Biological Sequence Analysis
(Ch 11. Background on probability)

Biointelligence Laboratory
School of Computer Sci. & Eng.
Seoul National University
Seoul 151-742, Korea

This slide file is available online at
http://bi.snu.ac.kr/

List of Contents

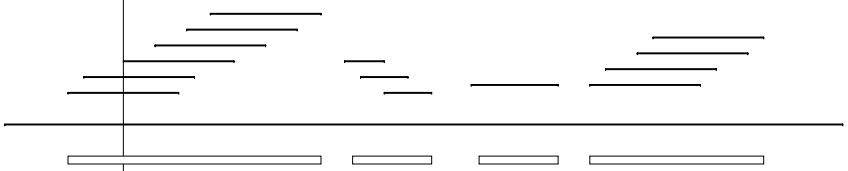
- Introduction
- Random variables
- Probability density, distributions
- Transformation
- Plots of statistical distributions
- Relative entropy and mutual information
- Many random variables
- One DNA sequence
- Maximum likelihood
- Sampling
- Metropolis sampling
- EM algorithm

Introduction

- Sequence similarity test:
  g g a g a c t g t a g a c a g c t a a t g c t a t a
  g a a c g c c c t a g c c a c g a g c c c t t a t c
  P (more than 10 matches) = ? (0.04)
- Parameter, data, hypothesis, random variable

One DNA Sequence

- Shotgun sequencing: Find long DNA sequence by many overlapping short sequences (500 bases).



One DNA Sequence (2)

- 1. What is the mean proportion of the genome covered by contigs?
- 2. What is the mean number of contigs?
- 3. What is the mean contig size?

(Discrete) Random Variables

- A discrete numerical quantity corresponding to an observed outcome of experiment
- (E.g.) experiment: rolling two six-sided dice
  - outcome: the two numbers on the dice
  - discrete random variables: the sum of the two numbers, the difference of the two numbers, the smaller number of the two,
- Random variables are usually represented by uppercase symbols (X,Y, …) and the realized values of the random variables are represented by lowercase symbols (x,y,…)

Probability Distributions

- For a finite set X : the probability distribution is simply an assignment of a probability p_x to each outcome x in X.

  (e.g.) the probability of outcomes of rolling a dice {1/12, 1/12, 1/12, 1/6, 1/4, 1/3}
- For a continuous set X : the probability density $f(x)$

$$p(x-\delta x/2 \le x \le x+\delta x/2) = f(x)\delta x$$

$$p(x_0 \le x \le x_1) = \int_{x_0}^{x_1} f(x)dx$$

$$f(x) \ge 0$$

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

Note: $f(x)$ can be greater than 0

Probability density, distribution (1/2)

- Relation of distribution and density

$$F(t) = P(X \le t) = \begin{cases} \sum_{x \le t} P(X = x) \\ \int_{x \le t} f(x)dx \end{cases}$$

- Expectation (mean)

$$E(X) = \begin{cases} \sum_{x} xP(X = x) \\ \int xf(x)dx \end{cases}$$

- For continuous X:

$$P(X = x) = 0, \qquad f(x) \text{ can be} > 1$$

Probability density, distribution (2/2)

- Relation of distribution and density

$$P(X = x) = \sum_{t \le x} P(X = t) - \sum_{t < x} P(X = t)$$

$$f(x) = \frac{d}{dt} F(t)\Big|_{t=x}$$

- Conditions for the probability density:

1. $f(x) \ge 0$
2. $\int f(x)dx = 1$

- Variance: $E(X-\mu)^2 = EX^2 - \mu^2$

Transformation

- One random variable:

$X_2 = g(X_1)$ *for monotone* g

$$F_2(t) = P(X_2 < t) = P(g(X_1) < t) = P(X_1 < g^{-1}(t)) = F_1(g^{-1}(t))$$

$$f_2(x) = f_1(g^{-1})\{g^{-1}\}'$$

- Many random variable: For 1-1 relations between ***X*** and ***U,***

$$\vec{X} = (X_1, \cdots, X_n)$$

$$\vec{U} = (U_1(X_1, \cdots, X_n), \cdots, U_n(X_1, \cdots, X_n))$$

$$f_U(u_1, \cdots, u_n) = f_X(x_1, \cdots, x_n)\left|J^{-1}\right| = f_X(x_1, \cdots, x_n)\left|J^*\right|$$

Statistical Independence

- Two events are independent if the outcome of one event does not affect the outcome of the other event
- Discrete random variables are independent if the value of one does not affect the probabilities associated with the values of another random variable
- (E.g.) Experiment: rolling a fair die

1. A: the number is even, B: the number is greater than or equal to 3.

$P(A \mid B) = P(A)$, $P(B|A) = P(B)$: A and B are independent

2. C: the number is greater than 3

$P(C) = 1/2$, $P(C \mid A) = 2/3$, $P(A \mid C) = 1/3$, $P(A) = 1/2$:

A and C are not independent

Uniform Distribution

- Discrete uniform : each outcome can occur equally likely.
- For N outcomes: $P(x) = 1/N$
- Continuous uniform: $f(x)=1/(b-a)$ *or* 1/area(A)

