CS 364 COMPUTATIONAL BIOLOGY

Sara Mathieson
Haverford College

Outline

Lab 7 *and* proposal due tonight!
Office hours moved to Tuesdays
2:30-3:30pm in H110
Lab 8 (last lab) posted

- Finish coalescent theory
- Application: Tajima's D for natural selection
- Begin: Markov models

Finish coalescent theory

Coalescent derivation from the Wright-Fisher model

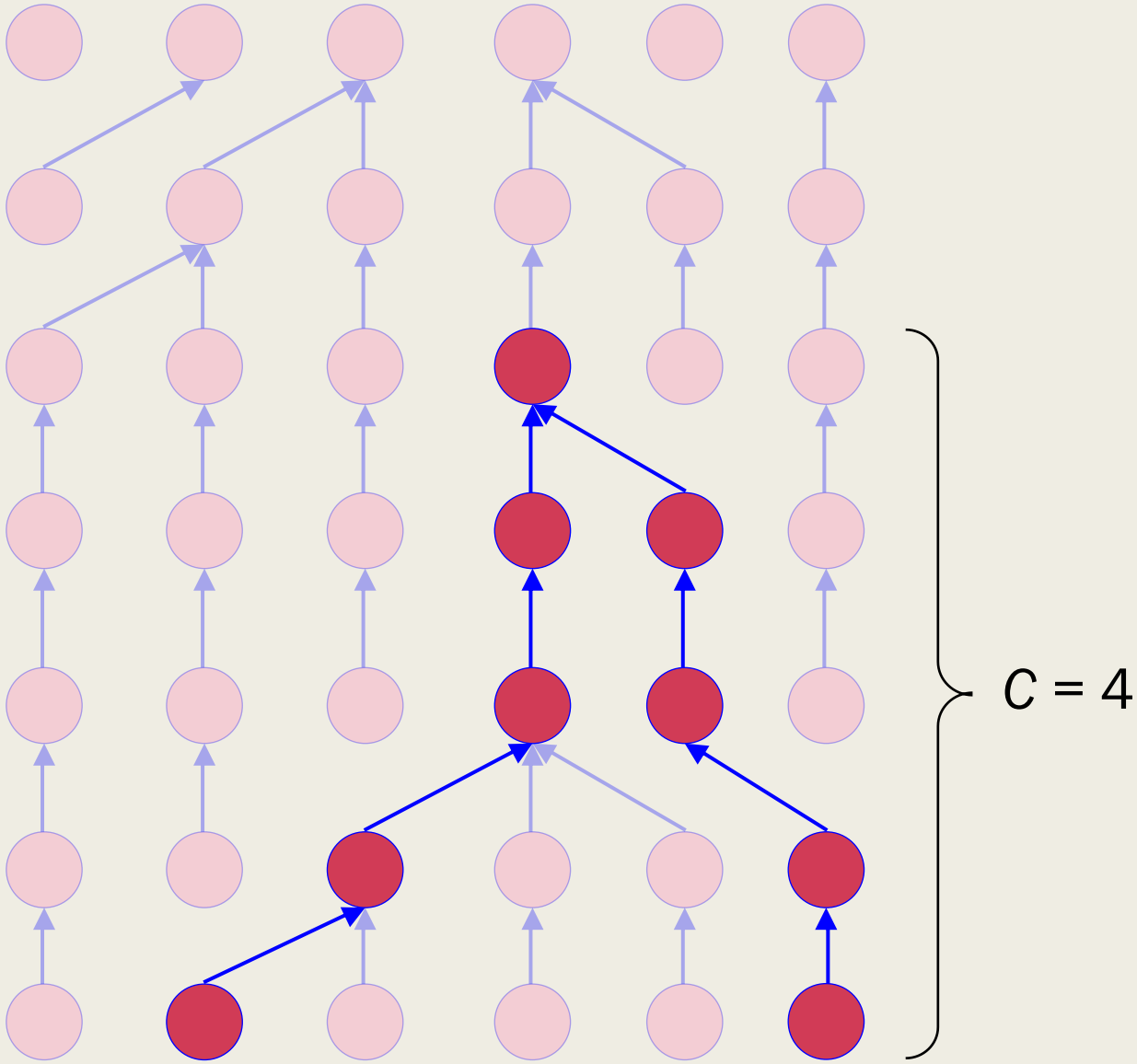
Probability two samples *coalesce* after *g* generations:

$$P_C(g) = \left(1 - \frac{1}{2N}\right)^{g-1} \frac{1}{2N}$$

Don't choose the same parent for *g*-1 generations

Choose same parent in the g^{th} generation

[Geometric distribution]



Population size $2N$=6, sample size n = 2

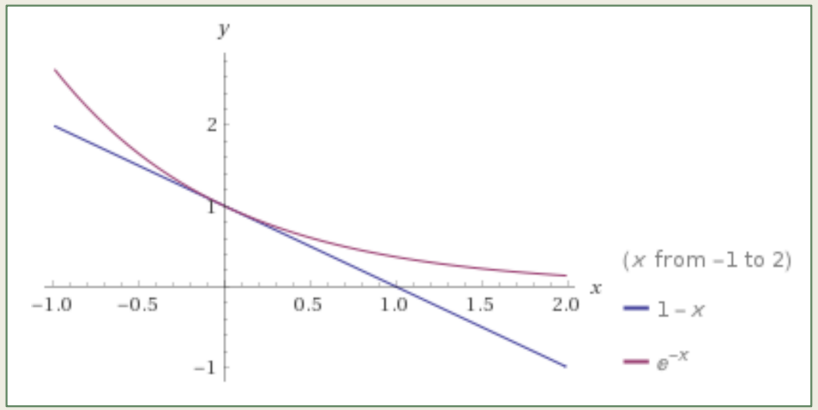
Coalescent derivation from the Wright-Fisher model

- We will make use of the Taylor series for e^{-x} around $x = 0$:

$$e^{-x} = 1 - x + \frac{x^2}{2!} - \frac{x^3}{3!} + \frac{x^4}{4!} + \cdots$$

- We will only use the first 2 terms:

$$e^{-x} \approx 1 - x$$



Created using WolframAlpha

- This allows us to rewrite our geometric coalescent probability

$$P_C(g) = \left(1 - \frac{1}{2N}\right)^{g-1} \frac{1}{2N}$$

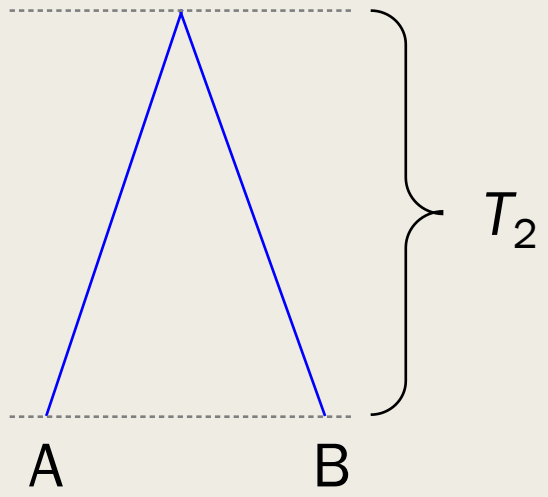
- as (drop the -1 since g is large):

$$P_C(g) \approx \frac{1}{2N} e^{-\frac{g}{2N}}$$

Coalescent for $n = 2$

- We let 1 coalescent unit = $2N$ generations, and let our new variable be t

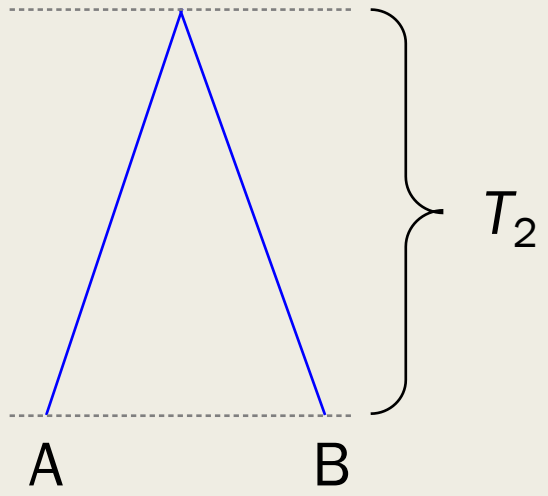
- We let T_i be a random variable representing the time when there are i lineages



