

Probabilistic models

Yuzhen Ye

School of Informatics, Computing and Computing

Indiana University, Bloomington

Spring 2018

Definitions

- Probabilistic models
 - A model means a system that simulates the object under consideration
 - A probabilistic model is one that produces different outcomes with different probabilities (BSA)

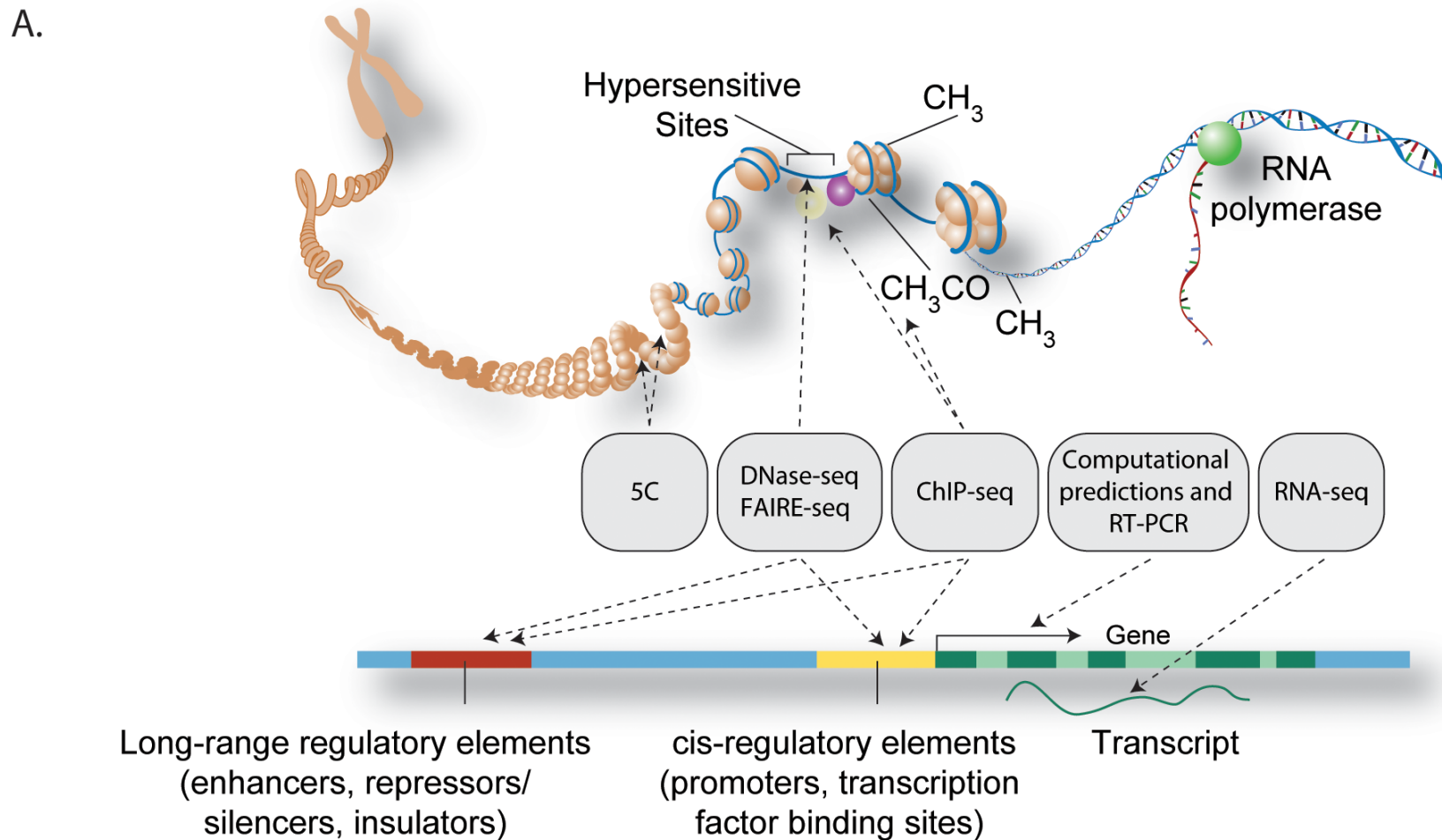


Why probabilistic models

- The biological system being analyzed is stochastic
- Or noisy
- Or completely deterministic, but because a number of *hidden* variables effecting its behavior are unknown, the observed data might be best explained with a probabilistic model



Figure 1. The Organization of the ENCODE Consortium.



