

# Lecture 16

## sequence statistics



Course: Practical Bioinformatics (BIOL 4220)  
Instructor: Michael Landis  
Email: [michael.landis@wustl.edu](mailto:michael.landis@wustl.edu)



# Lecture 16 outline

Last time: Python series

This time: sequence statistics

- base frequencies
- GC richness
- genetic code
- amino acid properties

# Sequence variation

		*		*	***
Felis_cattus_cytB	tccg	ttattca	t	tcaatc	
Mus_musculus_cytB	tccg	ttatc	c	cac	a
Homo_sapiens_cytB	tccg	ttattc	t	ctcaatc	
Bos_taurus_cytB	tct	t	gttattc	actcaatc	

*gene samples across species*

# Nucleotide composition

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

Reasons include:

- mutational bias
- repair bias
- synthesis costs
- limited resources
- selection on thermal tolerance
- selection on amino acids
- genetic code biases
- transcription factor motifs
- *etc.*

# Nucleotide composition

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

sp1\_gnA ACCTGT

sp2\_gnA ACTTGT

sp3\_gnA ACCTGA

*gene A*

sp1\_gnB TCGGGC

sp2\_gnB GCAGCC

sp3\_gnB GCACCT

*gene B*

# Composition per site (across sequences)

123456  
sp1\_gnA ACCTGT  
sp2\_gnA ACTTGT  
sp3\_gnA ACCTGA

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
<b>Total</b>	4	5	3	6

# Composition per site (across sequences)

123456  
sp1\_gnB TCGGGC  
sp2\_gnB GCAGCC  
sp3\_gnB GCACCT

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
<b>Total</b>	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
<b>Total</b>	4	5	3	6



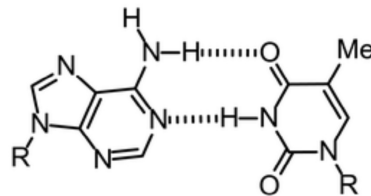
# Composition per sequence (across sites)

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

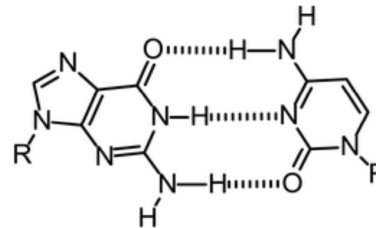
Species	A	C	G	T
sp1	-	2/6	3/6	1/6
sp2	1/6	3/6	2/6	-
sp3	1/6	3/6	1/6	1/6
<b>Total</b>	2	8	6	2

# GC-content bias

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



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 bias

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

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

sp1\_gnA ACCTGT

sp2\_gnA ACTTGT

sp3\_gnA ACCTGA

GC content in gene A  
 $8/18 = 44\%$

sp1\_gnB TCGGGC

sp2\_gnB GCAGCC

sp3\_gnB GCACCT

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

gene C

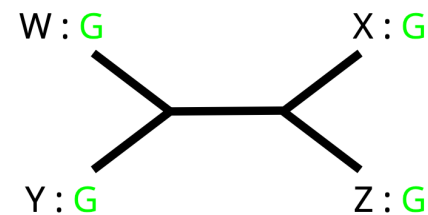
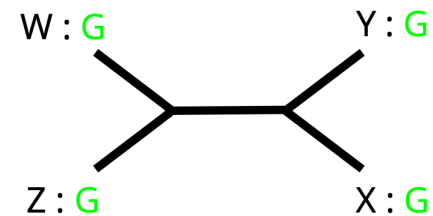
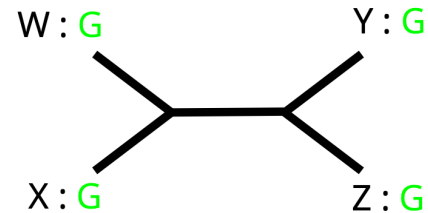
A site is ***phylogenetically informative*** if it can be used to estimate common ancestry