Bernoulli Distribution

- Bernoulli trial is a single trial with two possible outcomes (success / failure)
- Bernoulli random variable *Y*: number of successes in this trial

$$p(y \text{ succeses in a trial}) = p^y (1-p)^{1-y}, \quad y = 0,1$$

Binomial Distribution

- Defined on a finite set of all the possible results of N trials with a binary outcome ('0' or '1').
- The random variable is the number of success in the fixed *N* trials.

$$p(k \text{ succeses out of } N) = \binom{N}{k} p^k (1-p)^{N-k}$$

$$m = \sum k \; p(k) = \sum_{k=1}^{N} k \binom{N}{k} p^k (1-p)^{N-k} = Np$$

$$\sigma^2 = \sum (k-m)^2 \, p(k) = \sum_{k=1}^{N} (k-m)^2 \binom{N}{k} p^k (1-p)^{N-k} = Np(1-p)$$

Multinomial Distribution

- Extend the binomial to *K* independent outcomes with probabilities $\theta_i, \; (i = 1, \cdots, K)$

$$p(n_i, \; i = 1, \cdots, K \mid \theta) = \binom{n}{n_1 \cdots n_K} \prod_{i=1}^{K} \theta_i^{n_i}$$

(e.g.) Rolling a fair dice N times:

P (rolling 12 times and getting each number twice)

= 12! / 2!^{6} (1/6)^{12} = 3.4 x *10 $^{-3}$*

Geometric Distribution

- Random variable is the number of trials before the first failure (the length of a success run)
- Test of the significance of the long run in the sequences

$$P(y) = (1-p)\, p^y, \; (y = 0,\; 1,\; 2, \cdots)$$

$$F(y) = P(Y \le y) = 1 - p^{y+1}$$

- Geometric-like random variables in BLAST theory:

$$1 - F(y-1) = P(Y \ge y) \sim C p^y, \; (0 < C < 1)$$

Negative Binomial Distribution

- Random variable is the number of trials in the fixed number of successes
- *P* (*N*=*n*) = *P*(the first (*n*-1) *trials result in exactly* (*m*-1) *successes and* (*n*-*m*) *failures and trial n results in success)*

$$P(n) = \binom{n-1}{m-1} p^m (1-p)^{n-m}, \quad (n = m, m+1, m+2, \cdots)$$

Generalized Geometric Distribution

- Random variable is the number of trials before the (*k*+1)-*th* failure.

$$P(y) = \binom{y}{k} p^{y-k} (1-p)^{k+1}, \quad (y = k, k+1, k+2, \cdots)$$

Poisson Distribution

- A limiting form of the binomial distribution (as *n* becomes large, *p* becomes smaller) with moderate $np = \lambda$

$$P(y) = \frac{e^{-\lambda} \lambda^y}{y!}, \quad (y = 0,1,2, \cdots)$$

Exponential Distribution

- Exponential distribution is a continuous analogue of the geometric distribution

$$f(x) = \lambda e^{-\lambda x}, \quad x \geq 0$$

$$F(x) = 1 - e^{-\lambda x}, \quad x \geq 0$$

$$E(X) = 1/\lambda$$

$$Var(X) = 1/\lambda^2$$

$$Med(X) = \frac{\log 2}{\lambda} \approx 0.301 E(X) \ (\textit{right skewed})$$

Relation with Geometric Distribution

- Suppose X has an exponential distribution and Y is the integer part of X

$$P(Y = y) = P(y \le X < y+1) = (1-e^{-\lambda})e^{-\lambda y} = (1-p)\,p^{y}$$

$$(p = e^{-\lambda})$$

- The density function of the fractal part $D = X\text{-}Y$:

$$f(d) = \frac{\lambda e^{-\lambda d}}{1-e^{-\lambda}}$$

Gaussian Distribution

- As N gets larger, mean and variance of the binomial distribution increase linearly → rescale k by

$$u = \frac{k-m}{\sigma} = \frac{k-Np}{\sqrt{Np(1-p)}}$$

then the binomial becomes a Gaussian with the density as $f(u) = \frac{1}{\sqrt{2\pi}}\exp(-u^2/2)$

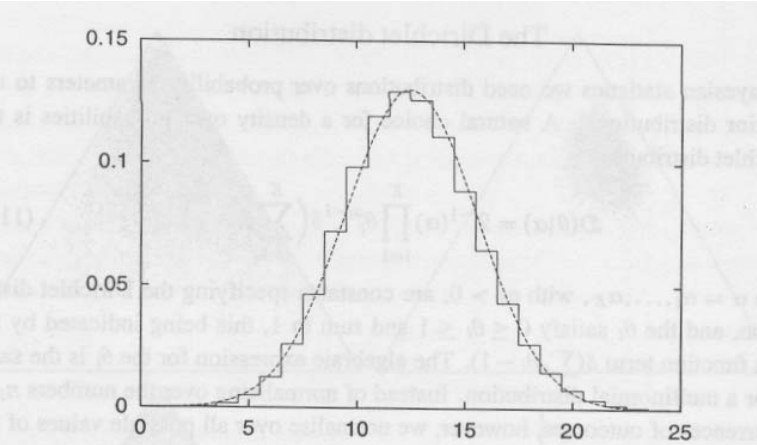


Figure 11.1 *The limit for large N of a binomial tends to a Gaussian. In this case $N = 40$ and $p = 1/4$ in (11.1).*

