上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

A scientist who has learned how to use probability theory directly as extended logic has a great advantage in power and versatility over one who has learned only a collection of unrelated ad hoc devices.

– E. T. Jaynes, 1996

Chapter 1: Probability, Statistics and InformationTheory

for Biological Sequence Analysis

Chaochun Wei

Spring 2019

Contents

- Reading materials
- Applications
- Introduction
  - Definition
  - Conditional, joint, marginal probabilities
  - Statistical inference
    - Bayesian statistical inference
    - Frequentist inference
  - Information theory
  - Parameter estimation

Reading

吴军

数学之美，人民邮电出版社，2014

Jaynes, E.T.,

Probability Theory: The logic of Science, Cambridge University Press, 2003

Probability theory

for biological sequence analysis

- Applications
  - BLAST significance tests
  - The derivation of BLOSUM and PAM scoring matrices
  - Position Weight Matrix (PWM or PSSM)
  - Hidden Markov Models (HMM)
  - Maximum likelihood methods for phylogenetic trees

- Definition
  - $P_i \geq 0; \sum_i P_i = 1$
  - $f(x) \geq 0; \int_{-\infty}^{+\infty} f(x)dx = 1$
- Examples:
  - A fair dice: $P_i = 1/6, i = 1,2,\ldots,6.$
  - A random nucleotide sequence: $P_A = P_C = P_G = P_T = 1/4$
- "i.i.d.": independent, identically distributed

- Conditional, joint and marginal probabilities
  - Joint probability: P(A,B): “probability of A and B”
  - Conditional probability: P(A|B) : “probability of A given B”
    - P(A|B) = P(A, B)/P(B)
  - Marginal probability: $P(A) = \sum_B [P(A \mid B) * P(B)] = \sum_B P(A, B)$
- Examples:
  - The occasionally dishonest casino. Two types of dice:

    99% are fair, 1% are loaded such that $P_6 = 0.5$

    Conditional P(6|loaded), joint P(6, loaded); marginal P(6)

- Statistical inference
  - Bayesian statistical inference
  - Maximum likelihood inference
  - Frequentist inference

- Bayesian statistical inference

  The probability of a hypothesis, H, given some data, D.

  - Bayes' rule: P(H|D) = P(H)*P(D|H)/P(D)

    H: hypothesis, D: data

    - P(H): prior probability
    - P(D|H) : likelihood
    - P(H|D): posterior probability
    - P(D): marginal probability: $P(D) = \sum_{H} P(D \mid H) P(H)$

- Bayesian statistical inference
  - Examples

1. The occasionally dishonest casino. We choose a die, roll it three times, and every roll comes up a 6. Did we pick a loaded die?

Ans: Let H stand for "picked a loaded die", then

P(H|6, 6, 6) = P(6, 6, 6|H) P(H)/P(6, 6,6) ~=0.21

- Maximum likelihood inference
  - For a model M, find the best parameter $\Theta=\{\Theta_i\}$ from a set of data D, i.e.,

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

  - Assume dataset D is created by model M with parameter Θ_0 : K observable outcome ω_i, i=1, …, K, with frequencies n_i, i=1, …, K. Then, the best estimation of $P(\omega_i\,|\Theta_0, M)$ is $n_i/\Sigma n_{k.}$

Probability theory

- Maximum likelihood inference
    - P(x|y): probability(of x) or likelihood(of y)
    - Likelihood ratios; log likelihood ratios (LLR)

      $P(D|\Theta_1, M)/P(D/\Theta_2, M)$; $\log(P(D|\Theta 1, M)/P(D/\Theta 2, M))$
    - Substitution matrices are LLRs
        - Derivation of BLOSUM matrices (Henikoff 1992 paper)
        - Interpretation of arbitrary score matrices as probabilistic models (Altschul 1991 paper)

- Maximum likelihood inference
  - Derivation of BLOSUM matrices (Henikoff 1992 paper)
    - aa pair frequency table f: {f_{ij} }
    - Compute a LLR matrix

$$q_{ij} = f_{ij} / (\sum_{i}^{20} \sum_{j}^{20} f_{ij})$$

$$p_i = q_{ii} + \sum_{i \neq j} q_{ij} / 2$$

Expected probability of each i,j pair:

$$e_{ij} = \begin{cases} p_i^2, i = j \\ 2 p_i p_j, i \neq j \end{cases}$$

substitution matrix: $s_{ij} = \log_2 (q_{ij} / e_{ij})$

- Frequentist inference
  - Statistical hypothesis testing and confidence intervals
  - Examples:
    - Blast p-values and E-values
    - P(S >= x)
    - Expectation value, E=NP(S>=x)

- Information theory
  - How to measure the degree of conservation?
  - Shannon entropy
  - Relative entropy
  - Mutual information

- Shannon entropy: A measure of uncertainty
  - Probability P(x_i) for discrete set of K events x_1, …, x_k, the Shannon entropy H(X) is defined as

$$H(X) = -\sum_i P(x_i)\log P(x_i)$$

  - Unit of Entropy: ‘bit’ (use logarithm base 2)
  - H(X) is maximized when P(x_i)=1/K for all i.
  - H(X) is minimized when P(x_k)=1, and P(x_i)=0 for all i≠K.