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

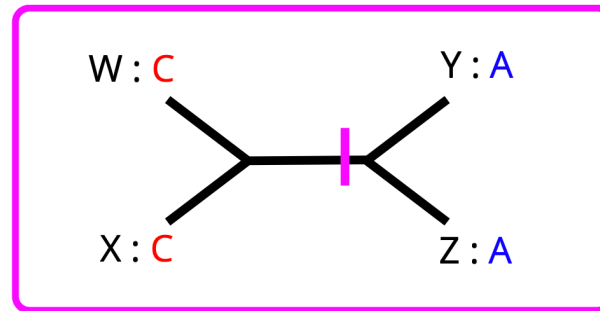
# Phylogenetically informative sites

	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

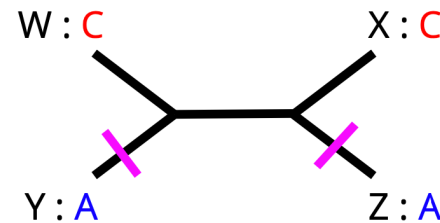
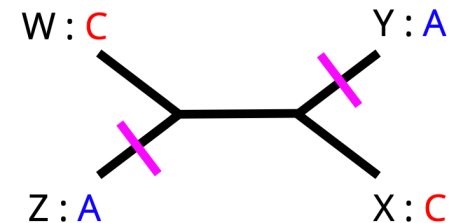


# Phylogenetically informative sites

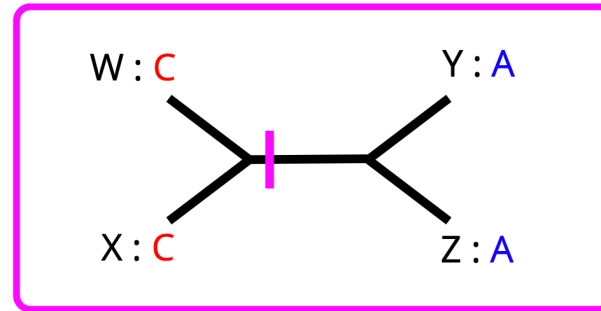


	1	2	3	4
spW_gnC	G	C	C	T
spX_gnC	G	C	T	G
spY_gnC	G	A	C	T
spZ_gnC	G	A	C	C

↑  
Yes, PI

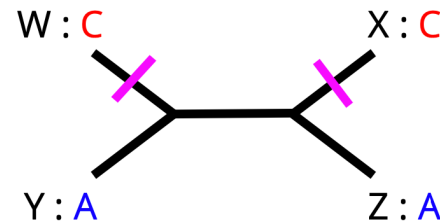
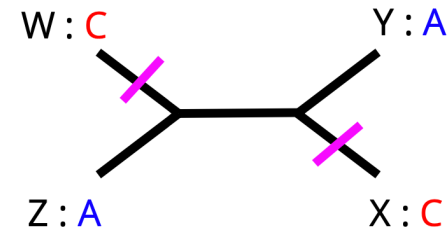


# Phylogenetically informative sites



	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

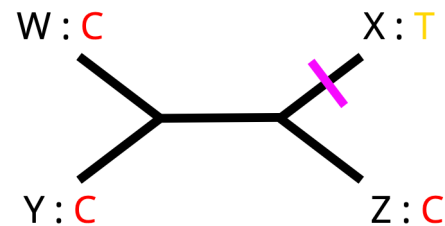
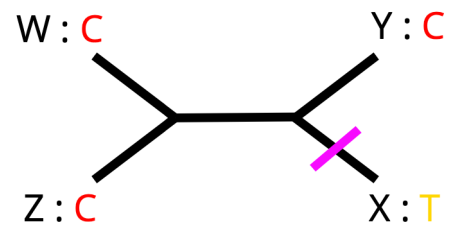
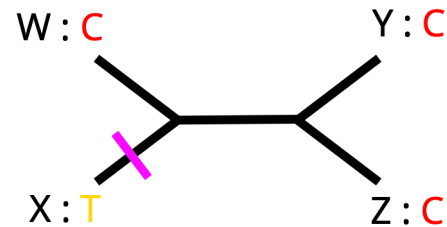
↑  
Yes, PI



# Phylogenetically informative sites

	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

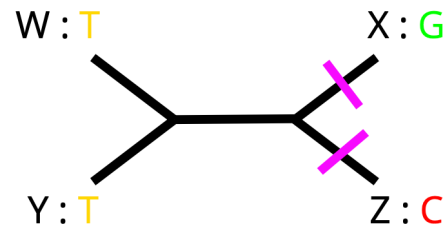
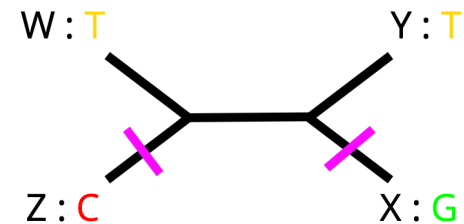
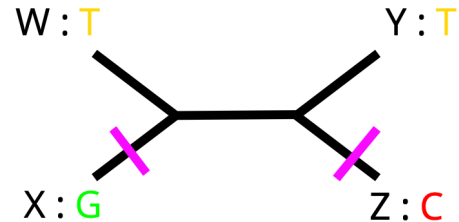