The ENCODE Project Consortium (2011) A User's Guide to the Encyclopedia of DNA Elements (ENCODE). PLoS Biol 9(4): e1001046. doi:10.1371/journal.pbio.1001046

<http://www.plosbiology.org/article/info:doi/10.1371/journal.pbio.1001046>

Probability

- **Experiment:** a procedure involving chance that leads to different results
- **Outcome:** the result of a single trial of an experiment
- **Event:** one or more outcomes of an experiment
- **Probability:** the measure of how likely an event is
 - Between 0 (will not occur) and 1 (will occur)



Example: a fair 6-sided dice

- Outcome: The possible outcomes of this experiment are 1, 2, 3, 4, 5 and 6
- Events: 1; 6; even
- Probability: outcomes are ***equally likely*** to occur
 - $P(A) = \frac{\text{The Number Of Ways Event A Can Occur}}{\text{The Total Number Of Possible Outcomes}}$
 - $P(1)=P(6)=1/6$; $P(\text{even})=3/6=1/2$;



Random variable

- Random variables are functions that assign a unique number to each possible outcome of an experiment
 - An example
 - Experiment: tossing a coin
 - Outcome space: {heads, tails}
 - $$X = \begin{cases} 1 & \text{if heads} \\ 0 & \text{if tails} \end{cases}$$
 - More exactly, X is a discrete random variable
 - $P(X=1)=1/2$, $P(X=0)=1/2$
-

Probability distribution

- Probability distribution: the assignment of a probability $P(x)$ to each outcome x .
 - A fair dice: outcomes are ***equally likely*** to occur → the probability distribution over the all six outcomes $P(x)=1/6$, $x=1,2,3,4,5$ or 6 .
 - A loaded dice: outcomes are ***unequally likely*** to occur → the probability distribution over the all six outcomes $P(x)=f(x)$, $x=1,2,3,4,5$ or 6 , but $\sum f(x)=1$.
-

Probability mass function (pmf)

- A probability mass function is a function that gives the probability that a **discrete** random variable is exactly equal to some value; it is often the primary means of defining a discrete probability distribution
- An example

$$P(X) = \begin{cases} 1/2 & \text{heads} \\ 1/2 & \text{tails} \\ 0 & \text{others} \end{cases}$$



Probability density function (pdf)

- Probability density functions (pdf) are for ***continuous*** rather than ***discrete*** random variables; $f(x)$
- A pdf must be integrated over an interval to yield a probability, since $P(X = x) = 0$

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

- Cumulative distribution function (cdf)

$$P(X \leq x) = \int_{-\infty}^x f(t)d(t)$$

Joint probability

- Two experiments (random variables) X and Y
 - $P(X, Y) \rightarrow$ joint probability (distribution) of X and Y
 - $P(X, Y) = P(X|Y)P(Y) = P(Y|X)P(X)$
 - $P(X|Y) = P(X)$, X and Y are ***independent***
 - Example: experiment 1 (selecting a dice), experiment 2 (rolling the selected dice)
 - $P(y)$: $y = D_1$ or D_2
 - $P(i, D_1) = P(i|D_1)P(D_1)$
 - $P(i|D_1) = P(i|D_2)$, independent events
-

The probability of a DNA sequence

- Event: Observing a DNA sequence $S=s_1s_2\dots s_n$:
 $s_i \in \{A,C,G,T\}$;
 - Random sequence model (or *Independent and identically-distributed, i.i.d.* model): s_i occurs at random with the probability $P(s_i)$, independent of all other residues in the sequence;
 - $P(S) = \prod_{i=1}^n P(s_i)$
 - This model will be used as a *background* model (or called a *null hypothesis*).
-