Coalescent for $n = 2$

- We let 1 coalescent unit = $2N$ generations, and let our new variable be t
- We let T_i be a random variable representing the time when there are i lineages
- For n=2, this gives us an exponential distribution with parameter 1

$$P_{T_2}(t) = e^{-t}$$

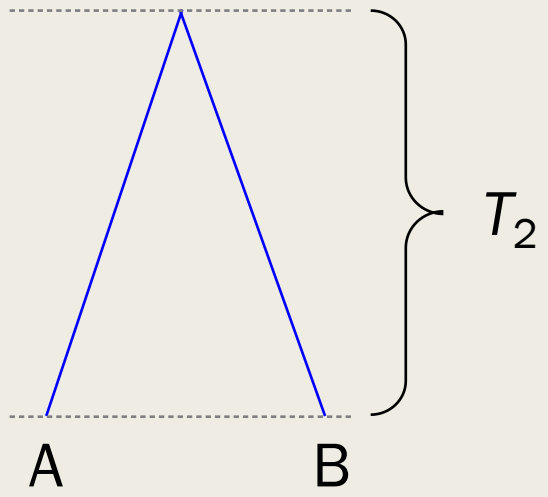


Coalescent for *n* = 2

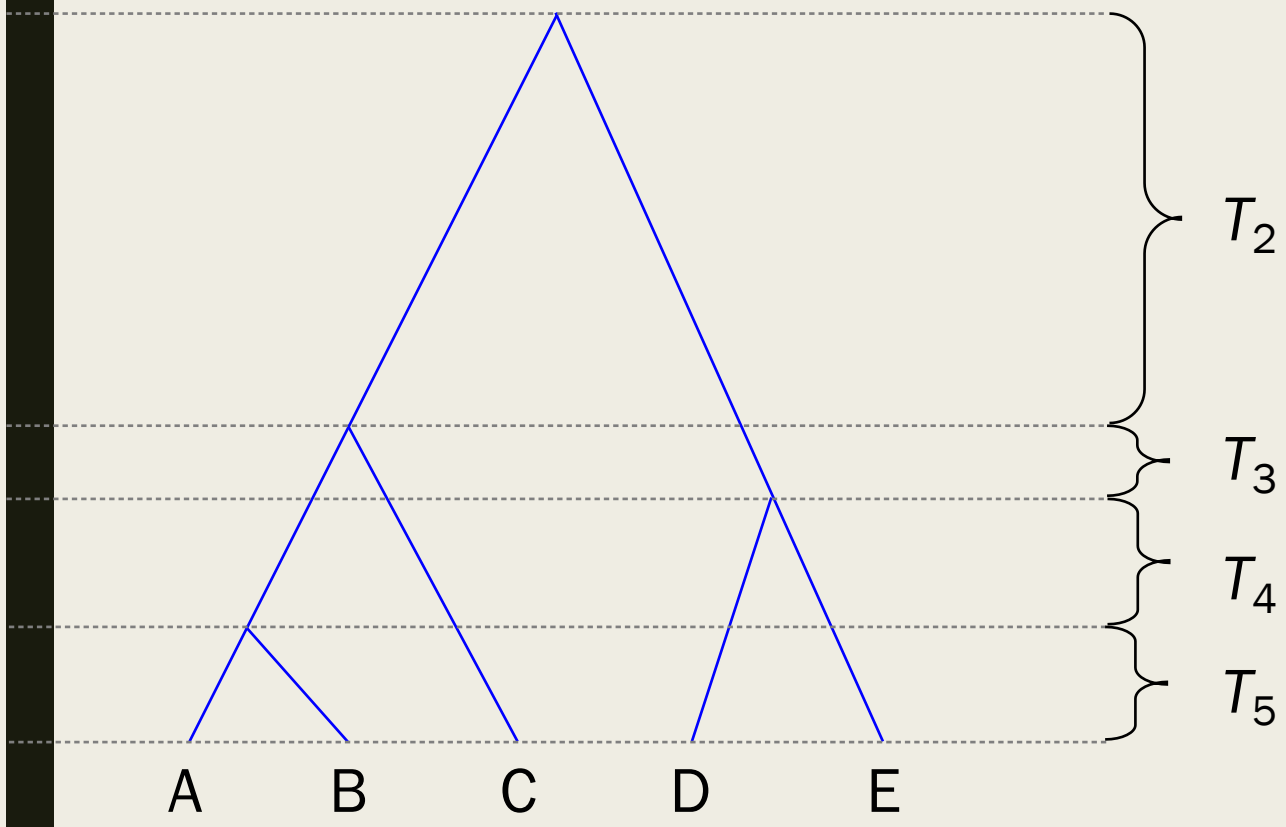
- We let 1 coalescent unit = 2*N* generations, and let our new variable be *t*
- We let T_i be a random variable representing the time when there are *i* lineages
- For *n*=2, this gives us an exponential distribution with parameter 1
- The expected time for 2 lineages to coalesce is 1 coalescent unit of time => 2N generations

$$P_{T_2}(t) = e^{-t}$$

$$E[T_2] = \int_0^\infty te^{-t}dt = 1$$



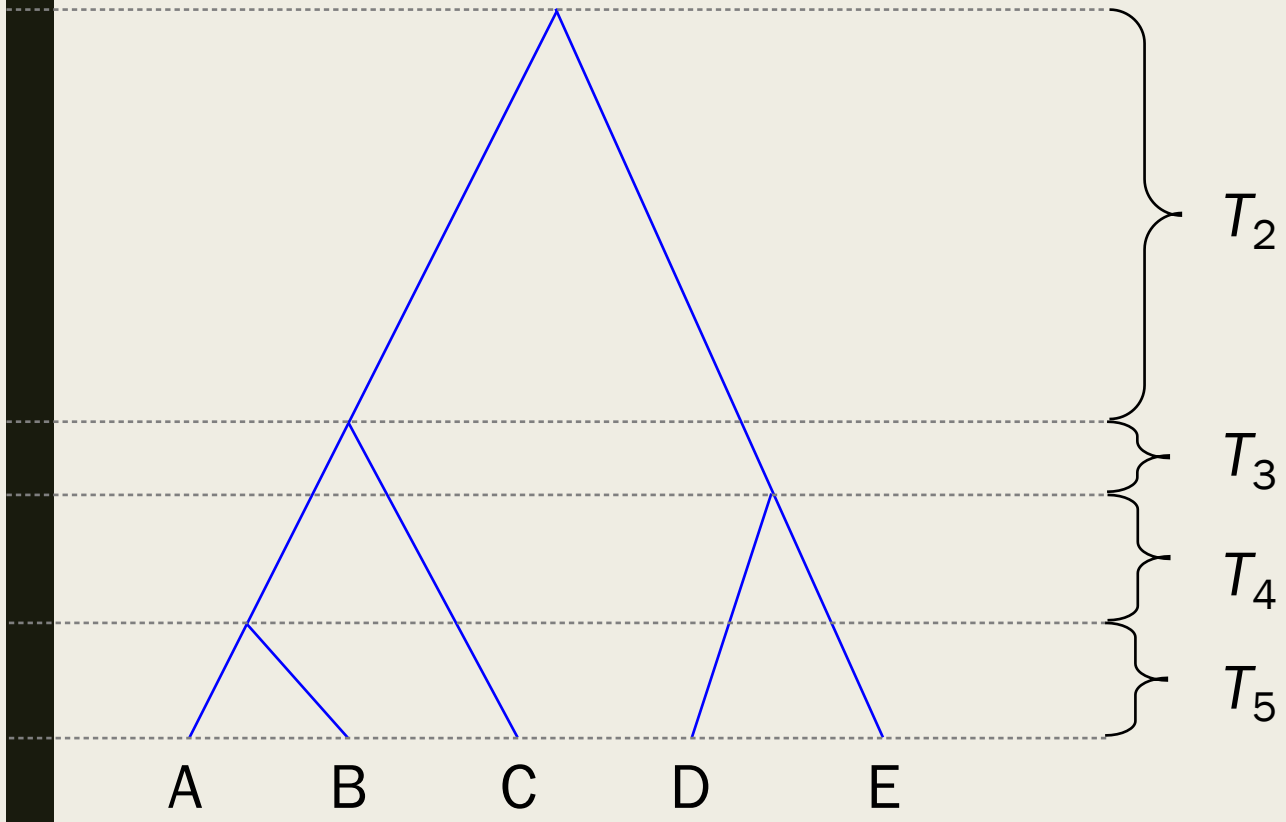
The Coalescent



- The larger our sample size n, the more pairs we have that can coalesce right away
- In general, the time when there are *i* lineages is also exponentially distributed with parameter *i*(*i*-1)/2 (*i* "choose" 2)

$$P_{T_i}(t) = \binom{i}{2} e^{-\binom{i}{2}t}$$

The Coalescent



- The larger our sample size n, the more pairs we have that can coalesce right away
- In general, the time when there are *i* lineages is also exponentially distributed with parameter *i*(*i*-1)/2 (*i* "choose" 2)

$$P_{T_i}(t) = \binom{i}{2} e^{-\binom{i}{2}t}$$

- Expected value (think: weighted average, mean)

$$E[T_i] = \int_0^\infty t \binom{i}{2} e^{-\binom{i}{2}t} dt = \frac{1}{\binom{i}{2}}$$

Deviations from neutrality: Tajima’s D

Tajima's D

- We often say a site/locus is "neutral" if it has no positive or negative effect on fitness
- More generally, "neutral" means agreeing with our Wright-Fisher model assumptions (constant population size, mutations have no consequences, random mating, etc)
- Deviations from neutrality could mean that any of these assumptions are wrong
- We will focus on two of them: allowing variable population size and allowing mutations with different selective advantages/disadvantages
- Tajima's D (1989) is a test statistic that compares different measures of sequence diversity that should be the same under neutrality
- If they are not the same, we can further investigate the causes

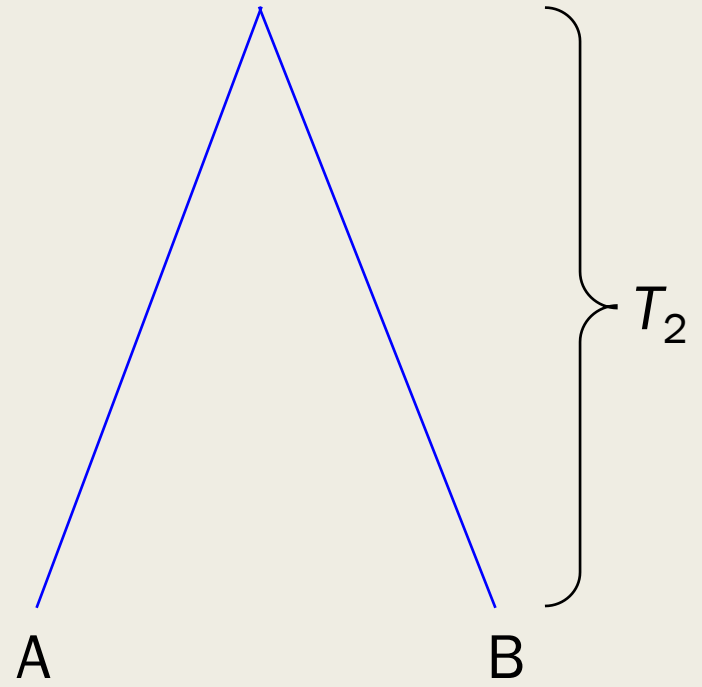
Expected values of S (number of segregating sites) and π (average pairwise heterozygosity)

