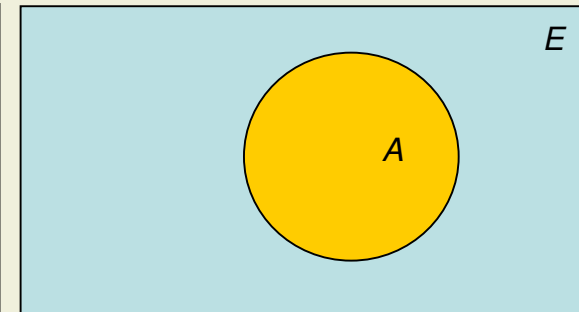


Probability Review

E : set of equally likely outcomes

A : an event

$$P(A) = \frac{n(A)}{n(E)} = \frac{\text{number of ways for event } A \text{ to happen}}{\text{number of equally likely outcomes}}$$



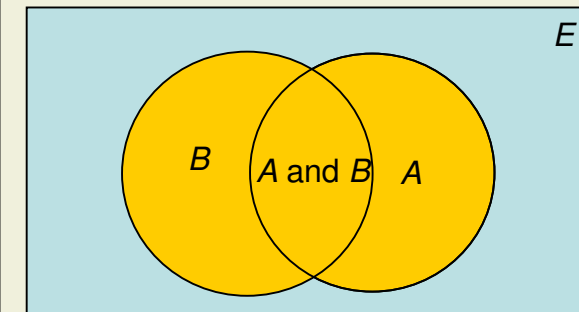
Combined Probability

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

Mutually Exclusive Events:

$$P(A \text{ and } B) = 0$$

$$\Rightarrow P(A \text{ or } B) = P(A) + P(B)$$



Conditional Probability (Probability of A given B)

$$P(A | B) = \frac{n(A \text{ and } B)}{n(B)} = \frac{P(A \text{ and } B)}{P(B)} \quad \Rightarrow \quad P(A \text{ and } B) = P(A | B)P(B)$$

Independent Events:

$$P(A | B) = P(A) \quad \Rightarrow \quad P(A \text{ and } B) = P(A)P(B)$$

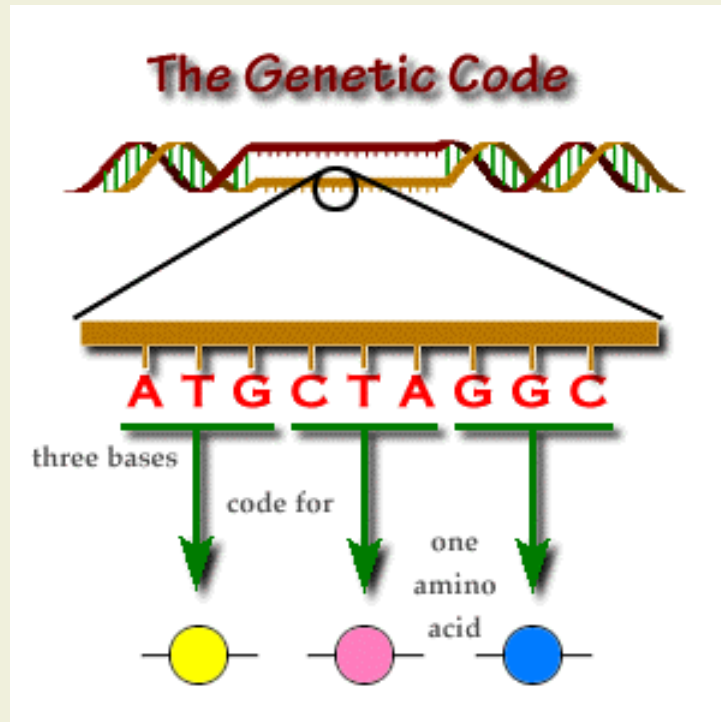
Probability Example

Pets in a Pet Store

	black	tawny	total
kitten	5	17	22
puppy	10	23	33
total	15	40	55

- $P(\text{black pet}) = 15/55 = 3/11$
- $P(\text{puppy}) = 33/55 = 3/5$
- $P(\text{black puppy}) = 10/55 = 2/11$
- $P(\text{puppy}|\text{black pet}) = 10/15 = 2/3$
- $P(\text{black pet} | \text{puppy}) = 10/33$
- $P(\text{puppy or black}) = (5+10+23)/55 = 38/55$

DNA Sequence



Bases (nucleotides)

A: adenine

G: guanine

C: cytosine

T: thymine

Purines

adenine and guanine

Pyrimidines

cytosine and thymine

Mutations to DNA Sequences

Base substitutions

S₀: GCC**C**ATCTG**A**A

S₁: GC**T**ATCTG**G**A

S₂: GC**T**ATCTG**A**A

(a) transition:

pur to pur or pyr to pyr

Eg A changed G

(b) transversion:

pur to pyr or pyr to pur

Eg A changed T

Deletions

S₀: GCC**C**ATCTGAA

S₁: GCATCTGAA

Insertions

S₀: GCCATCTGAA

S₁: GCC**G**ATCTGAA

Reversals

S₀: GCC**CATCTG**AA

S₁: GC**GTCTAC**AA

Probability of a base at a site of a sequence

S_0 : GCCATCTGAAGTACTTGGACCATGCTGTTCAGAGGGTCGTX

$n(A)=8$ $n(G)=12$ $n(C)=9$ $n(T)=11$ $n(E)=40$

Best estimate of the probability of each base at site X

- $P(A) = 8/40$
- $P(G) = 12/40$
- $P(C) = 9/40$
- $P(T) = 11/40$

S_1 : ACCACCTGAAGCACTAGGGCGATGCCGTTTAGAGAGTTGTX

$n(A)=10$ $n(G)=11$ $n(C)=9$ $n(T)=10$

Estimate of the probability of each base at site X (based on S_1)