Marginal probability

- The distribution of the ***marginal variables*** (the marginal distribution) is obtained by ***marginalizing*** over the distribution of the variables being discarded (so the discarded variables are marginalized out)
 - **Marginalizing** means considering all possible values the unknown variables may take, and averaging over them
 - $P(X) = \sum_Y P(X|Y)P(Y)$ $P(x) = \int P(x, y)dy$
 - Example: experiment 1 (selecting a dice), experiment 2 (rolling the selected dice)
 - $P(y)$: $y = D1$ or $D2$
 - $P(i) = P(i|D1)P(D1) + P(i|D2)P(D2)$
 - $P(i|D1) = P(i|D2)$, independent events
 - $P(i) = P(i|D1)(P(D1) + P(D2)) = P(i|D1)$
-

Conditional probability

- Conditioning the joint distribution on a particular observation
- Conditional probability $P(X|Y)$: the measure of how likely an event X happens under the condition Y ;

$$P(x|y) \equiv \frac{P(x, y)}{P(y)} = \frac{P(x, y)}{\int P(x, y)dy}$$

- Example: two dices D_1, D_2
 - $P(i|D_1) \rightarrow$ probability for picking i using dice D_1
 - $P(i|D_2) \rightarrow$ probability for picking i using dice D_2



Probability models

- A system that produces different outcomes with different probabilities.
- It can simulate a class of objects (events), assigning each an associated probability.
- Simple objects (processes) → probability distributions



Typical probability distributions

- Binomial distribution
- Gaussian distribution
- Multinomial distribution
- Poisson distribution
- Dirichlet distribution



Binomial distribution

- An experiment with binary outcomes: 0 or 1;
- Probability distribution of a single experiment:
 $P('1')=p$ and $P('0') = 1-p$;
- Probability distribution of N tries of the same experiment
- $\text{Bi}(k \text{ '1' s out of } N \text{ tries}) \sim \binom{N}{k} p^k (1-p)^{N-k}$



Gaussian distribution

- When $N \rightarrow \infty$, $Bi \rightarrow$ Gaussian distribution
- The Gaussian (normal) distribution is a continuous probability distribution with probability density function defined as:

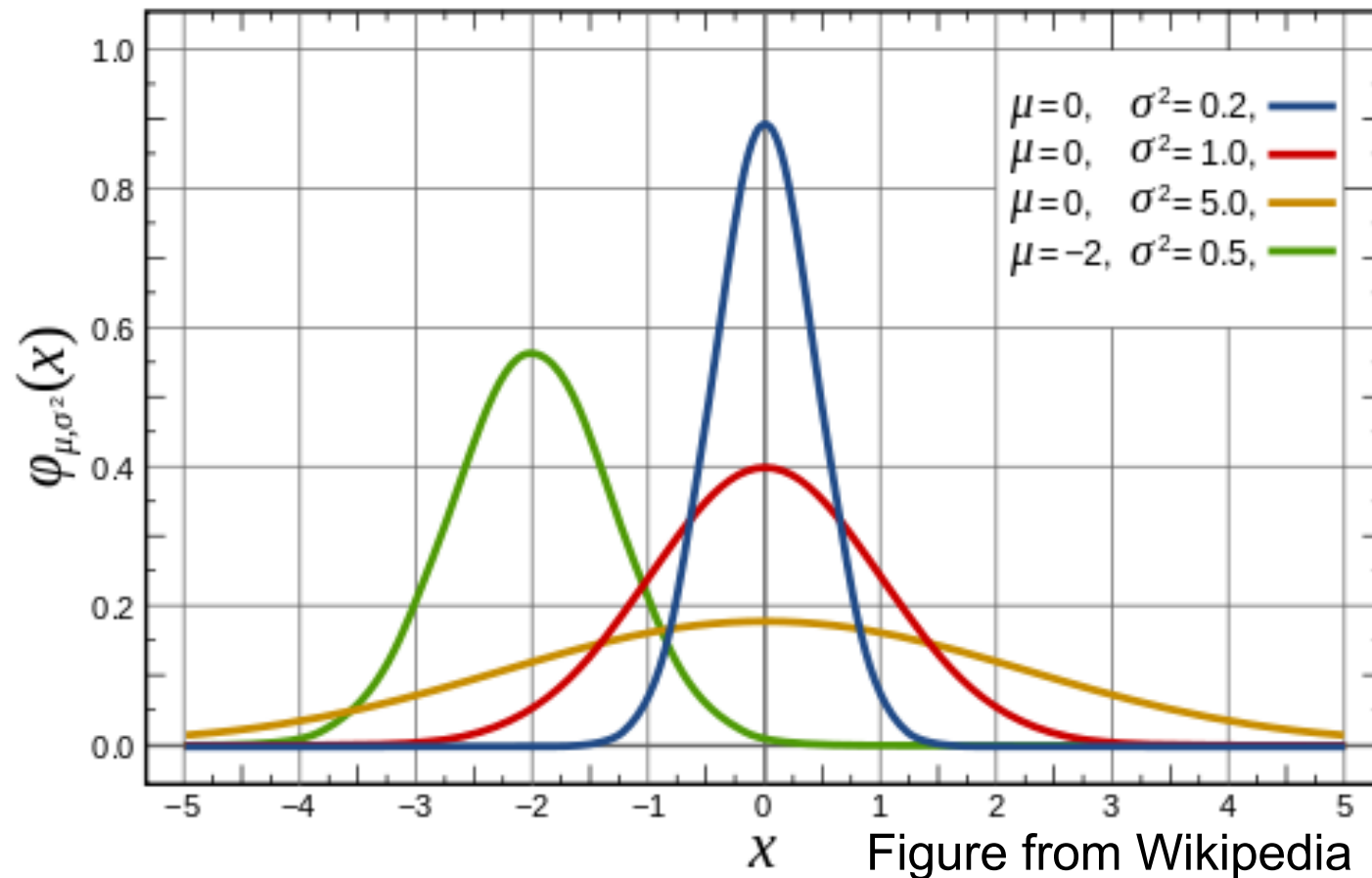
$$f(x; \mu, \sigma^2) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2} \left(\frac{x-\mu}{\sigma}\right)^2}$$

