Lecture 16
sequence statistics

Course: Practical Bioinformatics (BIOL 4220)
Instructor: Michael Landis
Email: 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 tccgttattcattcaatc
Mus_musculus_cytB tccgttatccacacaatc
Homo_sapiens_cytB tccgttattctctcaatc
Bos_taurus_cytB   tctgttattcactcaatc
```

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
Total	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
Total	2	8	6	2

Composition per sequence (across sites)

```
          123456
sp1_gnA   ACCTGT
sp2_gnA   ACTTGT
sp3_gnA   ACCTGA
```

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

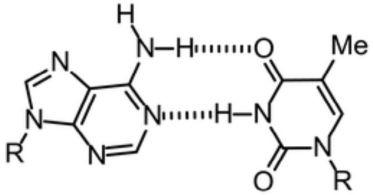
Composition per sequence (across sites)

```
          123456
sp1_gnB   TCGGGC
sp2_gnB   GCAGCC
sp3_gnB   GCACCT
```

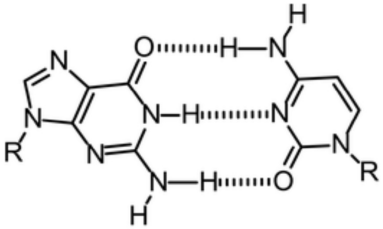
Species	A	C	G	T
sp1	-	2/6	3/6	1/6
sp2	1/6	3/6	2/6	-
sp3	1/6	3/6	1/6	1/6
Total	2	8	6	2

GC-content 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

```
          1234
spW_gnC   GCCT
spX_gnC   GCTG
spY_gnC   GACT
spZ_gnC   GACC
```

gene C

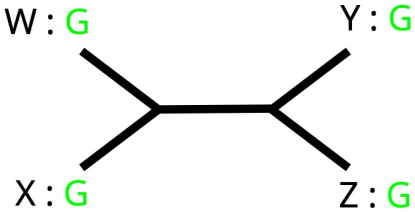
A site is ***phylogenetically informative*** if it can be used to estimate common ancestry

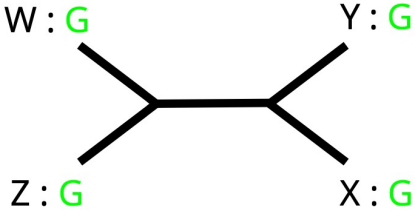
A phylogenetically informative site contains two variants, with at least two samples per variant

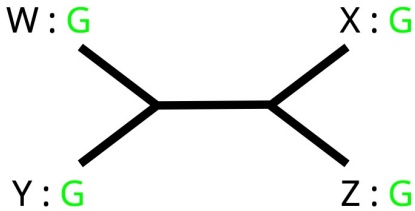
Phylogenetically informative sites

```
          1234
spW_gnC   GCCT
spX_gnC   GCTG
spY_gnC   GACT
spZ_gnC   GACC
```

Not PI



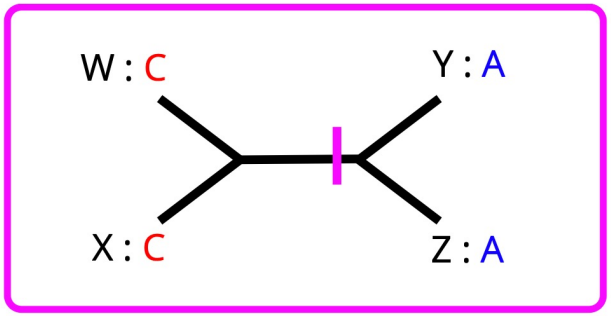


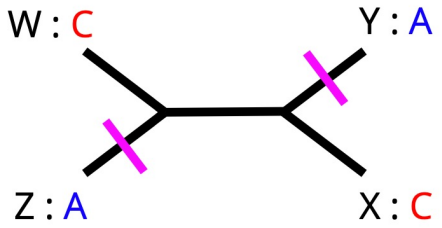


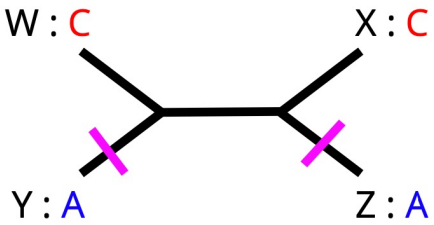
Phylogenetically informative sites

