Why probability and statistics in Bioinformatics?

- Sequence alignments: detecting similarities between DNA or protein sequences.
- Phylogenetic tree reconstruction („tree of life“)
- Gene prediction (hidden Markov models)
- Analysis of microarray data (multiple testing, multivariate analysis)
- BLAST searches (random walks, extreme values)
- Analysis of computer simulations, networks, etc.
- much more!

=> whenever you want to answer questions from data you will bump into probability and statistics!
Example: Evolution of DNA

- Given two sequences, are they similar enough to conclude that they have a common ancestor?

- Two statistical problems:
  1. find best alignment of the sequences
  2. decide whether the similarity of this alignment is significant
Sequence alignment of DNA

- Two DNA sequences of length 26, aligned.
- Matches at 11 of 26 positions.
- Is this sufficient to conclude that the two sequences are evolutionarily related?

=> We can answer this question with the help of probabilities

Ingredients:
- Random variables
- Probability distributions
- Independence
- Statistical testing
- ...
Probability

Examples

Throw a fair die. What is the probability to throw 1, 2, ... 6?

\[ p(i) = \frac{1}{6}, \quad 1 \leq i \leq 6 \]

*Frequentist probability*

What is the probability that you will do sports tonight?

\[ P("yes") = 0.4 \]
\[ P("no") = 0.6 \]

*Bayesian probability*
Discrete random systems

The **sampling space** $S$ contains discrete points (finite or infinite)

- **Coin tossing**: $S=\{\text{head, tail}\}, \{0,1\}, \{1,-1\}$
- **Die tossing**: $S=\{1,2,3,4,5,6\}$
- **Nucleotide die tossing (tetrahedron)**: $S=\{a,c,g,t\}$
- **Amino acid die tossing**: $S=\{A, C, ..., W, Y\}$
Calculating probabilities

Let $S$ be the set of possible outcomes of a chance experiment and $A, B \subseteq S$ events. Then

- $P(S) = 1$
- $P(\emptyset) = 0$
- $0 \leq P(A) \leq 1$
- For the complement $A^c$ of $A$:
  
  $$P(A^c) = 1 - P(A)$$

- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

- For mutually exclusive events $A$ and $B$ ($A \cap B = \emptyset$):
  
  $$P(A \cup B) = P(A) + P(B)$$
Random variables (RVs)

**Def:** Random variable
Numerical quantity, whose value depends on the outcome of a chance experiment.

**Examples**

- Coin tossing. The random variable $X$ is assigned
  $X = 0$ for tails
  $X = 1$ for heads

- Trow two dice. Then
  $X_1 = \text{the value of die 1},$
  $X_2 = \text{the value of die 2},$
  $X = X_1 + X_2 = \text{the sum of both dice}$
  are random variables.

- Choose at random two DNA sequences from a database and align them. Then
  $X = \text{number of matches between the sequences}$
  is a random variable.
Probability distribution of a RV

- The probability distribution is *the* important feature of a random variable.
- The probability distribution of a random variable $X$ is the mechanism (the function) which tells us, with what probability the random variable takes what values.
Distributions of discrete RVs

The distribution of a RV assigns probabilities to all possible outcomes of the chance experiment.

**Probability function**

\[ p_X(i) := P(X = i) \quad \in [0, 1], \quad i \in S; \]

**(Cumulative) distribution function**

\[ F_X(j) := P(X \leq j) = \sum_{i \in S; i \leq j} P(X = i) \quad \in [0, 1], \quad j \in S \]
Distributions of continuous RVs

There are similar formulas for continuous random variables $X$.

(Cumulative) distribution function

$$F_X(t) := P(X \leq t) = \int_{x \leq t} f_X(x) \, dx \quad \text{for } t \in S$$

for some probability density function $f_X$ with $f_X(x) \geq 0$.

Caution with the interpretation in the continuous case:

- $f_X(x) \neq P(X = x)$ (in fact: $P(X = x) = 0$ for continuous RVs!)
- $f_X(x)$ is NOT a probability (that is, $f_X(x)$ might be $> 1 \ldots$)

Discrete random variables are perhaps more important to bioinformatics.