μ : mean (expectation); σ^2 : variance (σ : the standard derivation)

- If we define a new variable $u=(x-\mu)/\sigma$

$$f(x) \sim \frac{1}{\sqrt{2\pi}} e^{-u^2/2}$$

Gaussian distribution



standard normal distribution when $\mu = 0$ and $\sigma^2 = 1$

Multinomial distribution

- An experiment with K independent outcomes with probabilities θ_i , $i = 1, \dots, K$, $\sum \theta_i = 1$.
- Probability distribution of N tries of the same experiment, getting n_i occurrences of outcome i , $\sum n_i = N$ ($n = \{n_i\}$).

$$P(n|\theta) = M^{-1}(n) \prod_{i=1}^K \theta_i^{n_i}$$

$$M(n) = \frac{n_1! n_2! \cdots n_K!}{(\sum_k n_k)!} = \frac{\prod_i n_i!}{(\sum_k n_k)!}$$

Example: a fair dice

- Probability: outcomes (1,2,...,6) are ***equally likely*** to occur
- Probability of rolling 1 dozen times (12) and getting each outcome twice:
 - $\frac{12!}{2^6} \left(\frac{1}{6}\right)^{12} \sim 3.4 \times 10^{-3}$



Example: a loaded dice

- Probability: outcomes (1,2,...,6) are ***unequally likely*** to occur: $P(6)=0.5$, $P(1)=P(2)=\dots=P(5)=0.1$
- Probability of rolling 1 dozen times (12) and getting each outcome twice:
 - $\frac{12!}{2^6} (0.5)^2 \times (0.1)^{10} \sim 1.87 \times 10^{-4}$



Poisson distribution

- Poisson gives the probability of seeing n events over some interval, when there is a probability p of an individual event occurring in that period.



Poisson distribution for sequencing coverage modeling



Assuming uniform distribution of reads:

Length of genomic segment: L

Number of reads: n Coverage $\lambda = n l / L$

Length of each read: l

How much coverage is enough (or what is sufficient oversampling)?