# Phylogenetically informative sites

	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

↑  
Not PI



# Genetic code

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

	123	123	123	123	
sp1_gnD	ATG	TGT	ATC	GTC	..
sp2_gnD	ATG	TGT	CTA	GTC	..
sp3_gnD	ATG	TGC	CTC	GTC	..
	<u>        </u>	<u>        </u>	<u>        </u>	<u>        </u>	
	<i>codons</i>				

# Genetic code

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

		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>	<u>          </u>	
		<i>amino acids</i>			

# Genetic code

(standard table; different species and genomes may use different codes)

		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 C A G
	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 C A G
	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 ]	U C A G
	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 ]	U C A G

start codon

stop codons


GUU, GUC, GUA, and GUG all encode Valine

AAA and AAG encode Lysine

# Reading frame

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

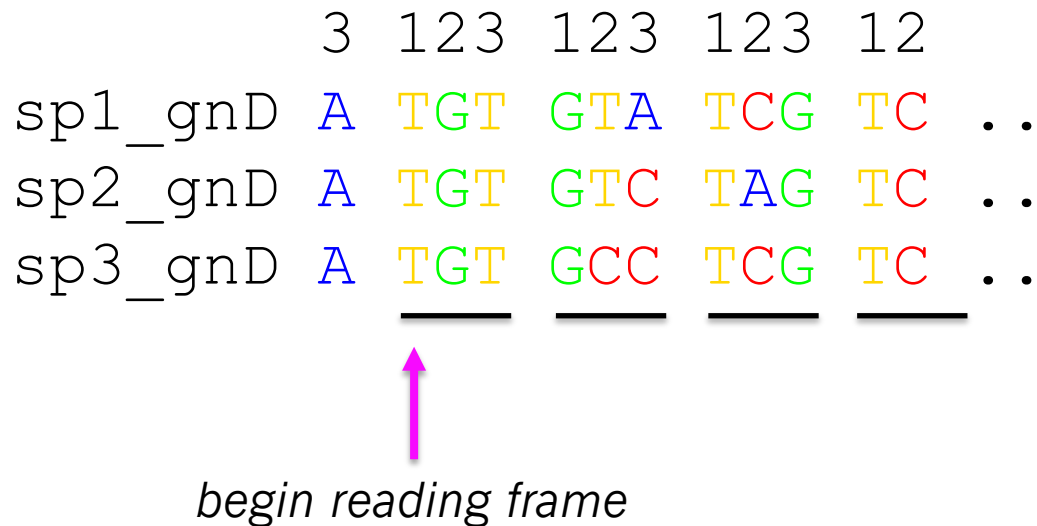
	123	123	123	123	
sp1_gnD	A T G	T G T	A T C	G T C	..
sp2_gnD	A T G	T G T	C T A	G T C	..
sp3_gnD	A T G	T G C	C T C	G T C	..



*begin reading frame*

# Reading frame


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



# Reading frame

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

	23	123	123	123	1	
sp1_gnD	AT	GTG	TAT	CGT	C	..
sp2_gnD	AT	GTG	TCT	AGT	C	..
sp3_gnD	AT	GTG	CCT	CGT	C	..



*begin reading frame*

# Amino acid translation differs due to frame shift

	123	123	123	123	
sp1_gnD	ATG	TGT	ATC	GTC	..
sp2_gnD	ATG	TGT	CTA	GTC	..
sp3_gnD	ATG	TGC	CTC	GTC	..



	123	123	123	123	
sp1_gnD	Met	Cys	Ile	Val	..
sp2_gnD	Met	Cys	Leu	Val	..
sp3_gnD	Met	Cys	Leu	Val	..