```
          1234
spW_gnC   GCCT
spX_gnC   GCTG
spY_gnC   GACT
spZ_gnC   GACC
```

Yes, PI



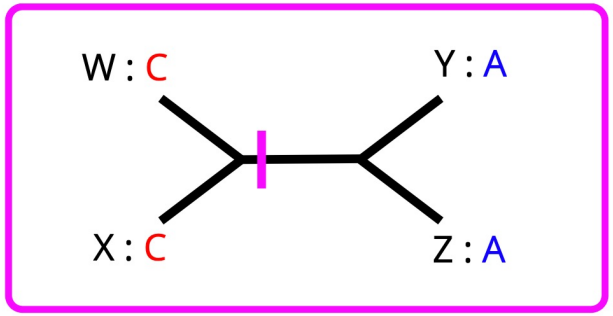


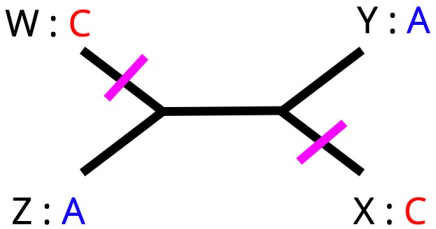


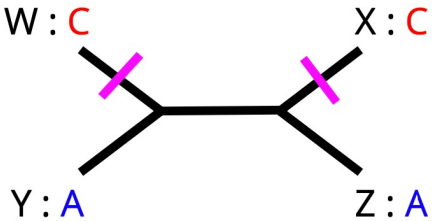
Phylogenetically informative sites

```
          1234
spW_gnC   GCCT
spX_gnC   GCTG
spY_gnC   GACT
spZ_gnC   GACC
```

Yes, PI



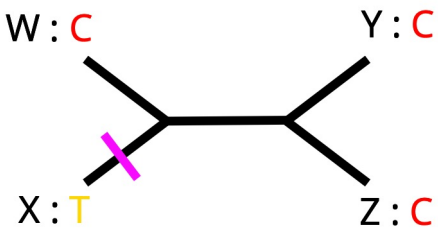


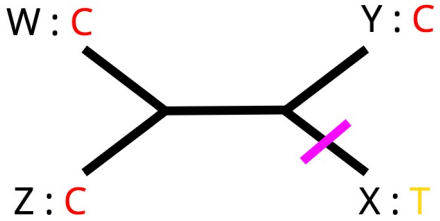


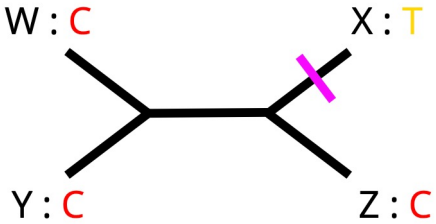
Phylogenetically informative sites

```
          1234
spW_gnC   GCCT
spX_gnC   GCTG
spY_gnC   GACT
spZ_gnC   GACC
```

Not PI



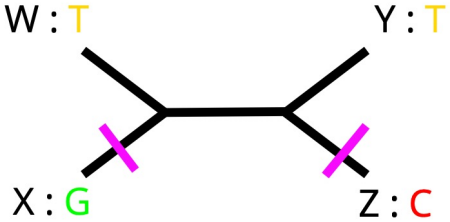


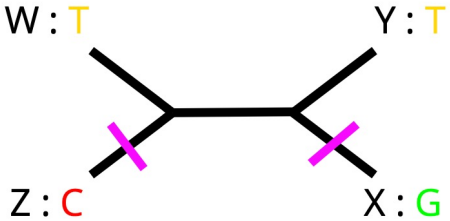


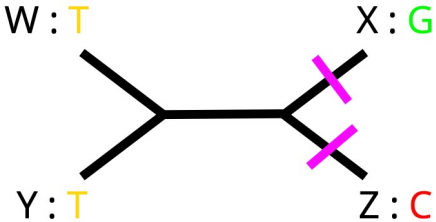
Phylogenetically informative sites