Dirichlet Distribution (1/3)

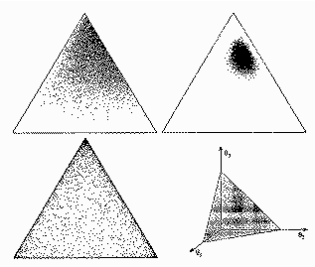
$$D(\theta_1, \cdots \theta_K \mid \alpha) = Z^{-1}(\alpha)\prod_{i=1}^{K}\theta_i^{\alpha_i - 1}\delta(\sum_{i=1}^{K}\theta_i - 1) \qquad (0 \le \theta_i \le 1)$$

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

- The Dirichlet distribution is a conjugate distribution for the multinomial distribution. (The Dirichlet prior wrt. multinomial distribution gives Dirichlet posterior again.)
- θ_i `s corresponds to the parameters of the multinomial distribution.

Dirichlet Distribution (2/3)

- The mean of Dirichlet is equal to the normalized parameters. $E[\theta_i] = \frac{\alpha_i}{\sum_k \alpha_k}$
- (e.g.) The three distributions all have the same mean (1/8, 2/8, 5/8) though with different values of α (1,2,5); (10,20,50); (.1, .2, .5)
- For *K=2*, Dirichlet reduces to the Beta distribution



Dirichlet Distribution (3/3)

- (e.g.) The dice factory: sampling $\theta_1, \cdots, \theta_6$ from Dirichlet parameterized by $\alpha_1, \cdots, \alpha_6$
- Factory A: all six parameters set to 10
  Factory B: all six parameters set to 2
- On average both factories produce fair dice (average = 1/6)
  But if we find a loaded dice with (.1 .1 .1 .1 .1 .6), it is More likely from factory B:
- The variance of Dirichlet is inversely proportional to the sum of parameters.

$$D(\theta \mid \alpha_A) = \frac{\Gamma(60)}{\Gamma(10)^6}(.1)^{5(10-1)}(.5)^{10-1} = 0.119$$

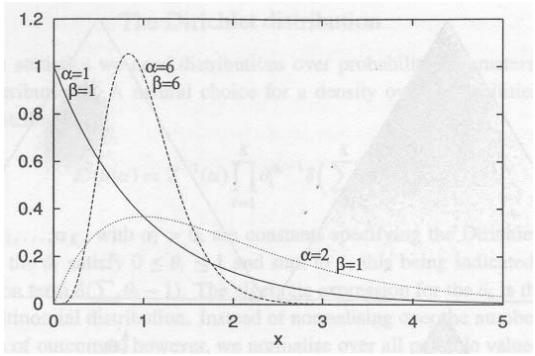
$$D(\theta \mid \alpha_B) = \frac{\Gamma(12)}{\Gamma(2)^6}(.1)^{5(2-1)}(.5)^{2-1} = 199.6$$

Gamma Distribution (1/2)

- $g(x, \alpha, \beta) = \frac{e^{-\beta x} x^{\alpha-1} \beta^{\alpha}}{\Gamma(\alpha)} \quad (0 < x, \alpha, \beta < \infty)$

$mean = \alpha / \beta$

$\text{var} = \alpha / \beta^2$



Gamma Distribution (2/2)

- The gamma distribution is conjugate to the Poisson: the probability of seeing n events over some interval when there is a probability *p* of an individual event occurring in that interval. $f(n) = \frac{e^{-p} p^n}{n!}$
- The gamma distribution is used to model the probabilities of rate.
- The gamma distribution is used to model the rate of evolution at different sites in DNA sequences.

Extreme Value Distribution (1/2)

- For N samples from $g(x)$, the probability that the largest of them are less than x is $G(x)^N$
  where $G(x) = \int_{-\infty}^{x} g(u)du$ $h(x) = Ng(x)G(x)^{N-1}$
- Extreme value density (EVD) for $g(x)$ is the limit of $h(x)$
- EVD is used to model the breaking point of a chain, to assessing the significance of the maximum score from a set of alignments

Extreme Value Distribution (2/2)

- EVD for exponential density $g(x) = \alpha\, e^{-\alpha x}$

$h(x) = \alpha e^{-\alpha x}(1-\alpha e^{-\alpha z}/N)^{N-1} \rightarrow \alpha e^{-\alpha z}\exp(-e^{-\alpha z}),\ (z = x-y)$

- (Gumbel distribution)
- (e.g.) For N=1,2,10,100, N >= 10 gives a good approx. to the EVD

