

Project & Homework

- Homework 2 due Nov 9
 - Team assignment
- Project meeting schedule
 - R1000-1030: Sarah + Daniel
 - R1130-1200: Sam + Jacob
 - R1200-1230: Kasturi + Andy
 - R1300-1330: Minh + Kyle
 - R1400-1430: James + Nazli
 - R1430-1500: Erdin + Cetin
 - F 1300-1330: Tim + Razvan
- Project description (due Friday)
 - 1/2 to 1 page of motivation.
 - 1/2 to 1 page of background/related work.
 - 1/2 to 1 page describing your progress.

Alignment statistics

Bioinformatics Fall 2011

1

Mathematical Basis

- Model matches as a sequence of coin tosses
- Let p be the probability of a “head”
 - For a “fair” coin, $p = 0.5$
- (Erdős-Rényi) If there are n throws, then the expected length R of the longest run of heads is

$$R = \log_{1/p}(n).$$

Alignment statistics

Bioinformatics Fall 2011

2

Example

- Example: Suppose $n = 20$ for a “fair” coin
 $R = \log_2(20) = 4.32$
- Trick is how to model DNA (or amino acid) sequence alignments as coin tosses.

Alignment statistics

Bioinformatics Fall 2011

3

Modeling Sequence Alignments

- To model random sequence alignments, replace a match with a “head” and mismatch with a “tail”.

AATCAT → HTHHHT
ATTCAG

- For DNA, the probability of a “head” is $1/4$
 - What is it for amino acid sequences?

Alignment statistics

Bioinformatics Fall 2011

4

Modeling Sequence Alignments

- So, for one particular alignment, the Erdős-Rényi property can be applied
- What about for all possible alignments and considering the best one?
 - Consider that sequences are being shifted back and forth
 - Length of effective string is approx. n^2
- The expected length of the longest match can be approximated as

$$R = 2 \log_{1/p}(n)$$

- Also holds with up to $\log(n) / \log \log(n)$ mismatches.

Alignment statistics

Bioinformatics Fall 2011

5

Extreme statistics of local alignment

- Need to compute scores instead of length of runs
 - An experiment obtains the maximum value of locally aligning a random string with the database string. Repeat with another random string and so on. Plot the distribution of these maximum values.
 - The resulting distribution is an extreme value distribution, called a *Gumbel distribution*.
 - General form: $\text{Prob}(X=x) = (1/\beta)e^{-(x-\mu)/\beta} - e^{-(x-\mu)/\beta}$, where μ is called the *location parameter* and β is called the *scale parameter*.
- $$\text{Prob}(X=x) = e^{-x} - e^{-x^2}, \text{ when } \mu = 0 \text{ and } \beta = 1, \text{ is called the standard Gumbel distribution}$$

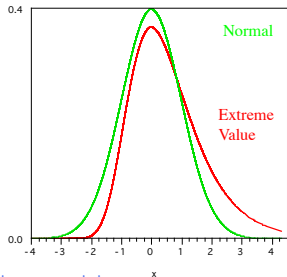
Contrast with central limit theorem

Alignment statistics

Bioinformatics Fall 2011

6

Normal vs. Extreme Value Distribution



Normal distribution:

$$y = (1/\sqrt{2\pi})e^{-x^2/2}$$

Extreme value distribution:

$$y = e^{-x} - e^{-x^2}$$

Alignment statistics

Bioinformatics Fall 2011

7

Extremal statistics of local alignment

- Possible to derive
 - $\text{Prob}(X \geq x) = 1 - e^{-e^{-(x-\mu)/\beta}}$, by integration ($y := e^{-(x-\mu)/\beta}$)
 - mean = $\mu + \gamma\beta$, γ = Euler's constant = 0.57
 - mode = μ
 - standard deviation = $\beta\pi/\sqrt{6}$
- Obtain μ and β experimentally, or by analyzing the score matrix
- Can compute the statistical significance of a given score

Alignment statistics

Bioinformatics Fall 2011

8

Obtaining μ and β from score matrix

- Assume that for a given score matrix
 - There exist a, b such that $\alpha_a \alpha_b s(a, b) > 0$
 - $\sum \alpha_a \alpha_b s(a, b) < 0$
- Let $\lambda^* > 0$ be the largest real root of
 - $f(\lambda) = \sum \alpha_a \alpha_b e^{\lambda s(a, b)} = 1$
 - $f(0) = 1$, $f'(0) < 0$, f tends to infinity in the limit
- $\mu = (\ln Kmn)/\lambda^*$, K is a constant > 0 → Stated without proof
- $\beta = 1/\lambda^*$
- $P(X \geq x) = 1 - e^{-e^{-(x-\mu)/\beta}} = 1 - e^{-Kmn e^{-\lambda^* x}}$
 - Approximate to $(Kmn e^{-\lambda^* x})$ for large values of x
- Expected score of an alignment = $\mu + \gamma\beta = \ln(Kmn)/\lambda^* + \gamma/\lambda^*$
- Standard deviation = $\pi/(\lambda^*\sqrt{6})$