Lander-Waterman model: $P(x) = (\lambda^x * e^{-\lambda}) / x!$

$$P(x=0) = e^{-\lambda}$$

where λ is coverage

Poisson distribution

c	$P_0=e^{-c}$	% not sequence	% sequenced (1- P_0)
1	0.37	37%	63%
2	0.135	13.5%	87.5%
3	0.05	5%	95%
4	0.018	1.8%	98.2%
5	0.0067	0.6%	99.4%
6	0.0025	0.25%	99.75%
7	0.0009	0.09%	99.91%
8	0.0003	0.03%	99.97
9	0.0001	0.01%	99.99%
10	0.000045	0.005%	99.995%



Dirichlet distribution

- Outcomes: $\theta = (\theta_1, \theta_2, \dots, \theta_K)$

- Density: $D(\theta|\alpha) = Z^{-1}(\alpha) \prod_{i=1}^K \theta_i^{\alpha_i-1} \delta\left(\sum_{i=1}^K \theta_i - 1\right)$

$$Z(\alpha) = \int \prod_{i=1}^K \theta_i^{\alpha_i-1} \delta\left(\sum_{i=1}^K \theta_i - 1\right) d\theta = \frac{\prod_i \Gamma(\alpha_i)}{\Gamma(\sum_i \alpha_i)}$$

- $(\alpha_1, \alpha_2, \dots, \alpha_K)$ are constants \rightarrow different α gives different probability distribution over θ .
 - $K=2 \rightarrow$ Beta distribution
-

Example: dice factories

- Dice factories produce all kinds of dices: $\theta(1)$, $\theta(2), \dots, \theta(6)$
 - A dice factory distinguish itself from the others by parameters $\alpha = (\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6)$
 - The probability of producing a dice θ in the factory α is determined by $\mathcal{D}(\theta|\alpha)$
-

Probabilistic model

- **Selecting a model**
 - A model can be anything from a simple distribution to a complex stochastic grammar with many implicit probability distributions
 - Probabilistic distributions (Gaussian, binominal, etc)
 - Probabilistic graphical models
 - **Markov models**
 - **Hidden Markov models (HMM)**
 - **Bayesian models**
 - Stochastic grammars
 - **Data → model (learning)**
 - The parameters of the model have to be *inferred* from the data
 - MLE (*maximum likelihood* estimation) & MAP (*maximum a posteriori* probability)
 - **Model → data (inference/sampling)**
-

MLE

- Estimating the model parameters (learning):
from *large* sets of *trusted* examples
- Given a set of data D (training set), find a model with parameters θ with the maximal likelihood $P(D|\theta)$

$$\hat{\theta}_{MLE} = \arg \max_{\theta} P(D|\theta)$$



Example: a loaded dice

- Loaded dice: to estimate parameters $\theta_1, \theta_2, \dots, \theta_6$, based on N observations $D=d_1, d_2, \dots, d_N$
- $\theta_i = n_i / N$, where n_i is the occurrence of i outcome (observed frequencies), is the maximum likelihood solution (BSA 11.5)

$$P(n|\theta_{MLE}) > P(n|\theta) \text{ for any } \theta \neq \theta_{MLE}$$

- Learning from counts
-

When to use MLE

- A drawback of MLE is that it can give poor estimations when the data are scarce
 - E.g, if you flip coin twice, you may only get heads, then $P(\text{tail}) = 0$
- It may be wiser to apply prior knowledge (e.g, we assume $P(\text{tail})$ is close to 0.5)
 - Use MAP instead



MAP

- Bayesian statistics

$$\begin{aligned} P(\theta|D) &= \frac{P(D|\theta)P(\theta)}{P(D)} \\ &= \frac{P(D|\theta)P(\theta)}{\sum_{\theta} P(D|\theta)P(\theta)} \end{aligned}$$

$P(\theta) \rightarrow$ prior probability

$P(\theta|D) \rightarrow$ posterior probability

$P(D|\theta) \rightarrow$ likelihood

- MAP
$$\begin{aligned} \hat{\theta}_{MAP} &= \arg \max_{\theta} P(\theta|D) \\ &= \arg \max_{\theta} \frac{P(D|\theta)P(\theta)}{P(D)} \\ &= \arg \max_{\theta} P(D|\theta)P(\theta) \end{aligned}$$