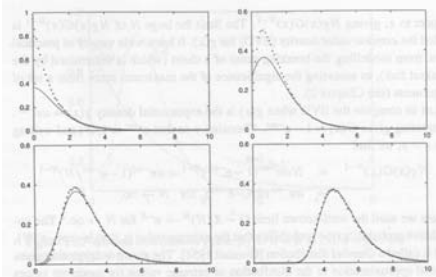
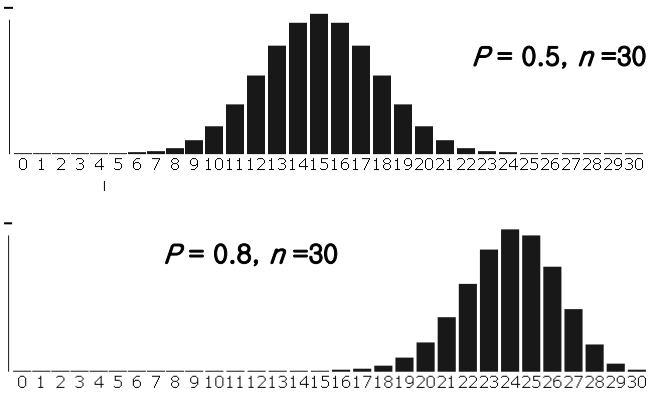


Figure 11.4 *Approximations to the extreme value distribution obtained by sampling N points from the distribution e^{-x} on $0 \le x < \infty$, and then taking the maximum. From the top left to bottom right, $N = 1, 2, 10, 100$.*

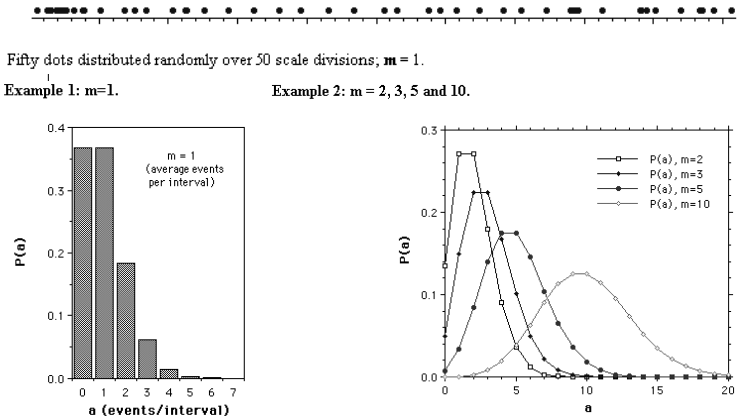
Statistical Distribution (1/7)

- Binomial Distribution



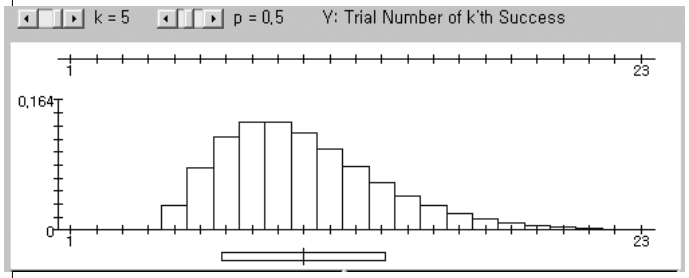
Statistical Distribution (2/7)

- Poisson Distribution



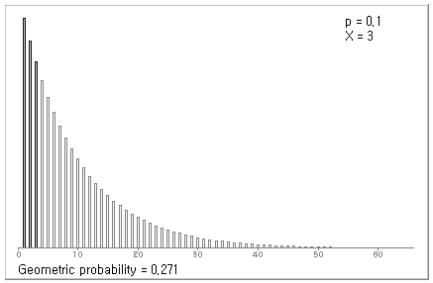
Statistical Distribution (3/7)

- Negative-Binomial Distribution



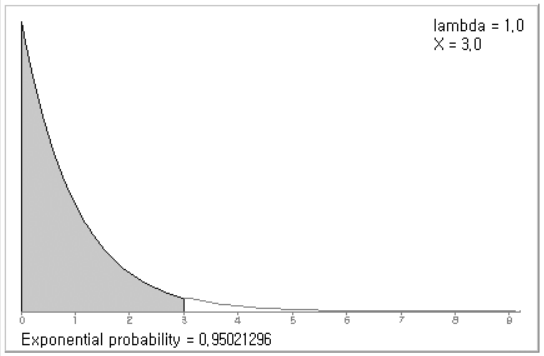
Statistical Distribution (4/7)

- Geometric Distribution



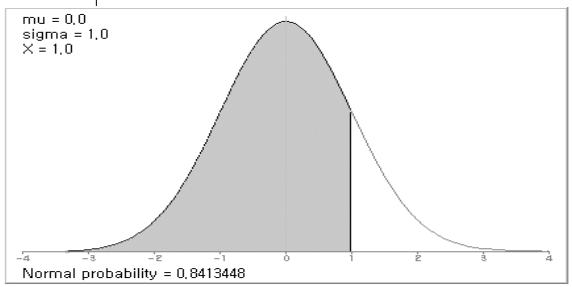
Statistical Distribution (5/7)

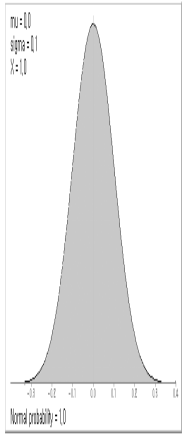
- Exponential Distribution



Statistical Distribution (6/7)