- Information: a measure of reduction of uncertainty
  - the difference between the entropy before and after a ‘message’ is received

$$I(X) = H_{before} - H_{after}$$

- Shannon entropy: A measure of uncertainty
  - Example: in a DNA sequence aϵ{A, C, G, T}, P_a=1/4; then

$$H(X) = -\sum_a P_a \log P_a = 2bits$$

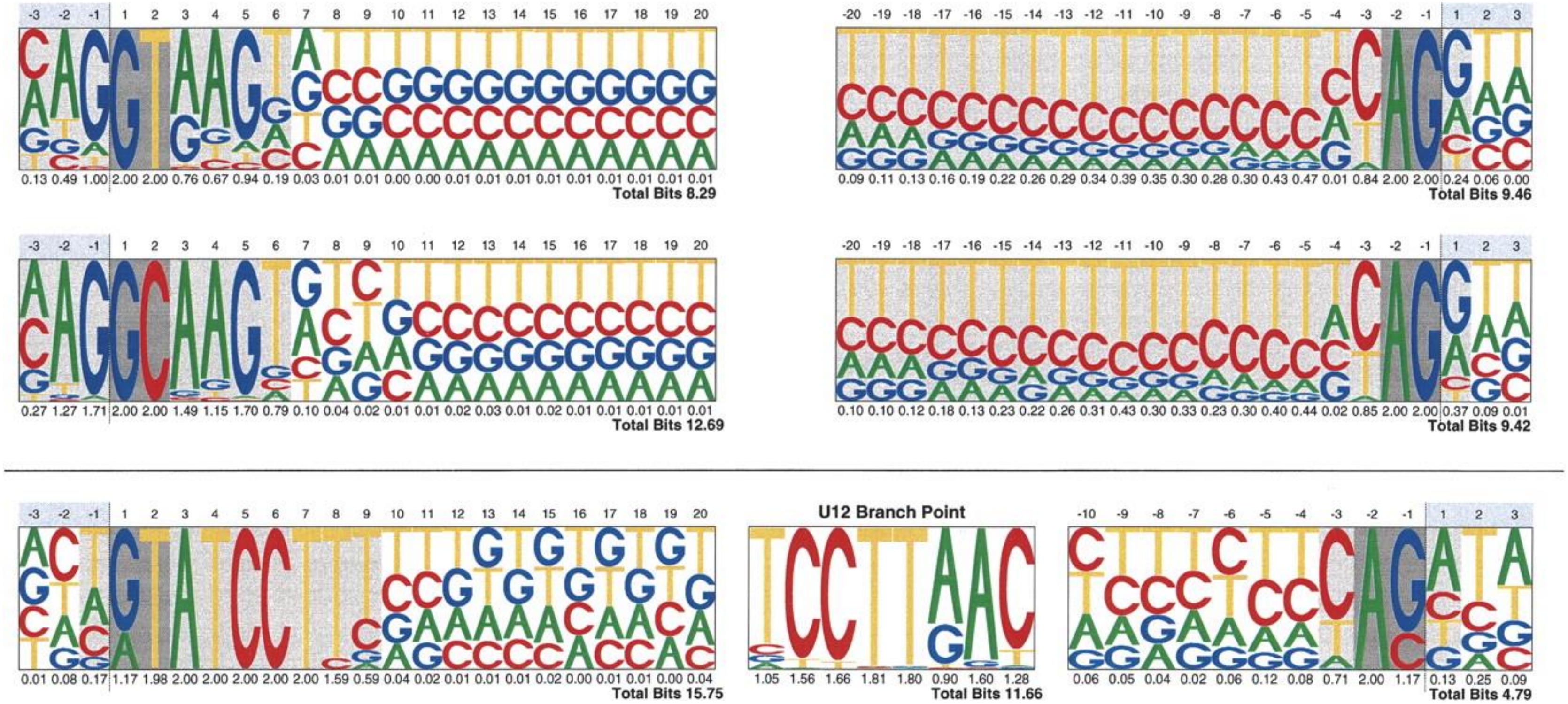
- Information: A measure of reduction in uncertainty
  - Example: measure the degree of conservation of a position in a DNA sequence

In a position of many DNA sequences, if P_C=0.5 and P_G=0.5, then
H_{after}= - 0.5log$_2$0.5 - 0.5log$_2$0.5 = 1 bits.
The information content of this position is
2-1=1 bits

Patterns in Splice Sites

Donor Sites

Acceptor Sites



Josep F. Abril et al. Genome Res. 2005; 15: 111-119

Sequence data from RefSeq of human, mouse, rat and chicken.

- Relative entropy: a measure of uncertainty
  - a different type of entropy

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

  - Property of a relative entropy
    - H(P||Q) ≠H(Q||P)
    - H(P||Q) ≥ 0
    - Can be viewed as the expected LLR.

- Proof of Relative entropy is always nonnegative

$$\because \log(x) \le x-1$$

$$\therefore -H(P \| Q) = \sum_i P(x_i)\log\frac{Q(x_i)}{P(x_i)} \le \sum_i P(x_i)(\frac{Q(x_i)}{P(x_i)}-1)$$

$$= \sum_i (Q(x_i) - P(x_i)) = 0$$

$$\therefore H(P \| Q) \ge 0$$

- Mutual information M(XY)

$$M(XY) = \sum_{xy} P(x,y) \log \frac{P(x,y)}{P(x)P(y)}$$

- Parameter estimation
  - Maximum likelihood estimation (ML)
  - Maximum a posterior estimation(MAP)
  - Expectation maximization (EM)

- Parameter estimation
  - Maximum likelihood estimation: use the observed frequencies as probability parameters, i.e.,

$$P(x) = \frac{count(x)}{\sum_y count(y)}$$

  - Maximum a posterior estimation(MAP)
    - “Plus-one” prior,
    - Pseudocounts

- Parameter estimation
  - EM: A general algorithm for ML estimation with “missing data”.
    - Iteration of two steps:
      - E-step: using current parameters, estimate expected counts
      - M-step: using current expected counts, re-estimate new parameters
  - Example: Baum-Welch algorithm for HMM parameter estimation.
  - Convergence guaranteed