Example: two die

- Prior probabilities: fair dice 0.99; loaded dice: 0.01;
 - Loaded dice: $P(6)=0.5$, $P(1)=\dots P(5)=0.1$
 - Data: 3 consecutive '6' es:
 - $P(\text{loaded} | 3 \text{ '6' s}) = P(\text{loaded}) * [P(3 \text{ '6' s} | \text{loaded}) / P(3 \text{ '6' s})] = 0.01 * (0.5^3 / C)$
 - $P(\text{fair} | 3 \text{ '6' s}) = P(\text{fair}) * [P(3 \text{ '6' s} | \text{fair}) / P(3 \text{ '6' s})] = 0.99 * ((1/6)^3 / C)$
 - Model comparison by using likelihood ratio:
 $P(\text{loaded} | 3 \text{ '6' s}) / P(\text{fair} | 3 \text{ '6' s}) < 1$
 - So fair dice is more likely to generate the observation.
-

Learning from counts: including prior

- Use prior knowledge when the data is scarce
- Use Dirichlet distribution as prior for the multinomial distribution:

- Posterior
$$P(\theta|n) = \frac{P(n|\theta)P(\theta)}{P(n)} = \frac{P(n|\theta)D(\theta|\alpha)}{P(n)}$$

- Posterior mean estimator (PME)

$$\theta_i^{PME} = \int \theta_i D(\theta|n + \alpha) d\theta = Z^{-1}(n + \alpha) \int \theta_i \prod_k \theta_k^{n_k + \alpha_k - 1} d\theta$$

$$\theta_i^{PME} = \frac{n_i + \alpha_i}{N + A}$$

- Equivalent to add α_j as pseudo-counts to the observation n_j (BSA 11.5) (Add-one smoothing; *Laplace estimator*)
 - **We can forget about statistics and use pseudo-counts in the parameter estimation!**
-

Sampling

- Probabilistic model with parameter $\theta \rightarrow P(x|\theta)$ for event x ;
 - Sampling: generate a large set of events x_i with probability $P(x_i|\theta)$;
 - Random number generator (function `rand()`) picks a number randomly from the interval $[0, 1)$ with the uniform density;
 - Sampling from a probabilistic model \rightarrow transforming $P(x_i|\theta)$ to a uniform distribution
 - For a finite set X ($x_i \in X$), find i s.t. $P(x_1) + \dots + P(x_{i-1}) < \text{rand}(0, 1) < P(x_1) + \dots + P(x_{i-1}) + P(x_i)$
-

Entropy

- Probabilities distributions $P(x_i)$ over K events
 - $H(x) = -\sum P(x_i) \log P(x_i)$
 - Maximized for uniform distribution $P(x_i) = 1/K$
 - A measure of average uncertainty
 - A sample application of entropy in bioinformatics: as a measurement for conservation
-

Mutual information

- Measure of independence of two random variable X and Y
 - $P(X|Y)=P(X)$, X and Y are independent \rightarrow
 $P(X,Y)/P(X)P(Y)=1$
 - $M(X;Y)=\sum_{x,y} P(x,y)\log[P(x,y)/P(x)P(y)]$
 - 0 \rightarrow independent
 - A sample application of mutual information:
 - Correlation between two residues
 - Application in RNA structure prediction
-

BRCA1 and BRCA2

■ A little background

- BRCA1 and BRCA2 are human genes that produce tumor suppressor proteins.
- Specific inherited mutations in BRCA1 and BRCA2 increase the risk of female breast and ovarian cancers, and they have been associated with increased risks of several additional types of cancer.
- Together, BRCA1 and BRCA2 mutations account for about 20 to 25 percent of hereditary breast cancers and about 5 to 10 percent of all breast cancers.

■ A simple calculation

- A rare mutation in an important gene is observed in only 2% of the population. A person that carries this mutation in his/her genome has 90% chance of developing a disease. On the other hand, a person that has a normal gene (without mutation) only has a 5% chance of developing this disease.
 - Question: If you tested having this disease, what's your chance of carrying this rare mutation?
-