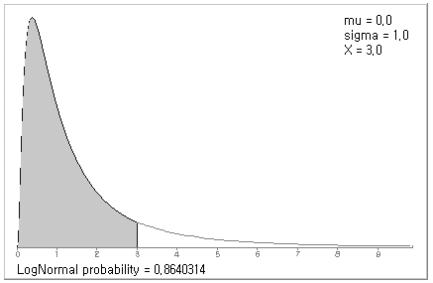
- Normal Distribution





Statistical Distribution (7/7)

- Lognormal Distribution



Entropy (1/2)

- A measure of average uncertainty of an outcome
- Shannon entropy: $H(X) = -\sum_i P(x_i) \log P(x_i)$
- Entropy is maximized when all the $P(x_i)$ are equal and the maximum is then $\log K$. If we are certain of the outcome, then the entropy is zero.
- *Information:* $I(X) = H_{before} - H_{after}$

Entropy (2/2)

- (E.g.) Entropy of equi-probable DNA symbol (*A*,*C*,*G*,*T*) is 2 bits.
- Information content of a conserved position: *A* particular position is always an *A* or a *G* with $p(A)=0.7$ and $p(G)=0.3$. Thus $H_{before} - H_{after} = 2 - .88 = 1.12$ *bits*

  (The more conserved the position, the higher the information content.)

Relative entropy and mutual information (1/4)

- Relative entropy (*KL* divergence) *P* wrt. *Q*:

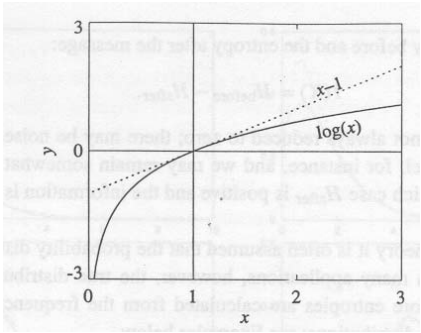
$$H(P \| Q) = \sum_i P(x_i) \log \frac{P(x_i)}{Q(x_i)}$$

- Information content and relative entropy are the same if Q represents the initial state
- $H(P \| Q) \neq H(Q \| P)$ (Not a metric)

Relative entropy and mutual information (2/4)

- Positive of relative entropy

$$\log(x) \le x - 1$$
$$-H(P \| Q) = \sum_i P(x_i) \log(Q(x_i)/P(x_i))$$
$$\le \sum_i P(x_i)(Q(x_i)/P(x_i) - 1) = 0$$



Relative entropy and mutual information (3/4)

- Independency can be measured by the relative entropy between $P(X,Y)$ and $P(X)P(Y)$.

$$M(X;Y) = \sum_{i,j} P(x_i, y_j) \log \frac{P(x_i, y_j)}{P(x_i)P(y_j)}$$

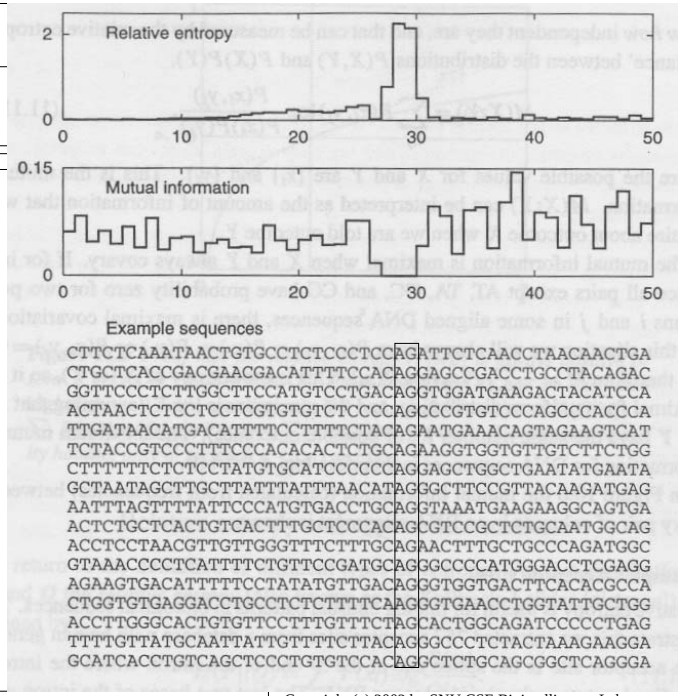
- $M(X, Y)$ can be interpreted as the amount of information that we acquire about outcome X when we are told outcome Y.

Relative entropy and mutual information (4/4)

- (E.g.) Acceptor sites: 757 acceptor sites from a database with human genes. 30 bases upstream, 20 bases down stream are extracted from each acceptor.
- Relative entropy: $\sum_a p_i(a) \log[p_i(a)/q_a]$

  where q_a *is the overall distribution of the four nucleotides in the sequences.*
- Mutual information:

$$\sum_{a,b} p_i(a,b) \log[p_i(a,b)/p_i(a)p_{i+1}(b)]$$



Many random variables (1/2)

- Marginal probability

$$P(Y_1 = y_1, \cdots, Y_i = y_i) = \sum_{y_{i+1}, \cdots, y_k} P(Y_1 = y_1, \cdots, Y_k = y_k)$$

$$f(x_1, \cdots, x_i) = \int \cdots \int f(x_1, \cdots, x_k) dx_{i+1} \cdots dx_k$$