- For now we will consider a single site

- Let μ be the per site, per generation mutation rate

Expected values of S (number of segregating sites) and π (average pairwise heterozygosity)

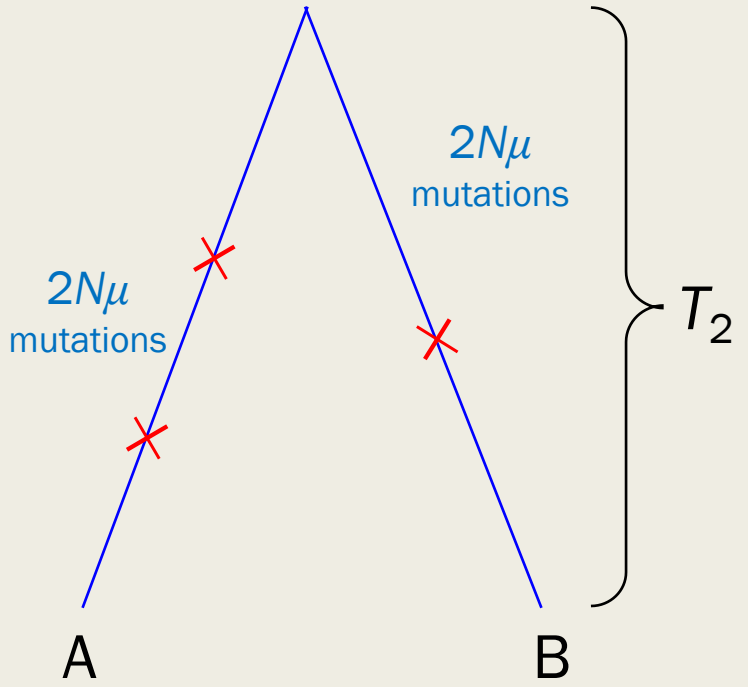
- For now we will consider a single site
- Let μ be the per site, per generation mutation rate
- Considering two samples, the expected time to coalescence is 1 coalescent unit or $2N$ generations



Expected values of S (number of segregating sites) and π (average pairwise heterozygosity)

- For now we will consider a single site
- Let μ be the per site, per generation mutation rate
- Considering two samples, the expected time to coalescence is 1 coalescent unit or 2*N* generations
- Therefore the expected number of mutations separating the two samples is

$$E[\pi] = 4N\mu = \theta$$



Expected values of S (number of segregating sites) and π (average pairwise heterozygosity)

- For E[S], we need to compute the total branch length

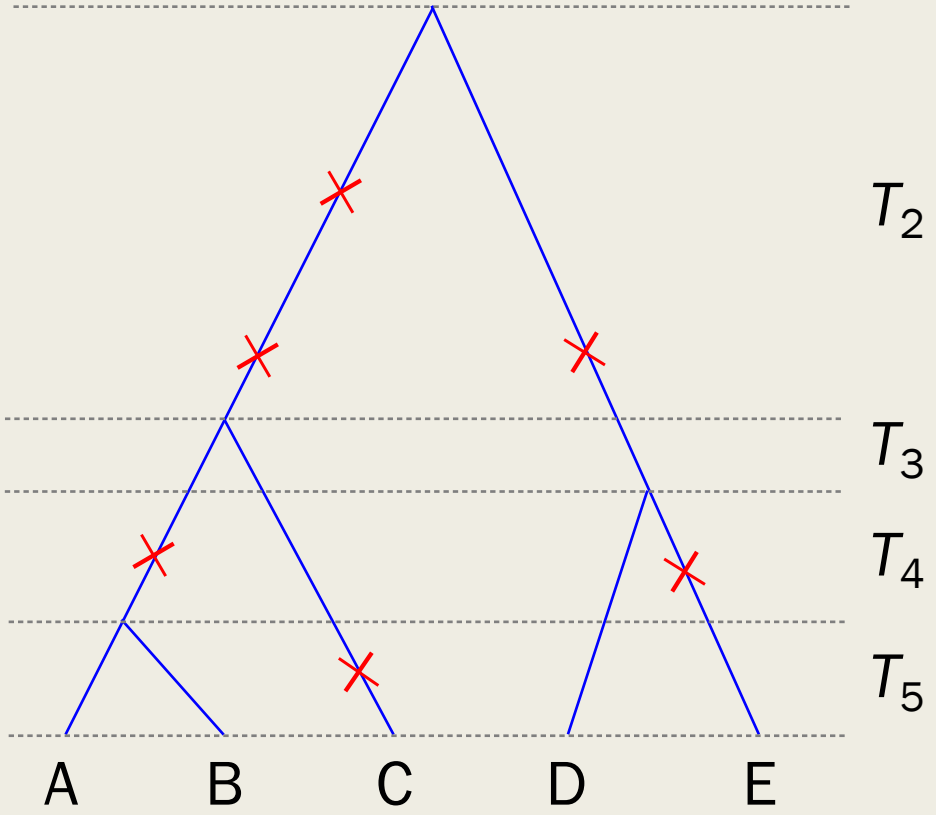
$T_{\text{total}} = \text{total length of all branches in the tree}$

$$E[T_{\text{total}}] = \sum_{i=n}^{2} E[T_i] \cdot i$$

$$= \sum_{i=n}^{2} \frac{2}{\cancel{i}(i-1)} \cdot \cancel{i}$$

$$= 2\sum_{i=1}^{n-1} \frac{1}{i}$$

$$= 2a_1$$



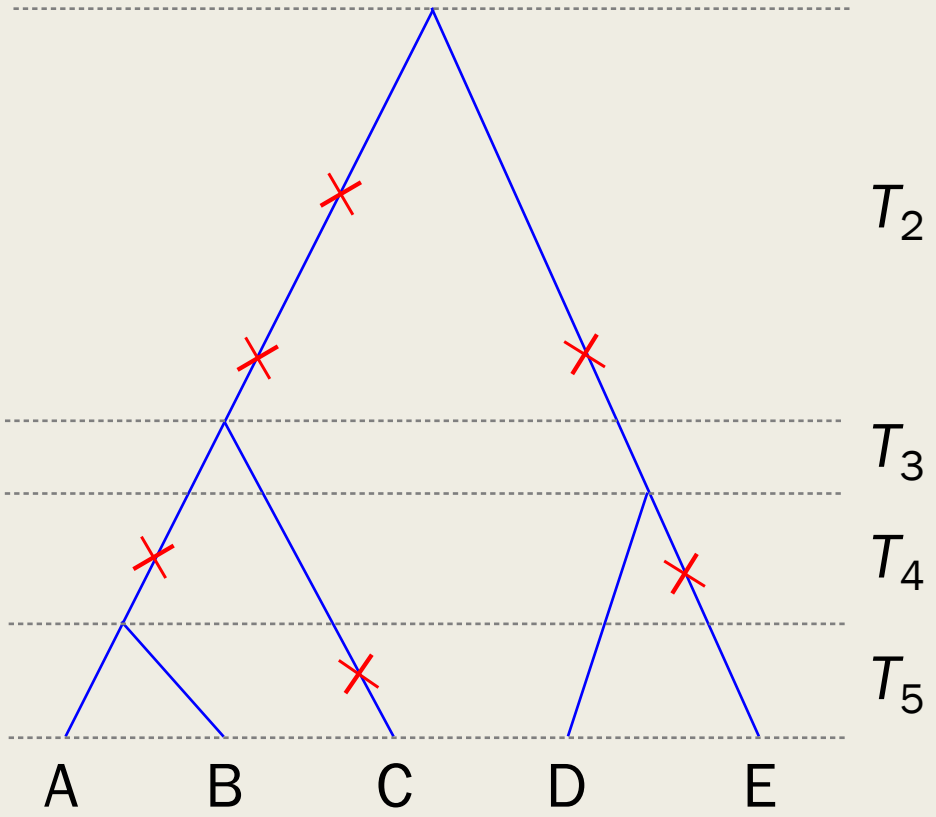
Expected values of S (number of segregating sites) and π (average pairwise heterozygosity)

- After we have the total branch length, we can multiple by $2N\mu$, the rate of mutations per unit of coalescent time

$$E[S] = E[T_{\text{total}}] \cdot (2N\mu)$$

- We can simplify this to get an expression similar to the expected value for π

$$E[S] = 4N\mu \cdot a_1 = \theta a_1$$



Putting this together, we get Tajima's d

- We will consider lowercase *d*, whose expectation is E[*d*] = 0

$$d = \pi - S/a_1$$

- Tajima's (capital) *D* is defined as:

$$D = \frac{d}{\sqrt{\mathrm{Var}(d)}}$$

- We will mainly focus on the sign of *d* so we'll ignore the denominator

What do deviations from *d*=0 mean?