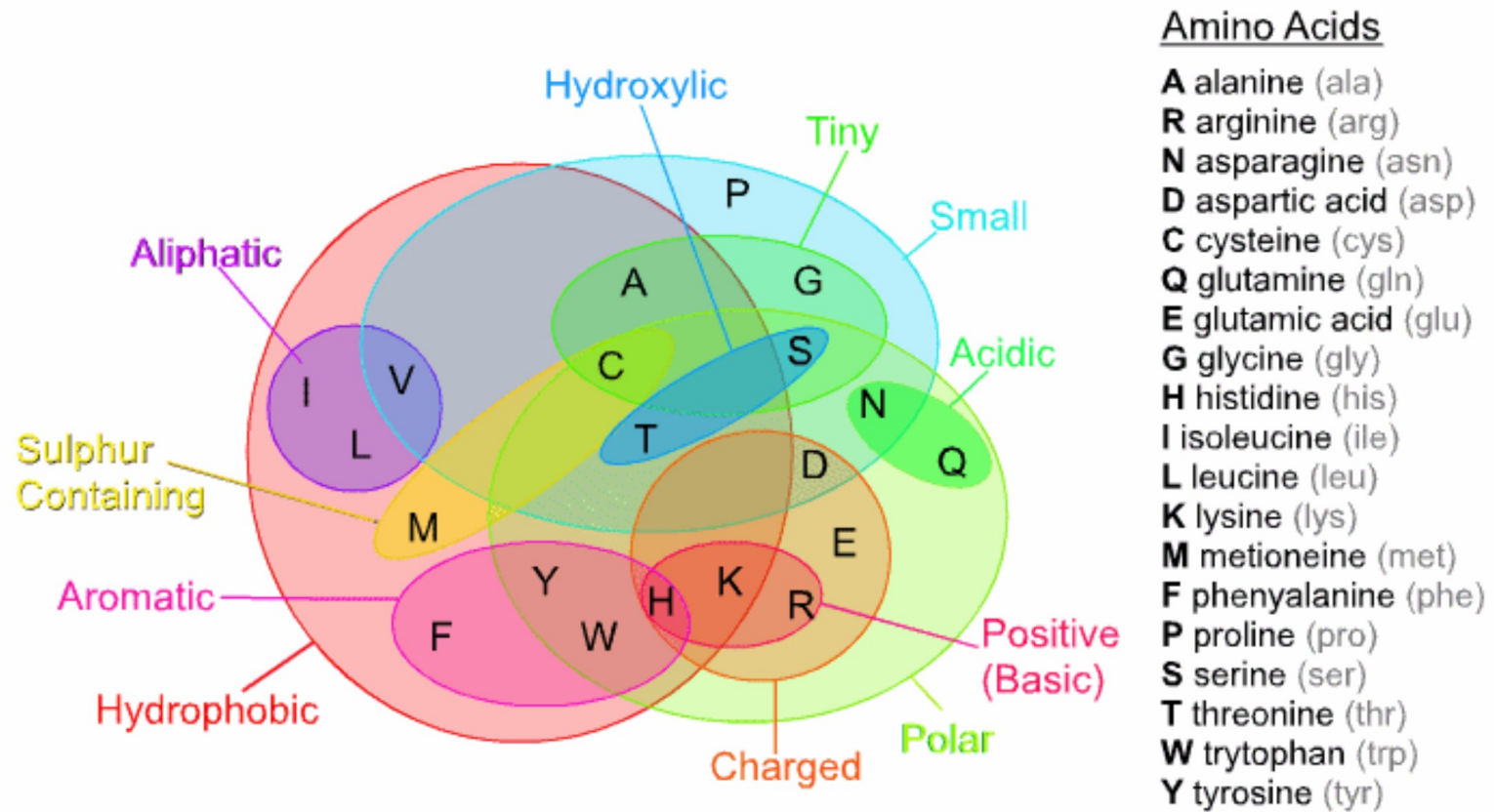
	23	123	123	123	1	
sp1_gnD	AT	GTG	TAT	CGT	C	..
sp2_gnD	AT	GTG	TCT	AGT	C	..
sp3_gnD	AT	GTG	CCT	CGT	C	..



	123	123	123	
sp1_gnD	..	Val	Tyr	Arg ...
sp2_gnD	..	Val	Ser	Ser ...
sp3_gnD	..	Val	Pro	Arg ...



# Amino acid physicochemical properties



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? And where do they occur in the sequence?*

Arg: positive, polar, charged  
Ser: small, tiny, hydroloxic, 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

Some codons may be used significantly more often than others to encode the same amino acid; 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

# Overview for Lab 16