- Conditional probability

$$P(Y_{i+1} = y_{i+1}, \cdots, Y_k = y_k \mid Y_1 = y_1, \cdots, Y_i = y_i)$$

$$= \frac{P(Y_1 = y_1, \cdots, Y_k = y_k)}{P(Y_1 = y_1, \cdots, Y_i = y_i)}$$

$$f(x_{i+1}, \cdots, x_k \mid x_1, \cdots, x_i) = \frac{f(x_1, \cdots, x_k)}{f(x_1, \cdots, x_i)}$$

Many random variables (2/2)

- Covariance

$$\sigma_{Y_1 Y_2} = \sum_{y_1, y_2} (y_1 - \mu_1)(y_2 - \mu_2) P(Y_1 = y_1, Y_2 = y_2)$$

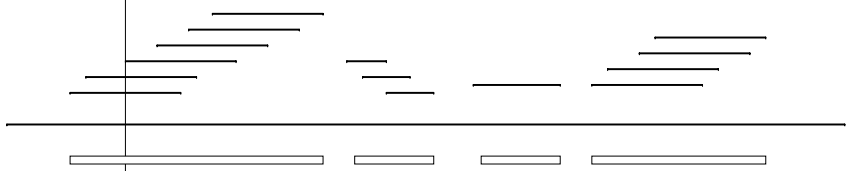
$$\sigma_{X_1 X_2} = \iint (x_1 - \mu_1)(x_2 - \mu_2) f(x_1, x_2) dx_1 dx_2$$

- Correlation:

$$\rho_{12} = \frac{\sigma_{Y_1 Y_2}}{\sigma_1 \sigma_2}$$

One DNA Sequence (1/10)

- Shotgun sequencing: Find long DNA sequence by many overlapping short sequences.



One DNA Sequence (2/10)

- 1. What is the mean proportion of the genome covered by contigs?
- 2. What is the mean number of contigs?
- 3. What is the mean contig size?

One DNA Sequence (3/10)

- There are *N* fragments, each of length *L*, the whole DNA is of length *G*.
- The position of the left hand end of any fragment is uniformly distributed in *(0, G)*
- The number of fragments falling in the interval length h becomes *Np=Nh/G* from binomial distribution.
- For large *N* and small interval *h*, this distribution becomes Poisson with mean *a = Nh/G*

One DNA Sequence (4/10)

- Mean proportion of genome covered by one or more fragment

  = *P* (a point chosen at random is covered at least by one fragment)

  = *P* (the left end of at least one fragment is in the interval L starting from that point), (by poisson)

  = $1-e^{-NL/G} = 1-e^{-a}, \quad (a = NL/G)$

One DNA Sequence (5/10)

- The length of fragments to cover the whole genome:

  P = .99 → *a* = *NL*/*G* = 4.6 (*a* : coverage)

  P = .999 → *a* = 6.9 (about 7 times the genome length)
- Human genome is 3x10^9 nucleotides : 3,000,000 of them are missing with *P*=.999
- What is the mean number of contigs ?

  success number: a unique right most fragment

  trial number: *N*

  p = *P* (a fragment is the right-most member of a contig)

  = *P* (no other fragment has its left-end point on the fragment in question)

  = e^{-a}
- Mean number of contigs $Ne^{-a} = Ne^{-NL/G}$

One DNA Sequence (6/10)

- For *G* = 100,000 *L* = 500 :

a	0.5	0.75	1	1.5	2	3	4	5	6	7
Mean number of contigs	60.7	70.8	73.6	66.9	54.1	29.9	14.7	6.7	3.0	1.3

- If there is a small number of fragments , there must be a small number of contigs
- A large number of fragments tend to form a small number of large contigs.

One DNA Sequence (7/10)

- Mean contig size ?
- Expectation of a sum of random variables where *N* is random
- Moment generating function of a random variable
- Probability generating function of a random variable
- Geometric distribution
- Exponential distribution

One DNA Sequence (8/10)

- Moment generating function (mgf) of a random variable *Y*:

$$m(\theta) = E(e^{\theta Y}) = \begin{cases} \sum_y e^{\theta y} P(y) \\ \int e^{\theta y} f(y) dy \end{cases}$$

- Finding mean and variance from mgf.:

$$\mu = \left(\frac{dm(\theta)}{d\theta} \right)_{\theta=0}, \quad \sigma^2 = \left(\frac{d^2 m(\theta)}{d\theta^2} \right)_{\theta=0} - \mu^2$$

or

$$\mu = \left(\frac{d \log m(\theta)}{d\theta} \right)_{\theta=0}, \quad \sigma^2 = \left(\frac{d^2 \log m(\theta)}{d\theta^2} \right)_{\theta=0}$$

One DNA Sequence (9/10)

- Probability generating function (pgf) of a (discrete) random variable *Y*:

$$p(t) = E(t^Y) = \sum_y P(y) t^y$$

$$\begin{cases} \mu = \left(\frac{d}{dt} p(t) \right)_{t=1} \\ \sigma^2 = \left(\frac{d^2}{dt^2} p(t) \right)_{t=1} + \mu - \mu^2 \end{cases}$$