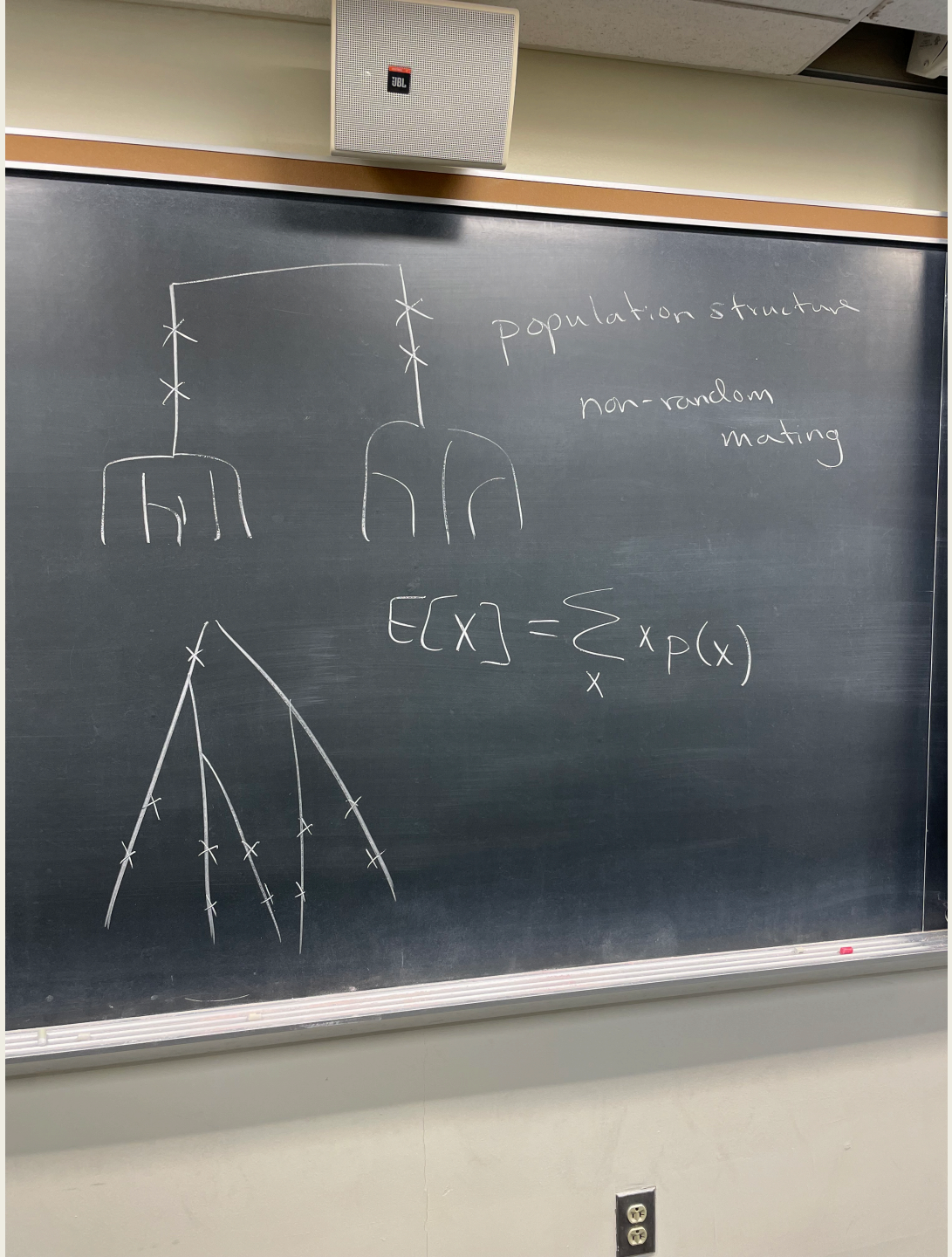
- If d is close to 0, neutral expectations (probably) hold (i.e. constant population size, random mating, no natural selection

- If d > 0, the pairwise heterozygosity is higher than we expect relative to the number of segregating sites => excess of **middle** frequency SNPs

  - Bottleneck or population decline
  - Population structure or isolation with migration

- If d < 0, the number of segregating sites is higher than we expect relative to the pairwise heterozygosity => excess of **rare** variation

  - Population growth
  - Natural selection


JBL
population structure
non-random
mating
$E[X] = \sum_x x\,p(x)$

Tajima’s D in practice

Example of Tajima's D from the literature

Genomic regions exhibiting positive selection identified from dense genotype data

Christopher S. Carlson,[1,3] Daryl J. Thomas,[2] Michael A. Eberle,[1] Johanna E. Swanson,[1] Robert J. Livingston,[1] Mark J. Rieder,[1] and Deborah A. Nickerson[1]

[1]Department of Genome Sciences, University of Washington, Seattle, Washington 98195-7730, USA; [2]Center for Biomolecular Science and Engineering, University of California, Santa Cruz, California 95064-1099, USA

- Why is Tajima's D greater than 0?
- Hypothesis: bottleneck in European and Asian populations is still affecting patterns of variation
- Population structure is playing a role in African populations

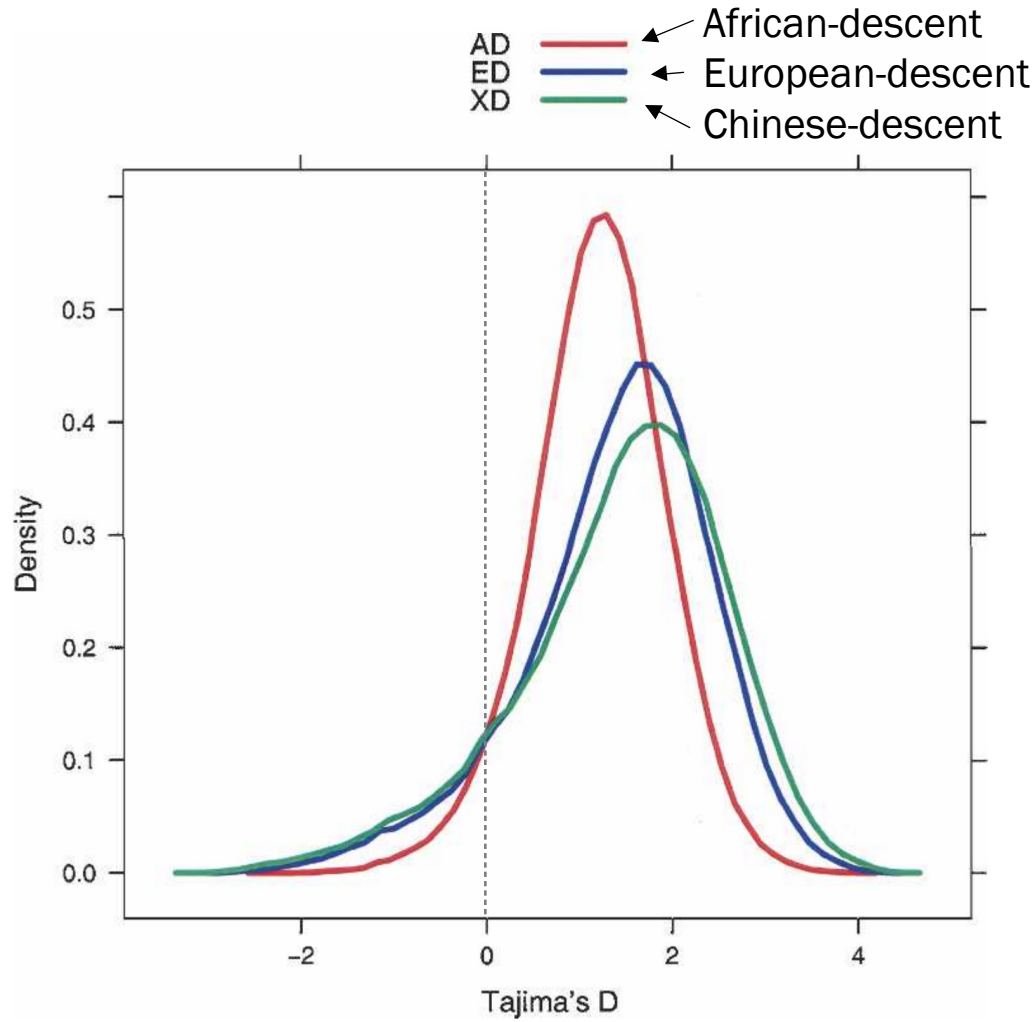


Figure 2. A probability density plot of the distribution of Tajima's D in the sliding windows is shown for each population. All three distributions depart significantly from a normal distribution, most noticeably in the heavy tail at low values in each population.

Example of Tajima's D from the literature

Genomic regions exhibiting positive selection identified from dense genotype data

Christopher S. Carlson,[1,3] Daryl J. Thomas,[2] Michael A. Eberle,[1] Johanna E. Swanson,[1] Robert J. Livingston,[1] Mark J. Rieder,[1] and Deborah A. Nickerson[1]

[1]Department of Genome Sciences, University of Washington, Seattle, Washington 98195-7730, USA; [2]Center for Biomolecular Science and Engineering, University of California, Santa Cruz, California 95064-1099, USA

- Regions where Tajima's *D* < 0, probably natural selection (could be random)