- $P(A) = 10/40$
- $P(G) = 11/40$
- $P(C) = 9/40$
- $P(T) = 10/40$

Probability of a base at a site in aligned sequences

S_0 : **G**CCATCT**GAA**GTA**CTTGG**ACCAT**GCTG**TTCAG**AGGG**TC**G**TX
 S_1 : **A**CCACCTGAAGCACTAGGGCGATGCCGTTTAGAG**A**GTTGTX

Best estimate of the probability of a base in one sequence given a base in another

- $P(A_1/A_0) = 7/8$
- $P(A_1/G_0) = 2/12$

Table comparing bases at site X in aligned sequences S_0 and S_1

$S_1 \setminus S_0$	A	G	C	T	Total
A	7	2	0	1	10
G	1	10	1	0	12
C	0	0	6	3	9
T	0	0	2	7	9
Total	8	12	9	11	40

Conditional Probability

Table comparing bases at site X in aligned sequences S_0 and S_1

$S_1 \setminus S_0$	A	G	C	T	Total
A	7	2	0	1	10
G	1	10	1	0	12
C	0	0	6	3	9
T	0	0	2	7	9
Total	8	12	9	11	40

Estimate of the conditional probability of a base at a site in sequence S_1 given a particular base at that site in sequence S_0

$S_1 \setminus S_0$	A	G	C	T	
A	7/8	2/12	0	1/11	
G	1/8	10/12	1/9	0	
C	0	0	6/9	3/11	
T	0	0	2/9	7/11	
Total	1	1	1	1	

Conditional Probability

Table of conditional probability of a base at a site in sequence S_1 given a particular base at that site in sequence S_0

$S_1 \setminus S_0$	A	G	C	T	
A	$P(A_1 A_0)$	$P(A_1 G_0)$	$P(A_1 C_0)$	$P(A_1 T_0)$	
G	$P(G_1 A_0)$	$P(G_1 G_0)$	$P(G_1 C_0)$	$P(G_1 T_0)$	
C	$P(C_1 A_0)$	$P(C_1 G_0)$	$P(C_1 C_0)$	$P(C_1 T_0)$	
T	$P(T_1 A_0)$	$P(T_1 G_0)$	$P(T_1 C_0)$	$P(T_1 T_0)$	
Total	1	1	1	1	

Objective: Create a model for the conditional probabilities in the above table and use the table to predict the probabilities of a particular base at a site in the future. For example, $P(A_1)$

$$P(A_1) = P(A_1 \text{ and } A_0) \text{ or } P(A_1 \text{ and } G_0) \text{ or } P(A_1 \text{ and } C_0) \text{ or } P(A_1 \text{ and } T_0)$$

$$P(A_1) = P(A_1 | A_0)P(A_0) + P(A_1 | G_0)P(G_0) + P(A_1 | C_0)P(C_0) + P(A_1 | T_0)P(T_0)$$

Building a model to predict how sequences change

The probability of a base at a site in the future can be written as a matrix product. For example:

$$P(A_1) = P(A_1 | A_0)P(A_0) + P(A_1 | G_0)P(G_0) + P(A_1 | C_0)P(C_0) + P(A_1 | T_0)P(T_0)$$

$$P(A_1) = \begin{pmatrix} P(A_1 | A_0) & P(A_1 | G_0) & P(A_1 | C_0) & P(A_1 | T_0) \end{pmatrix} \begin{pmatrix} P(A_0) \\ P(G_0) \\ P(C_0) \\ P(T_0) \end{pmatrix}$$

$$\text{Let } \vec{P}_0 = \begin{pmatrix} P(A_0) \\ P(G_0) \\ P(C_0) \\ P(T_0) \end{pmatrix} \quad \text{and} \quad M = \begin{pmatrix} P(A_1 | A_0) & P(A_1 | G_0) & P(A_1 | C_0) & P(A_1 | T_0) \\ P(G_1 | A_0) & P(G_1 | G_0) & P(G_1 | C_0) & P(G_1 | T_0) \\ P(C_1 | A_0) & P(C_1 | G_0) & P(C_1 | C_0) & P(C_1 | T_0) \\ P(T_1 | A_0) & P(T_1 | G_0) & P(T_1 | C_0) & P(T_1 | T_0) \end{pmatrix}$$

$$\text{Then } \vec{P}_1 = M\vec{P}_0 \quad \text{and} \quad \vec{P}_t = M^t\vec{P}_0$$

The matrix M is called a transition matrix. Entries are probabilities. Columns sum to one. This is an example of a Markov Model.