One DNA Sequence (9/10)

- Expectation of a sum of random variables $S = X_1 + \cdots + X_N$ when *N* is random :

$$P(S = y) = \sum_n P_n P(S = y \mid n)$$

$$pgf\ of\ N : p(t) = \sum_n P_n t^n$$

$$pgf\ of\ S \mid n : (q(t))^n$$

$$P(S = y) = coeff.\ of\ t^y\ in\ \sum_n P_n (q(t))^n$$

$$= coeff.\ of\ t^y\ in\ p(q(t))$$

$$thus\ pgf\ of\ S : p(q(t)),$$

$$and\ by\ chain\ rule\ E(S) = E(N)E(X)$$

One DNA Sequence (10/10)

- Contig length = sum of *n* successful left end segments of fragments overlapped in the contig + L
- The distance between two successful left-hand points ~ *Exp* (λ) ,λ=*N*/*G* (known from Poisson process)
- The number of overlapping fragments ~ Geo (p)

$$p = \int_0^L \lambda e^{-\lambda x} dx = 1 - e^{-a}$$

- Mean of the random distance (any left-left end) ~ conditional exponential ($0 < X < L$)

$$f(x \mid 0 < x < L) = \frac{f(x)}{\int_0^L f(x)dx}$$

$$f(x) = \frac{\lambda e^{-\lambda x}}{1 - e^{-\lambda L}}$$

$$E(X \mid 0 < X < L) = \frac{1}{\lambda} - \frac{L}{e^{\lambda L} - 1}$$

One DNA Sequence (11/10)

- Mean Contig length:

$$E(N)E(X \mid 0 < X < L) + L$$

$$= (e^a - 1)\left(\frac{1}{\lambda} - \frac{L}{e^a - 1}\right) + L$$

- (e.g.) For L = 500, a = 2 ,G = 100,000 : the mean contig size = 1600

Maximum likelihood

$$\theta^{ML} = \arg\max_{\theta} P(D \mid \theta, M)$$

- Consistent : parameter value used to generate the data set will also be the value that maximizes the likelihood in the limit.
- (E.g.) For K observable outcomes of the model, the frequency of occurrence w will tend to $P(\omega \mid \theta_0, M)$ and log-likelihood for parameter θ: $\sum_i (n_i / \sum n_k) \log P(\omega_i \mid \theta, M)$ tends to $\sum_i P(\omega_i \mid \theta_0, M) \log P(\omega_i \mid \theta, M)$

  By the positivity of the relevant entropy implies that for all θ

  $\sum_i P(\omega_i \mid \theta_0, M) \log P(\omega_i \mid \theta_0, M) \geq \sum_i P(\omega_i \mid \theta_0, M) \log P(\omega_i \mid \theta, M)$

  Thus the likelihood is maximized by θ_0
- Drawbacks: can give poor estimate for scanty data

Posterior probability distribution

- Bayes theorem $P(\theta \mid D, M) = \frac{P(D \mid \theta, M) P(\theta \mid M)}{P(D \mid M)}$
- Use of posteriors: MAP (maximum a posteriori probability) estimate

$$\theta^{MAP} = \arg\max_{\theta} P(D \mid \theta, M) P(\theta \mid M)$$

- PME (posterior mean estimator) : parameters weighted by the posterior

$$\theta^{PME} = \int \theta \, P(\theta \mid n) d\theta$$

Change of variables

- Given a density of *x: f*(*x*) and a change of variable $x = \phi(y)$, the density of *y: g*(*y*) becomes

$$g(y) = f\{\phi(y)\}\ |\phi'(y)|$$

- The maximum of *g*(*y*), MAP or PME may shift from that of *f*(*x*).

Sampling

- Given a finite set with probabilities *P*(*x*), to sample from this set means to pick elements *x* randomly with probability *P*(*x*).
- Sampling by random number generator in [*0 1*]:

$$P(\textit{selecting } x_i) = P\{p(x_1) + \cdots + p(x_{i-1})$$
$$< \textit{rand } [0,\ 1] < p(x_1) + \cdots + p(x_{i-1}) + p(x_i)\}$$
$$= P(x_i)$$

Sampling by transformation from a uniform distribution

- For a uniform density *f*(*x*) and a map $x = \phi(y)$,

$$g(y) = f(\phi(y))\phi'(y) = \phi'(y),\ \ \phi(y) = \int_b^y g(u)du$$
$$y = \phi^{-1}(x)$$

- (E.g.) Sampling from a Gaussian:

$g(y)$, *x is given. Find* y:

Define $\phi(y) = \int_{-\infty}^{y} e^{-u^2/2} / \sqrt{2\pi}\ du$

And let $y = \phi^{-1}(x)$

Sampling with the Metropolis algorithm (1/3)

- For sampling when the analytic methods are not available
- Generate a sequence $\{y_i\}$ to approximate *P* as close as we like from $\tau(y \mid y_{i-1})$ under a condition:

(detailed balance) $P(x)\tau(y \mid x) = P(y)\tau(x \mid y)$

- Detailed balance implies:

$$\lim_{N\to\infty} \frac{1}{N} \#(y_i = x) = P(x)$$

- The sequence under this transition process will sample *P* correctly

Sampling with the Metropolis algorithm (2/3)