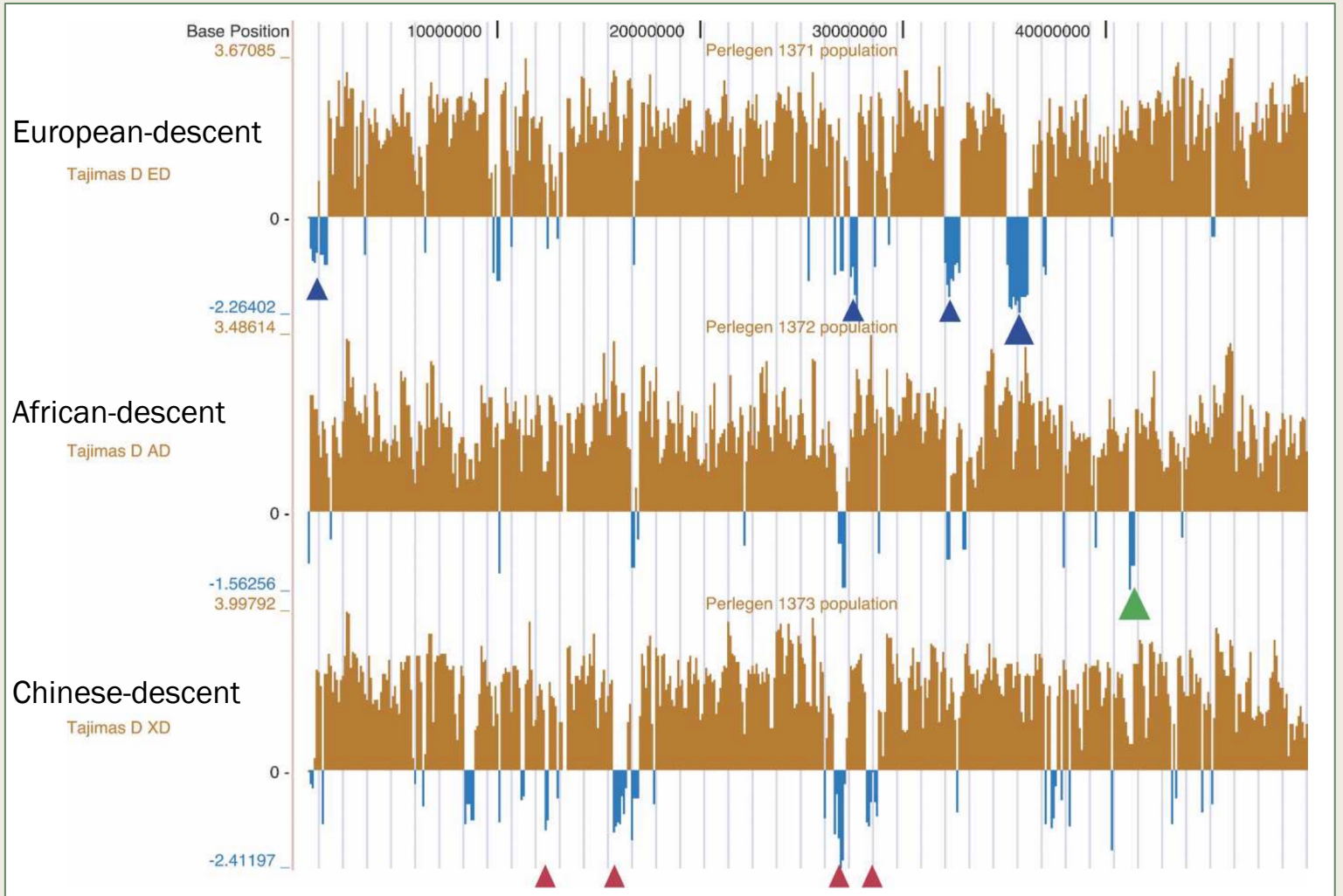


Figure 3. Tajima's D in 100-kbp sliding windows with 10-kbp steps is shown across the first 50 megabases of chromosome 1. Several CRTRs are visible, including a region near 35M in the ED population containing *CLSPN* (large blue arrowhead) and a region near 41M in the AD population spanning *CTPS*, *FLJ23878*, and *SCMH1* (large green arrowhead). CRTRs at the less stringent 5% level are also indicated in the ED population as small blue arrowheads and in the XD population as small red arrowheads.

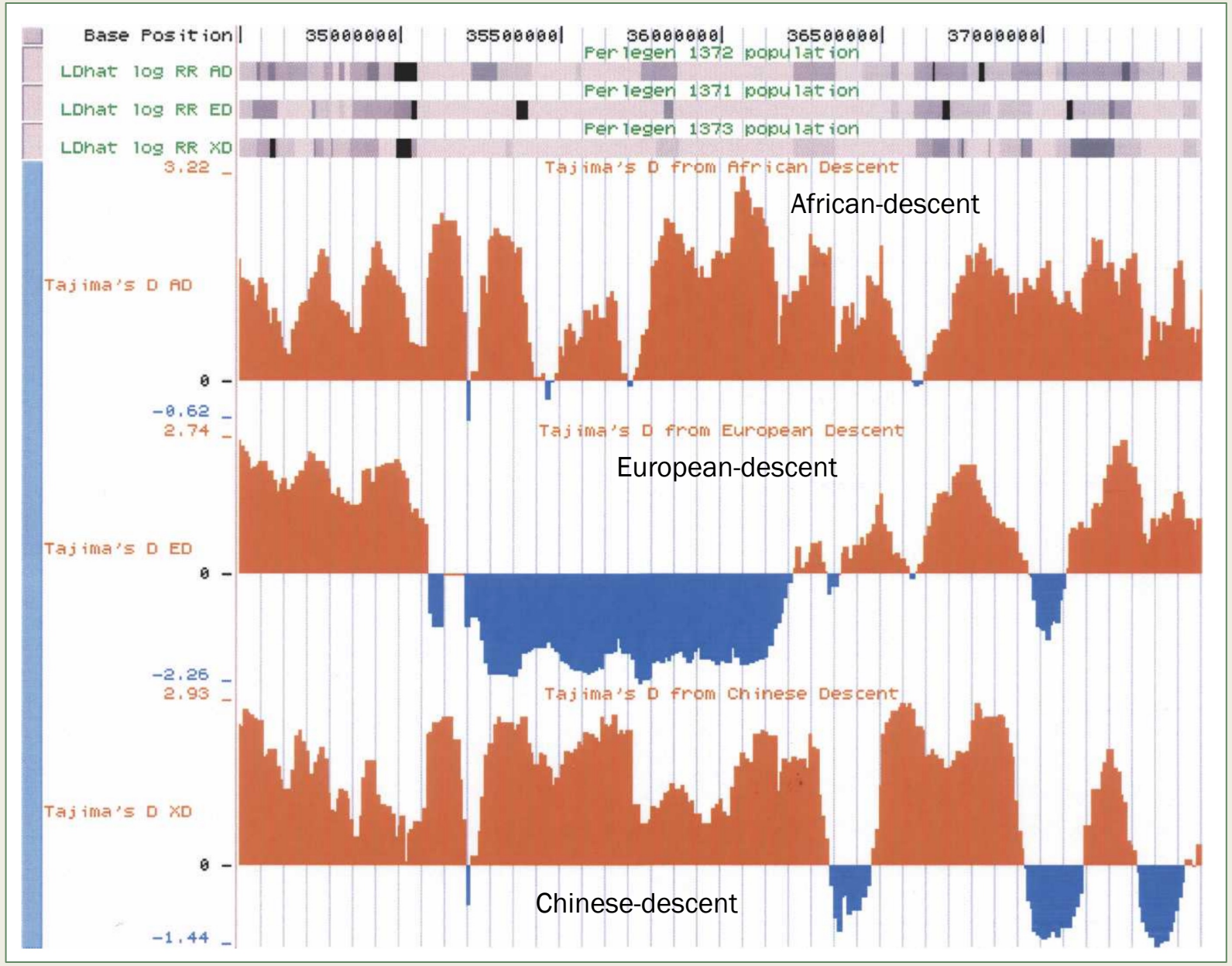
Example of Tajima's D from the literature

Genomic regions exhibiting positive selection identified from dense genotype data

Christopher S. Carlson,[1,3] Daryl J. Thomas,[2] Michael A. Eberle,[1] Johanna E. Swanson,[1] Robert J. Livingston,[1] Mark J. Rieder,[1] and Deborah A. Nickerson[1]

[1]Department of Genome Sciences, University of Washington, Seattle, Washington 98195-7730, USA; [2]Center for Biomolecular Science and Engineering, University of California, Santa Cruz, California 95064-1099, USA

- Extended regions of low *D*, could be strong selection
- This paper found several regions under selection in European and Chinese populations that are linked to drug metabolism



Example of Tajima's D from the literature

Genomic regions exhibiting positive selection identified from dense genotype data

Christopher S. Carlson,[1,3] Daryl J. Thomas,[2] Michael A. Eberle,[1] Johanna E. Swanson,[1] Robert J. Livingston,[1] Mark J. Rieder,[1] and Deborah A. Nickerson[1]

[1]Department of Genome Sciences, University of Washington, Seattle, Washington 98195-7730, USA; [2]Center for Biomolecular Science and Engineering, University of California, Santa Cruz, California 95064-1099, USA

Note: not our formulas!
But exactly the same idea/goal

Nucleotide diversity analysis

There are several statistics that can be used to describe nucleotide diversity, including θ_s (equation 1), π (equation 2), and θ_H (equation 3). These statistics can be calculated for a given resequencing data set by using the following parameters: n is the number of chromosomes resequenced, Sn is the number of polymorphic sites observed, p_i is the derived (nonancestral) allele frequency of the ith SNP, and q_i is the ancestral allele frequency of the ith SNP.

