias

Codons are not necessarily used in equal proportions when encoding amino acids; this is called ***codon usage bias***

sp1_gnE GTT GTT GTG GTT

sp1_gnE GTA GTT GTT GTT

sp1_gnE GTT GTT GTT GTC



GTT overrepresented
for Valine

sp1_gnE Val Val Val Val

sp2_gnE Val Val Val Val

sp3_gnE Val Val Val Val

Lab 10A

github.com/WUSTL-Biol4220/home/labs/lab_10A.md