Lecture No. 4
Basic Probability

January 14, 2003

- The basic problems are:
  what does a random sequence look like?
  How does a given DNA or Protein sequence compares to a random sequence?

- Random sequence- Each nucleotide or each amino acid will have an equal opportunity to appear at any position in the sequence.

- Unequal distribution for DNA sequence.
  ATGC is approximately distributed as:

$A = 30\%$	$T = 30\%$	$G = 20\%$	$C = 20\%$

  There is a linear correspondence between the nucleotides in a gene's DNA and the amino acids in the protein it encodes.

- The mRNA transcribed from DNA also have the same linear correspondence.

- The coding unit for each amino acid most likely consisted of three consecutive nucleotides in RNA, such coding units are termed as codons.

- Example- How many different codons are possible?
  Answer is: 4 items taken 3 at a time with repetition allowed i.e., $4^3 = 64$.

4 (ACGT)	4 (ACGT)	4 (ACGT)

- Permutation: n things taken k at a time with repetition is n^k. In last example $n = 4$ and $k = 3$.

- Example: - How many different PCR primers are possible of a sequence of 16 nucleotides (a nucleotide can be used more than once)?
  $n = 4$ and $k = 16$ so $n^k = 4^{16} = 3.2$ billion

- Primer: Short preexisting polynucleotide chain to which new deoxyribonucleotide can be added by DNA polymerase.

- DNA or RNA Polymerase: -Enzymes that catalyzes the synthesis of nucleic acid on preexisting nuclei acid templates assembling RNA from ribonucleotide or DNA from deoxyribonucleotide.

- PCR: Polymerase Chain Reaction - Amplifying a DNA base sequence using DNA polymerase.

1). First, the DNA is denatured by heating.

2). Single strand then mixed with excess of two kinds of short strands of DNA, and allowed to anneal.

3). Each single strand forms base pair with its complementary sequence in one of the two strands of the unwound DNA at the boarder of the segment.

4). These two short strands serve as initiators for copying the target sequence, and extends by the action of DNA polymerase.

- permutations without repetition: n things taken k at a time without repetition

$$P_k^n = \frac{n!}{(n-k)!}$$

- Example: - How many octapeptide can be there without repeating an amino acid. There are 20 amino acids taken 8 at a time:

$$\frac{20!}{(20-8)!} = 5 \text{ billion}$$

- Example: - n things taken all at a time

$$n! = n \times (n-1) \times (n-2), \ldots, \times 2 \times 1$$

- Example: - Given 5 different amino acids how many different peptide can be made? An amino acid can be used only once.
  Here $n = 5$ | 5 | 4 | 3 | 2 | 1 |
  $5! = 5 \times 4 \times 3 \times 2 \times 1 = 120$

- Example: - How many different circular peptide can be made using 5 amino acids? No starting point. Fewer different molecules.
  $n = 5$.
  $(n-1)! = 4! = 24$ instead of 120.

- Combination: - The number of ways a subset of k objects can be selected from a larger set of n objects.

$$C_k^n = \frac{P_k^n}{k!} = \frac{n!}{(n-k)!\,k!}$$

This is the selection without regard to order.

- Example: In how many ways

1). a nucleotide? $C_1^4 = \binom{4}{1} = 4$

2). 2 nucleotides ? $C_2^4 = \binom{4}{2} = 6$

3). 3 nucleotides ? $C_3^4 = \binom{4}{3} = 6$

4). 4 nucleotides ? $C_4^4 = \binom{4}{4} = 1$

- Probability Distributions: Probability theory is used as a model for situations for which the outcomes occur randomly. Such situation are called statistical experiments. E.g., tossing a die is a statistical experiment.

1). Sample Space: - The set of all possible outcomes.
Sample space of die tossing experiment is $S = \{1, 2, 3, 4, 5, 6\}$

2). sample point: - The elements in set S are called sample points.

3). Event: A subset of possible outcomes. A_1 is an event for the numbers < 3, i.e., $A_1 = \{1, 2\}$.

4). Null event ϕ: An event that will not occur.

- Probability: -Probability is a quantitative measurement of how likely an event will happen. The probability of an event A denoted by $p(A)$ is defined as:

$$\begin{aligned} p(A) &= \frac{n(A)}{n(S)} \\ &= \frac{\text{no. of sample points in } A}{\text{no. of sample points in } S} \end{aligned}$$

- Example: - Consider the die roll example and define
  Event A: get a 2 $A = \{2\}$
  Event B: get < 3 $B = \{1, 2\}$
  Event C: get > 2 $C = \{3, 4, 5, 6\}$
  Event D: get > 6 $D = \phi$
  Assume fair die: $p(1) = p(2) = p(3) = p(4) = p(5) = P(6) = 1/6$
  $p(A) = p(2) = 1/6$
  $p(B) = p(< 3) = p(1) + p(2) = 2/6$
  $p(C) = p(3) + p(4) + p(5) + p(6) = 4/6$
  $p(D) = p(\phi) = 0$

Probability Facts:

1). $p(S) = 1$

2). $0 \leq p(A) \leq 1$

3). If $p(A \text{and} B) = \phi$ and $p(A \text{or} B) = P(A) + p(B)$
So event A and B are mutually exclusive.