$$\theta_s = \frac{Sn}{\sum_{i=1}^{n-1}\frac{1}{i}} \quad (1)$$

$$\pi = \frac{n}{n-1}\sum_{i=1}^{Sn} 2p_i q_i \quad (2)$$

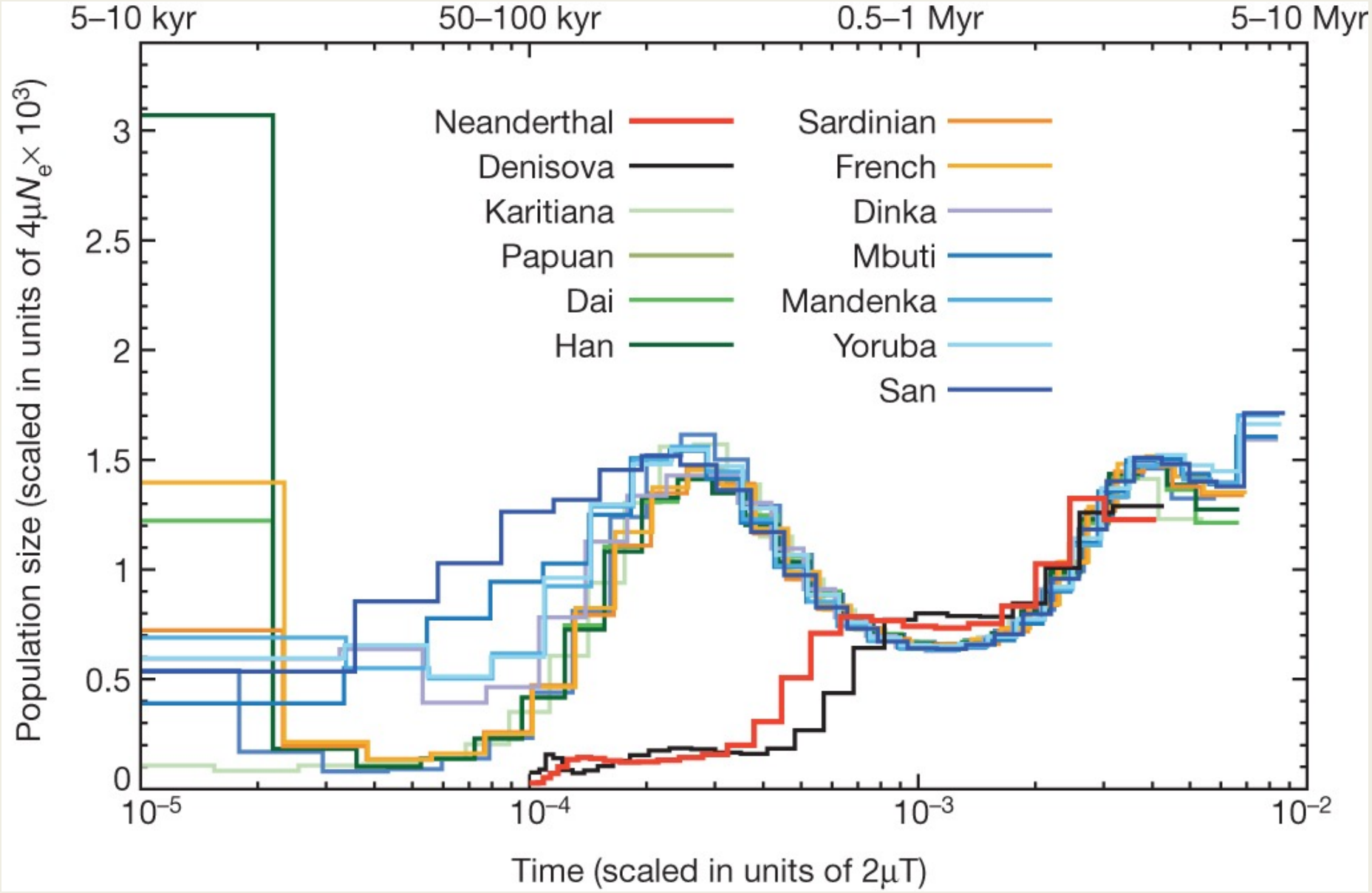
$$\theta_H = \frac{n}{n-1}\sum_{i=1}^{Sn} 2p_i^2 \quad (3)$$

There are many statistics that can evaluate departures from the expected patterns of neutral variation. One of these is Tajima's D (Tajima 1989), equation 4:

$$D = \frac{\pi - \theta_s}{\sqrt{Var(\pi - \theta_s)}} \quad (4)$$

Markov Chains

HMM example from the literature (similar to Lab 8)



K Prüfer *et al. Nature* **000**, 1-7 (2013) doi:10.1038/nature12886

Conditional probability

- Idea of conditional probability

$$P(A, B) = P(A)P(B|A)$$

- Bayes' Theorem

$$P(A)P(B|A) = P(B)P(A|B)$$

OH: Zubrow

u = bring umbrella

r = rain

s = sun

$W \in \{r, s\}$

↑ weather

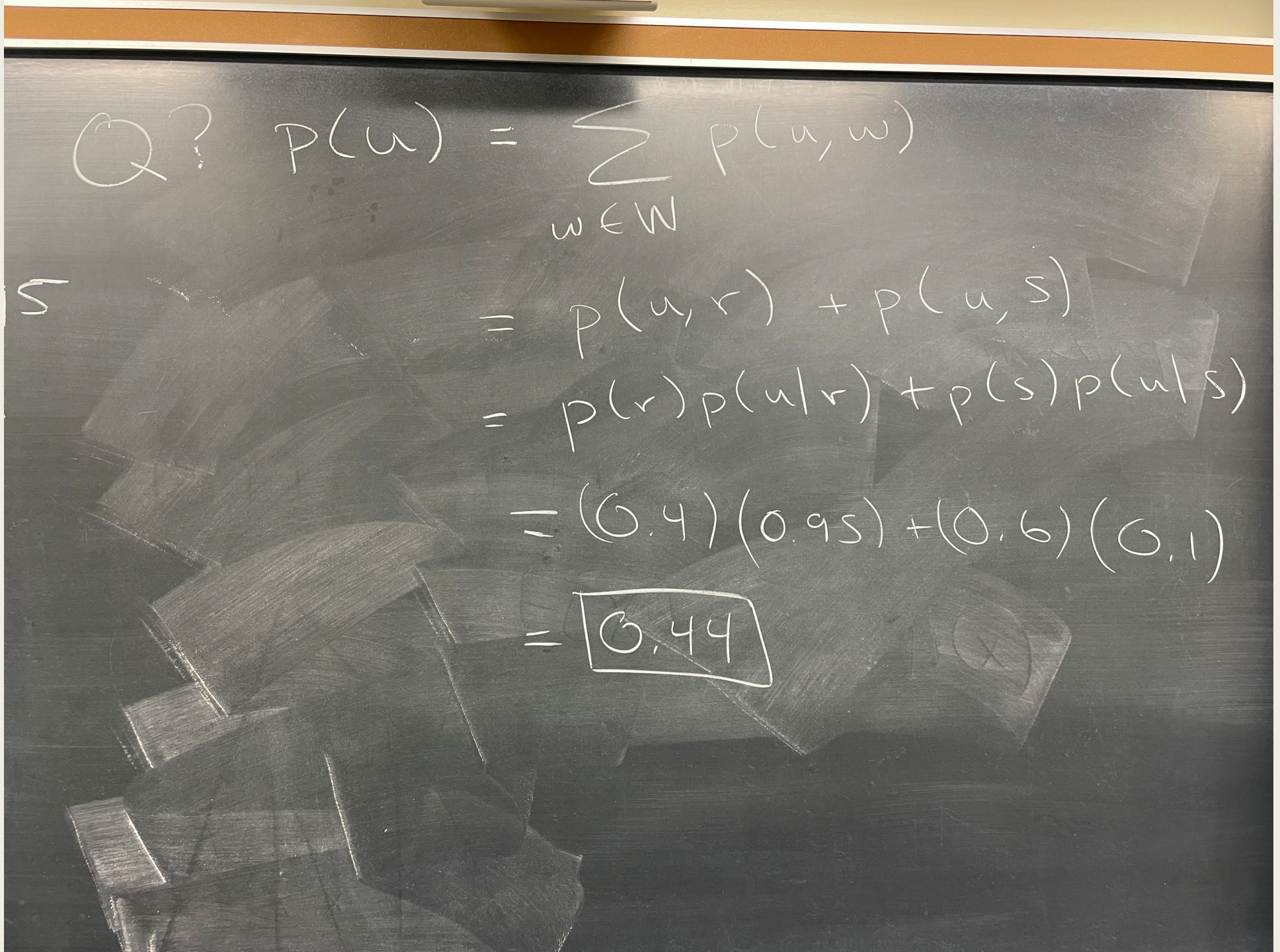
give

$p(u|r) = 0.95$

$p(u|s) = 0.1$

add to 1 $\begin{cases} p(r) = 0.4 \\ p(s) = 0.6 \end{cases}$

$$p(A) = \sum_{b \in vals(B)} p(A, B = b)$$


Q? $p(u) = \sum_{w \in W} p(u,w)$
$= p(u,r) + p(u,s)$
$= p(r)p(u|r) + p(s)p(u|s)$
$= (0.4)(0.95) + (0.6)(0.1)$
$= \boxed{0.44}$

Markov Chains

- Markov assumption: current state only depends on the previous state

$$P(z_i|z_0, z_1, \cdots, z_{i-1}) = P(z_i|z_{i-1})$$

Markov Chains

- Markov assumption: current state only depends on the previous state

$$P(z_i|z_0, z_1, \cdots, z_{i-1}) = P(z_i|z_{i-1})$$

- This allows us to simplify the probability of observing a Markov chain:

$$P(z_0, z_1, \cdots, z_L) = P(z_0) \prod_{i=1}^{L} P(z_i|z_{i-1})$$

Markov Chains

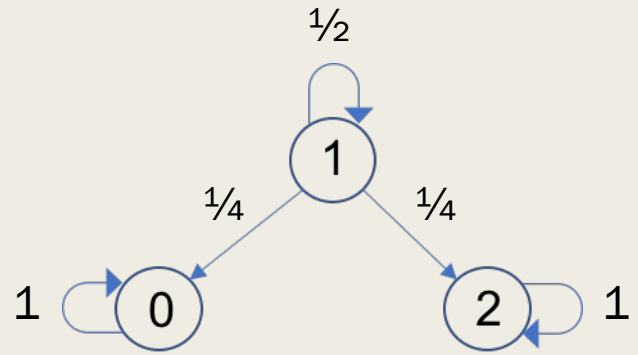
- Markov assumption: current state only depends on the previous state

$$P(z_i|z_0, z_1, \cdots, z_{i-1}) = P(z_i|z_{i-1})$$

- This allows us to simplify the probability of observing a Markov chain:

$$P(z_0, z_1, \cdots, z_L) = P(z_0) \prod_{i=1}^{L} P(z_i|z_{i-1})$$

- Note the difference between the state diagram (right) and an observed state sequence