Alignment statistics

Bioinformatics Fall 2011

9

P-value and E-value

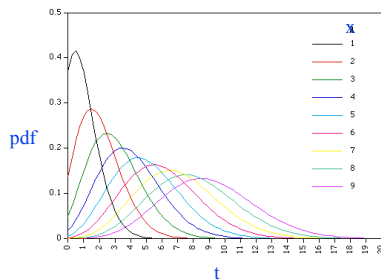
- The distribution of local alignments with score at least s is a Poisson distribution
 - $e^{-x} x^t / t!$ As score s increases, the mean x decreases
 - $P(t > 0) = 1 - e^{-x} = P(\text{score} \geq s) = P(X \geq s) = 1 - e^{-Kmn e^{-\lambda^* s}}$ From previous slide
 - mean = $x = Kmn e^{-\lambda^* s} = \text{E-value (in BLAST) / default} = 10^*/$
 - $1 - e^{-Kmn e^{-\lambda^* s}}$ is called the P-value Linear in length of query string
 - $(\lambda^* s - \ln K) / \ln 2$ is called the Bit score
 - E-value = $mn 2^{-\text{bit score}}$
 - m and n are the lengths of the database and the query string
 - K is an empirical constant < 1 ; λ^* is also in the same range

Alignment statistics

Bioinformatics Fall 2011

10

Poisson distribution



With high score thresholds, the mean x is very small

Alignment statistics

Bioinformatics Fall 2011

11

Another explanation

- Ewens and Grant, Statistical Methods in Bioinformatics
- Alignments of two sequences generates a random walk
 - Ladder points
 - Excursions
- Y = random variable denoting the highest value reached by an excursion
- $\text{Prob}(Y \geq y) \sim Ce^{-by}$
 - Geometric-like distribution
- Let X denote the maximum of a number of such iid random variables
 - $\text{Prob}(X \geq x) \approx 1 - e^{-Kmn e^{-\lambda^* x}}$
- The number of excursions scoring higher than s is a binomial distribution over a large number of trials with a probability of success $O(e^{-\lambda^* s})$.
- This is approximated by a Poisson distribution with mean $Kmn e^{-\lambda^* s} = \text{E-value}$.

Alignment statistics

Bioinformatics Fall 2011

12

Effect of gaps and short strings

- Model also extends to a gapped model
 - Scores tend to be higher
 - Estimation of parameters using random sequences
- Edge effects
 - Theory only for asymptotically long strings
 - Higher scores tend not to occur towards the edges
 - Correction for short strings
 - No correction needed for strings longer than 200

Alignment statistics

Bioinformatics Fall 2011

13

Reexamining the score matrix

- Define $p(a,b) = \alpha_a \alpha_b e^{-\lambda^* s(a,b)}$, called the target frequency
 - $\sum p(a,b) = 1$, from definition of λ^*
 - $s(a,b) = (1/\lambda^*) \ln [p(a,b)/(\alpha_a \alpha_b)]$
 - log-odds ratio !
- Every score matrix is uniquely determined by its target frequencies
- A given class of alignments is best distinguished by a score matrix whose target frequencies characterize the class.
- PAM and BLOSUM matrices attempt to do this.

Alignment statistics

Bioinformatics Fall 2011

14

Global alignment statistics

- S_n = global alignment score of two strings A and B of length n each drawn iid.
 - $E(S_n)/n \rightarrow \rho$
 - $\rho \geq E(s(A,B))$
 - Little is known about the value of ρ .
- $\text{Var}(S_n) \leq n(1-p)c$ Much harder problem
 - $c = \max\{0, \min\{s^*+2g(1), s^*-s_*\}\}$
 - $p = \text{prob of a match} = \sum \alpha_a^2$ Identical distribution for query and database
 - $s^* = \max\{s(a,b), \text{ for all } a,b\}$
 - $s_* = \min\{s(a,b), \text{ for all } a,b\}$
 - $g(k) = \text{score of a gap of length } k$
- $P(S_n - E(S_n) \geq kn) \leq e^{-k^2 n / 8c^2}$
- Generate distributions using random sequences and compute deviation from the mean to guess significance

Alignment statistics

Bioinformatics Fall 2011

15