- Example: - How often an EcoRI site would be expected to appear by chance in a random sequence.
EcoRI is a restriction enzyme. Its recognition is GAATTC. Restriction enzymes cleave DNA molecules at specific recognition sites.
Event $A = \{GAATTC\}$ and $n(A) = 1$
No. of sample points in S: $n(S) = 4^6$.
Therefore

$$p(A) = \frac{1}{4^6}$$

- Some common restriction enzymes are

BamHI	GGATCC	CCTAGG
HindIII	AAGGTT	TTCCAA
TaqI	TCGA	AGCT
XhoI	CTCGAG	GAGCTC

- Independent events: If A and B are two events defined on the same sample space S and

$$p(A \text{and} B) = p(A)p(B)$$

then A and B are called independent events.

- One of the principal difficulties in analyzing DNA sequences is that the frequencies of neighboring bases are not independent.

  In particular, the frequencies of adjacent bases are generally different from the products of the frequencies of single bases.

- If p_u is the frequency of base type u in the sequence, and p_{uv} is the frequency with which successive bases are types u and v, then

$$p_{uv} \neq p_u p_v$$

Example Dinucleotide counts for the chicken $\beta-$globin sequence:

		Second Base				
		A	C	G	T	Total
First Base	A	23	26	23	15	87
	C	37	51	14	41	143
	G	25	38	36	19	118
	T	2	29	44	14	89
	Total	87	144	117	89	437

- Set the null hypothesis $H_0 : p_{ij} = p_i p_j$.
- Let n_{ij} be the count in $(ij)^{th}$ cell, observed count.
- Let $n_i = \sum_{j=1}^{4} n_{ij}$ and $n_j = \sum_{i=1}^{4} n_{ij}$.
- Let $n = \sum_{i=1}^{4} n_i = \sum_{j=1}^{4} n_j$.

- Under H_0 the expected counts will be:

$$e_{ij} = \frac{n_i \times n_j}{n}$$

- The form of χ^2 statistic is:

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

Pearson's Chi-squared test
data: x.mat X-squared = 59.4042, df = 9, p-value = 1.746e-09

- Conditional Probability: If A and B defined on S, the probability of B after A has occurred is a conditional probability:

$$p(B/A) = \frac{p(A \text{ and } B)}{p(A)}.$$

If A and B are independent events, then

$$p(A/B) = p(B)$$

Example: In dice example

$$S = \{1, 2, 3, 4, 5, 6\}$$

Let A be the event that the number is > 2 and B be the event that the number is 5:

$A = \{3, 4, 5, 6\},\ p(A) = 4/6$

$B = \{5\},\ p(A \text{and} B) = p(B) = 1/6$

$p(B/A) = \frac{p(A \text{and} B)}{p(A)} = \frac{1/6}{4/6} = 1/4$

- Bayesian Analysis: A method for including additional information from previous experience with similar data.
  Example: If you are in an alpha helix, what is the probability that the next amino acid is an alanine or a protine.

- Bayes Theorem: Sample space S is divided into a set of mutually exclusive and collectively exhaustive events $A_1, A_2, \ldots, A_N$.
  Event B is also defined in S:

$$\begin{aligned} p(A_k/B) &= \frac{p(A_k \text{and} B)}{p(B)} \\ &= \frac{p(B/A_k)p(A_k)}{p(B)} \\ &= \frac{p(B/A_k)(p(A_k)}{\sum_{i=1}^{N} p(B/A_i)p(A_i)} \end{aligned}$$

  Where
  $p(B) = \sum_{i=1}^{N} p(B \text{and} A_i) = \sum_{i=1}^{N} p(B/A_i)p(A_i)$

Example: Diagnostic Testing for Tuberculosis:

XRay	Tuberculosis No	Yes	Total
Negative	1,739	8	1,747
Positive	51	22	73
Total	1,790	30	1,820

- Let D_1 be the event that an individual suffering from tuberculosis.

- Let D_2 be the event that he or she is not.

- D_1 and D_2 are mutually exclusive and exhaustive.

- Let T^+ be the event that individual has positive Xray.

- Question:
  Find the Probability that an individual who tests positive for tuberculosis actually has the disease i.e., $P(D_1|T^+)$.

- This is the positive predictive value of the Xray. Using the Bayes' Theorem,

  $$p(D_1|T^+) = \frac{p(D_1)p(T^+|D_1)}{p(D_1)p(T^+|D_1) + p(D_2)p(T^+|D_2)}$$

  To solve for $p(D_1|T^+)$ we must first know $p(D_1)$, $P(D_2)$, $P(T^+|D_1)$, and $p(T^+|D_2)$.

- $p(D_1)$ is the probability that an individual in the general population has tuberculosis; since the 1820 individuals were not chosen from the population at random, the prevalence of disease cannot be obtained from above table.

However, in a past study there were 9.3 cases of tuberculosis per 100,000 population.

$$p(D_1) = 0.000093$$

and

$$p(D_2) = 0.999907.$$

$$\begin{aligned} p(T^+|D_1) &= 22/30 \\ &= 0.7333 \\ p(T^+|D_2) &= 1 - p(T^-|D_2) \\ &= 1 - \frac{1739}{1790} \\ &= 0.0285 \end{aligned}$$

Now put in the values in Bayes formula we get
$p(D_1|T^+) = 0.00239$
This means for every 100,000 positive Xrays,only 239 signal true cases of tuberculosis.
$p(D_1) = 9.3/100,000$ is called the prior probability.
$p(D_1|T^+) = 239/100,000$ is called posterior probability.