Markov Chains

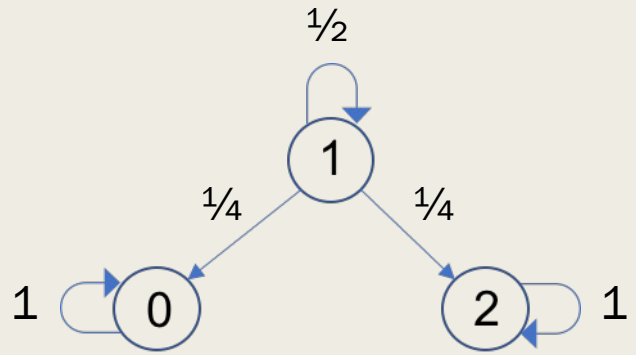
- Markov assumption: current state only depends on the previous state

$$P(z_i|z_0, z_1, \cdots, z_{i-1}) = P(z_i|z_{i-1})$$

- This allows us to simplify the probability of observing a Markov chain:

$$P(z_0, z_1, \cdots, z_L) = P(z_0) \prod_{i=1}^{L} P(z_i|z_{i-1})$$

Note that the sum of *outgoing* probabilities should be 1

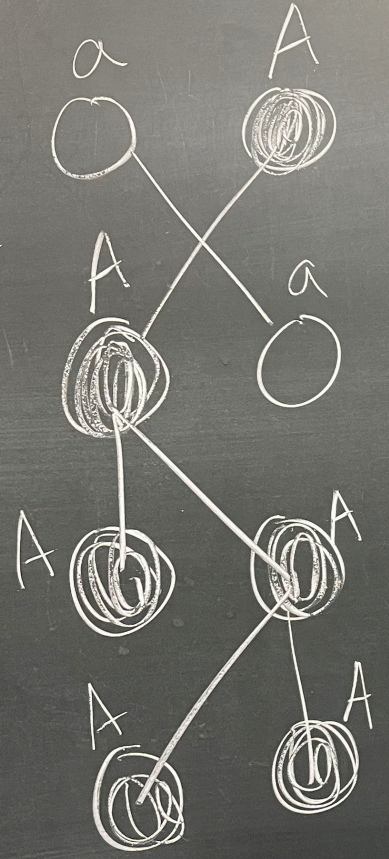


- Note the difference between the state diagram (right) and an observed state sequence

Markov Chain

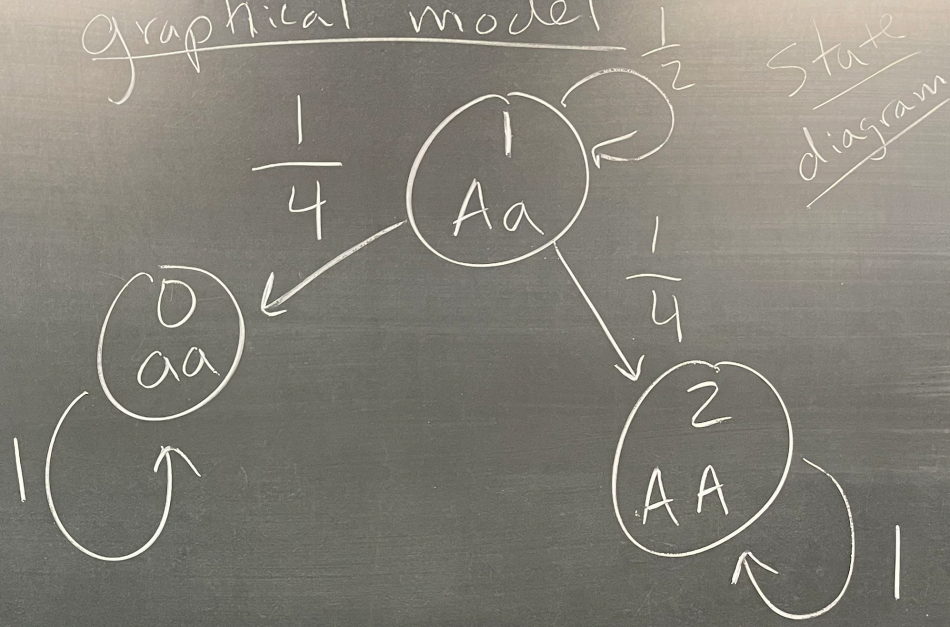
N = 1

2N = 2



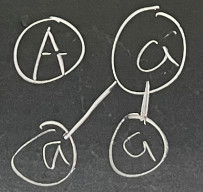
(no new mutation)

graphical model



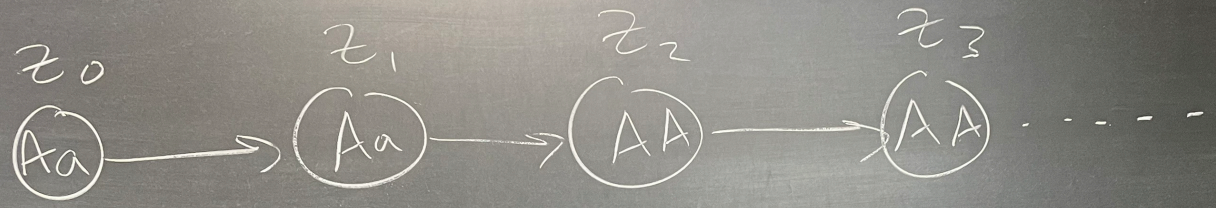
State diagram

Sum of outgoing arrows should be 1



Aa
aA
aa
AA

Chain (state sequence)



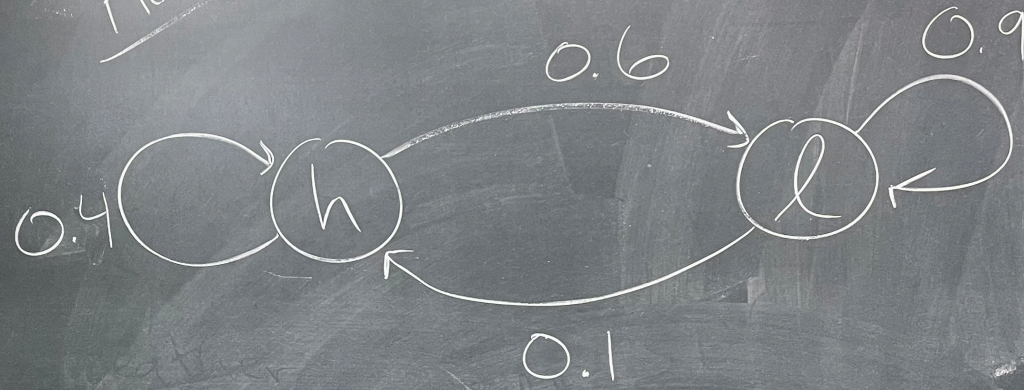
Markov property: current state only depends on the previous state