- Metropolis algorithm:
  - Symmetric proposal mechanism: Given a point *x*, this selects a point *y* with probability $F(y|x)$ and symmetric
  - Acceptance mechanism : accepts proposed *y* with probability *min* $(1, P(y)/P(x))$

    (A point *y* with larger posterior probability than the current x is always accepted, and one with lower probability is accepted randomly with probability $P(y)/P(x)$)

Sampling with the Metropolis algorithm (3/3)

- Metropolis algorithm (balance):

$$\begin{aligned} P(x)\tau(y \mid x) &= P(x)F(y \mid x)\min(1, P(y)/P(x)) \\ &= F(y \mid x)\min(P(x), P(y)) \\ &= F(x \mid y)\min(P(y), P(x)) \\ &= P(y)\tau(x \mid y) \end{aligned}$$

Gibbs sampling

- Sampling from conditional distributions

  $P(x_i \mid x_1, \cdots, x_{i-1}, x_{i+1}, \cdots, x_N)$ for each i.

- Proposal distribution is the conditional distribution:

  Always accept the sample.

Estimation of probabilities from counts (1/2)

- $\theta_i^{ML} = n_i / N$

$P(n \mid \theta^{ML}) > P(n \mid \theta)$ *for any* $\theta \neq \theta^{ML}$ *and an observation* n:

$$\begin{aligned} \log \frac{P(n \mid \theta^{ML})}{P(n \mid \theta)} &= \log \frac{\prod_i (\theta_i^{ML})^{n_i}}{\prod_i \theta_i^{n_i}} \\ &= \sum_i n_i \log \frac{\theta_i^{ML}}{\theta_i} \\ &= N \sum_i \theta_i^{ML} \log \frac{\theta_i^{ML}}{\theta_i} > 0 \end{aligned}$$

Estimation of probabilities from counts (2/2)

- For scarce data: use prior

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

$$D(\theta \mid \alpha) = \frac{1}{Z(\alpha)} \prod_{i=1}^{K} \theta_i^{n_i - 1}$$

$$P(\theta \mid n) = \frac{P(n \mid \theta) D(\theta \mid \alpha)}{P(n)}$$

$$= \frac{1}{P(n) M(n) Z(\alpha)} \prod_i \theta_i^{n_i + \alpha_i - 1} \qquad \theta_i^{PME} = \frac{n_i + \alpha_i}{N + A}$$

$$= \frac{Z(n+\alpha)}{P(n) M(n) Z(\alpha)} D(\theta \mid n+\alpha) \qquad \alpha_i : \text{ pseudo counts}$$

$$P(\theta \mid n) = D(\theta \mid n+\alpha)$$

Mixtures of Dirichlets

$$P(\theta \mid \alpha^1, \cdots, \alpha^m) = \sum_k q_k D(\theta \mid \alpha^k), \;\; q_k = P(\alpha^k)$$

$$P(\theta \mid n) = \sum_k P(\theta \mid \alpha^k, n) P(\alpha^k \mid n)$$

$$= \sum_k P(\alpha^k \mid n) D(\theta \mid n + \alpha^k)$$

$$\theta_i^{PME} = \sum_k P(\alpha^k \mid n) \frac{n_i + \alpha_i^k}{N + A}$$

EM algorithm (1/3)

- A general algorithm for ML estimation with missing data
- Baum-Welch algorithm for estimating hidden Markov model probabilities is a special case of the EM algorithm.

Maximize:

$$\log P(x \mid \theta) = \log \sum_y P(x, y \mid \theta)$$

x : *observation*

y : *miss ing data*

EM algorithm (2/3)

$$\log P(x \mid \theta) = \log P(x, y \mid \theta) - \log P(y \mid x, \theta)$$

$$= \sum_y P(y \mid x, \theta^t) \log P(x, y \mid \theta) - \sum_y P(y \mid x, \theta^t) \log P(y \mid x, \theta)$$

$$Q(\theta \mid \theta^t) = \sum_y P(y \mid x, \theta^t) \log P(x, y \mid \theta)$$

$$\log P(x \mid \theta) - \log P(x \mid \theta^t) =$$

$$Q(\theta \mid \theta^t) - Q(\theta^t \mid \theta^t) + \sum_y P(y \mid x, \theta^t) \log \frac{P(y \mid x, \theta^t)}{P(y \mid x, \theta)}$$

Thus,

$$\log P(x \mid \theta) - \log P(x \mid \theta^t) \geq Q(\theta \mid \theta^t) - Q(\theta^t \mid \theta^t)$$

EM algorithm (3/3)

- Choosing $\theta^{t+1} = \arg\max_{\theta} Q(\theta \mid \theta^t)$ will always makes the difference positive and thus the likelihood of the new model is larger than the likelihood of the old one.
- EM Algorithm:

(E-step) Calculate Q function

(M-step) Maximize $Q(\theta \mid \theta^t)$ wrt. θ

- The likelihood increases in each iteration
- Instead of maximizing in the (M-step), algorithms that increase Q are called generalized EM (GEM).