```
          1234
spW_gnC   GCCT
spX_gnC   GCTG
spY_gnC   GACT
spZ_gnC   GACC
```

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 ..
```

codons

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 ..
```

amino acids

Genetic code

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

Second letter

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

stop codons

start codon

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  ATG TGT ATC GTC ..
sp2_gnD  ATG TGT CTA GTC ..
sp3_gnD  ATG TGC CTC GTC ..
```

begin reading frame

Reading frame

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

```
         3  123  123  123  12
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 (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 ..
```

→

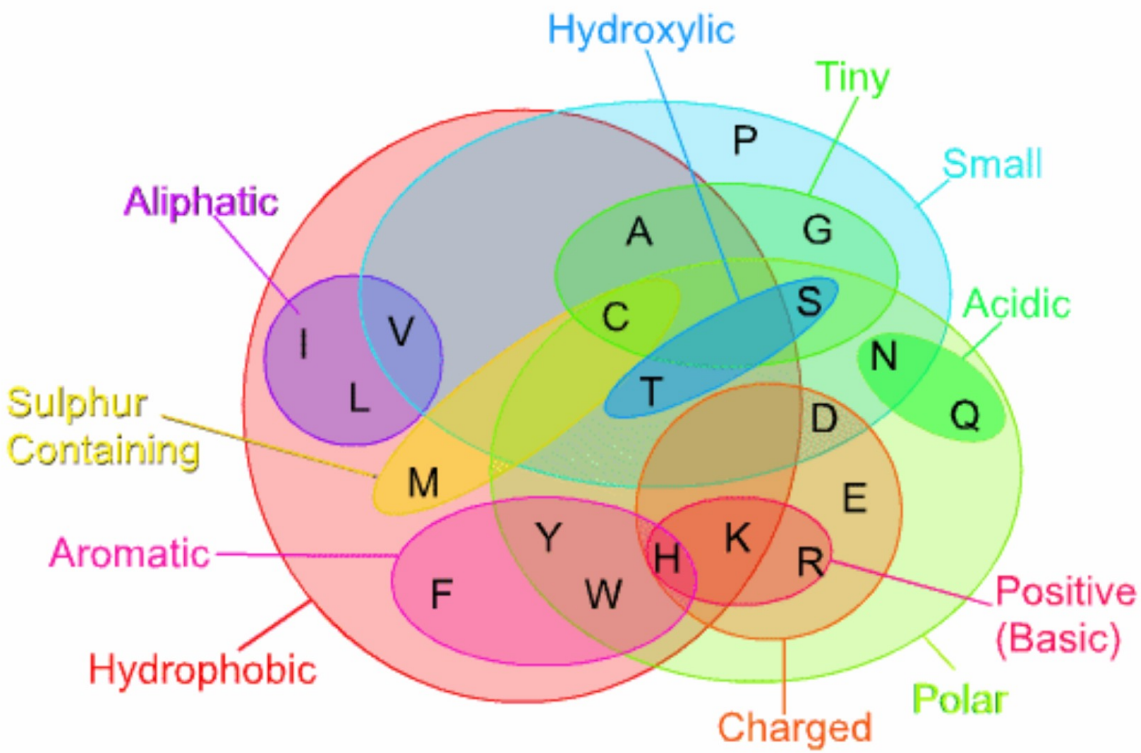
```
         123 123 123 123
sp1_gnD  Met Cys Ile Val ..
sp2_gnD  Met Cys Leu Val ..
sp3_gnD  Met Cys Leu Val ..
```

```
         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



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)

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

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

Arg: positive, polar, charged
Ser: **small**, tiny, hydroloxic, polar
Tyr: aromatic, hydrophobic, polar
Pro: **small**
Val: **small**, hydrophobic, aliphatic

Amino acid frequencies

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

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

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