$$p(z_i \mid z_0, z_1 \cdots z_{i-1}) = p(z_i \mid z_{i-1})$$

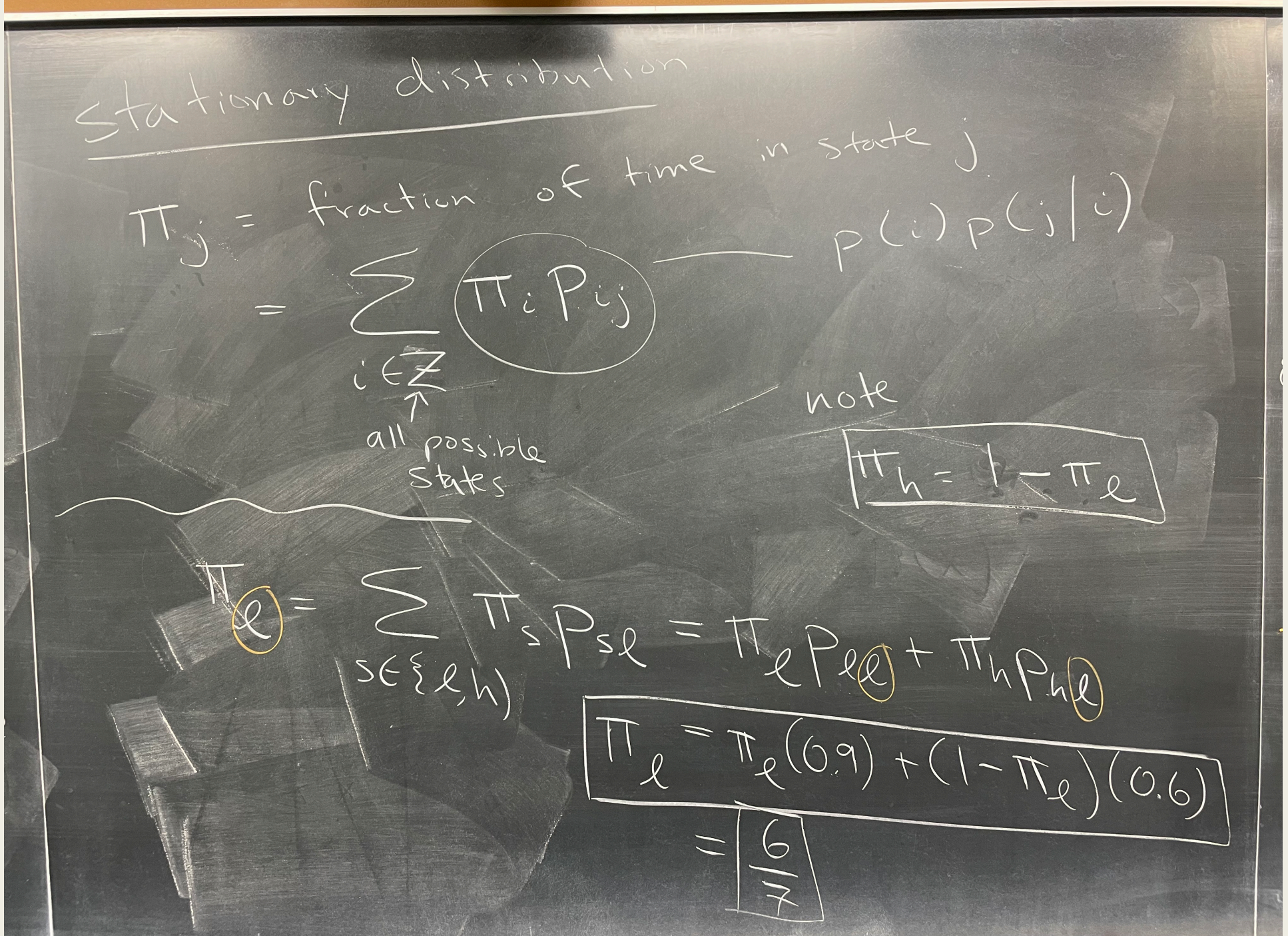
Handout 19

OH: Zubrow

Handout 19



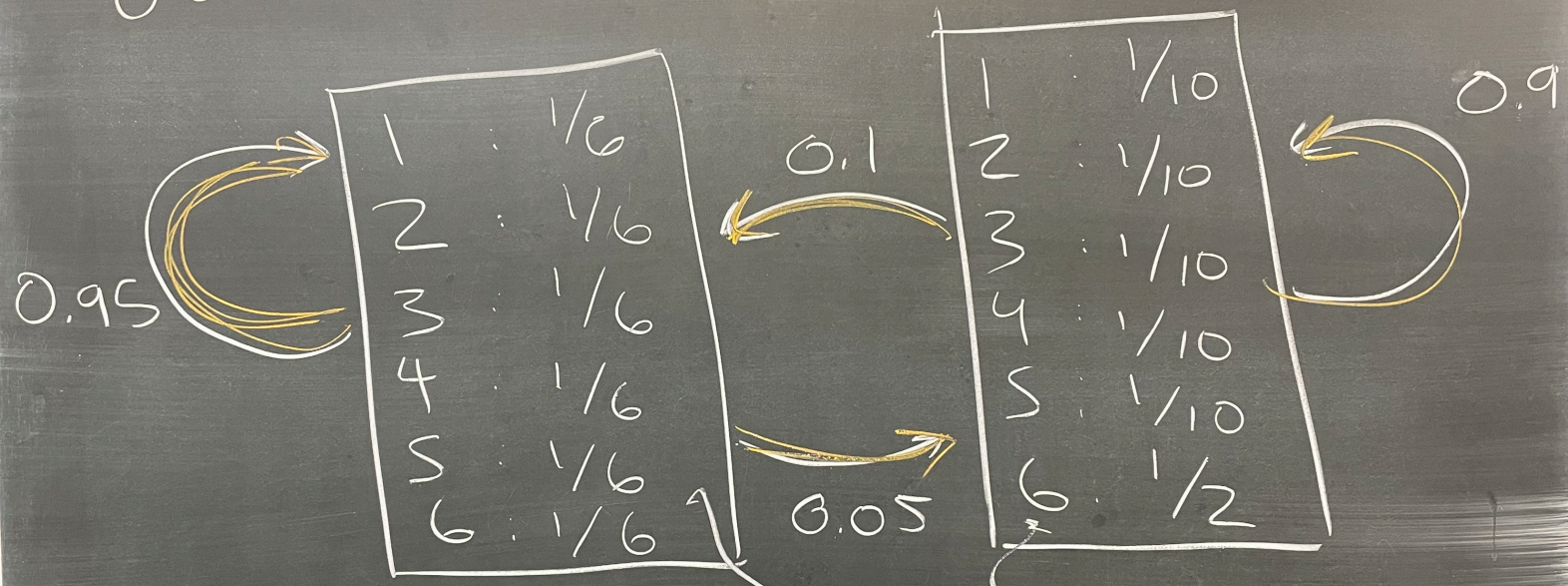
Q: What fraction of nights are spent in low sleep


Stationary distribution
π_j = fraction of time in state j
$= \sum_{i \in S} \pi_i P_{ij}$
$p(i)\,p(j|i)$
all possible states
note
$\pi_h = 1 - \pi_\ell$
$\pi_\ell = \sum_{s \in \{\ell, h\}} \pi_s P_{s\ell} = \pi_\ell P_{\ell\ell} + \pi_h P_{h\ell}$
$\pi_\ell = \pi_\ell (0.9) + (1 - \pi_\ell)(0.6)$
$= \frac{6}{7}$

Hidden Markov Models

HMMs hidden Markov models

"Occasionally dishonest casino"

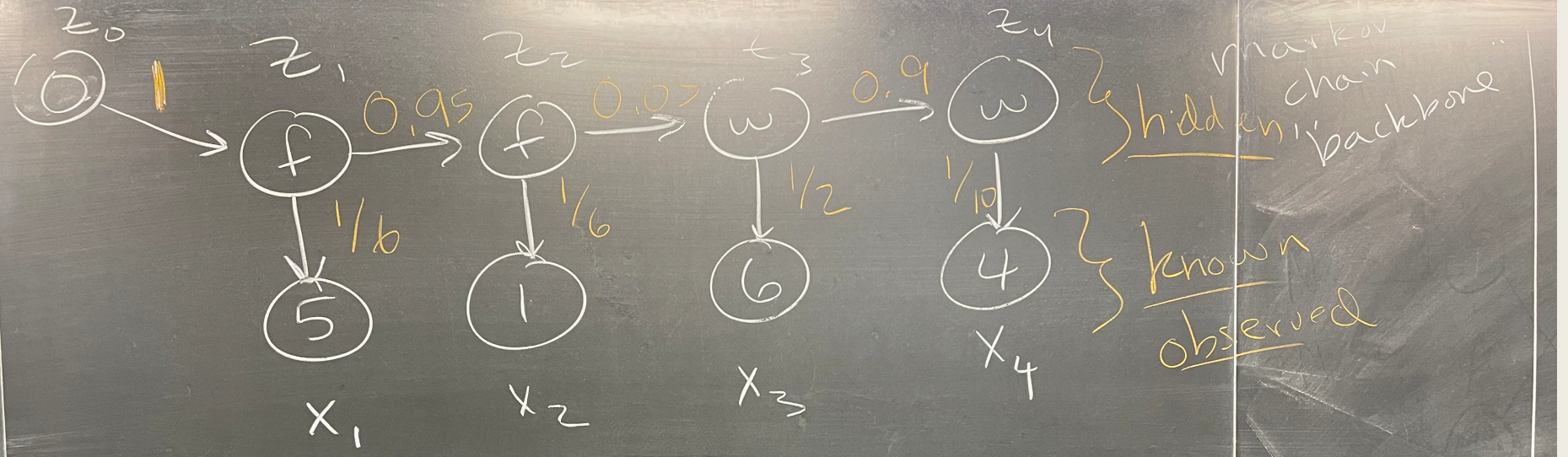


fair die
"f"

weighted die
"w"

transition probabilities

emission probabilities



HMM definition

K = # hidden states
i.e. $K=2$

B = # possible emissions
i.e. $B=6$

L = sequence length.
i.e. $L=4$

transition probabilities

$$a_{kl} = P(z_i = l \mid z_{i-1} = k)$$

$\forall\ k \neq l$, pairs

$$p(\vec{x}, \vec{z}) = \prod_{i=1} a_{z_{i-1} z_i} \cdot e_{z_i}(x_i)$$

joint

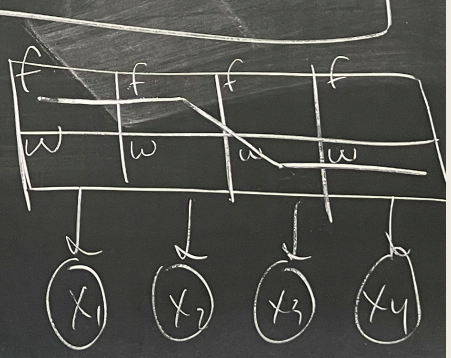
$$= \underbrace{a_{0f}\, e_f(5)}_{i=1} \cdot \underbrace{a_{ff}\, e_f(1)}_{i=2} \cdot \underbrace{a_{fw}\, e_w(6)}_{i=3} \cdot \underbrace{a_{ww}\, e_w(4)}_{i=4}$$

$$= (1 \cdot \tfrac{1}{6})(0.95 \cdot \tfrac{1}{6})(0.05 \cdot \tfrac{1}{2})(0.9 \cdot \tfrac{1}{10})$$

emission probabilities

$$e_k(b) = P(\underbrace{x_i = b}_{\text{observation}} \mid \underbrace{z_i = k}_{\text{hidden state}})$$

$\forall$ k & b pairs



Viterbi Algorithm

Q: How to we know state sequence?

A: Viterbi Algorithm

$V_k(i)$ = probability of the best path that ends in state k

$$= e_k(x_i) \cdot \max_\ell \{ V_\ell(i-1) a_{\ell k} \}$$