

# CSE182-L6

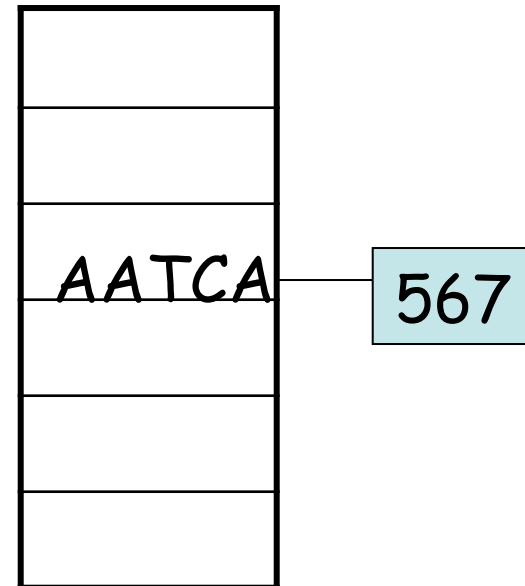
P-value and E-value  
Dictionary matching  
Pattern matching

# Why is BLAST fast?

- Assume that keyword searching does not consume any time and that alignment computation the expensive step.
- Query  $m=1000$ , random Db  $n=10^7$ , no TP
- $SW = O(nm) = 1000 * 10^7 = 10^{10}$  computations
- BLAST,  $W=11$ 
  - $E(\#11\text{-mer hits}) = 1000 * (1/4)^{11} * 10^7 = 2384$
  - Number of computations =  $2384 * 100 * 100 = 2.384 * 10^7$
  - Ratio =  $10^{10} / (2.384 * 10^7) = 420$
- Further speed improvements are possible

# Keyword Matching

- How fast can we match keywords?
- Hash table/Db index? What is the size of the hash table, for  $m=11$
- Suffix trees? What is the size of the suffix trees?
- Trie based search. We will do this in class.



# Silly Quiz



Skin patterns  
Facial Features



# Expectation?

- Some quantities can be reasonably guessed by taking a statistical sample, others not
  - Average weight of a group of 100 people
  - Average height of a group of 100 people
  - Average grade on a test
- Give an example of a quantity that cannot.
- When the distribution, and the expectation is known, it is easy to see when you see something significant.
- If the distribution is not well understood, or the wrong distribution is chosen, a wrong conclusion can be drawn

# P-value computation

- BLAST: The matching regions are expanded into alignments, which are scored using SW, and an appropriate scoring matrix.
- The results are presented in order of decreasing scores
- The score is just a number.
- How significant is the top scoring hits if it has a score  $S$ ?
- Expect/E-value (score  $S$ )= Number of times we would expect to see a random query generate a score  $S$ , or better
- How can we compute E-value?

# What is a distribution function

- Given a collection of numbers (scores)
  - 1, 2, 8, 3, 5, 3, 6, 4, 4, 1, 5, 3, 6, 7, ....
- Plot its distribution as follows:
  - X-axis = each number
  - Y-axis (count/frequency/probability) of seeing that number
  - More generally, the x-axis can be a range to accommodate real numbers

# P-value

- P-value: probability that a specific value (11) is achieved by chance.
- Compute an scores obtained by chance
  - 1, 2, 8, 3, 5, 3,6,12,4, 4,1,5,3,6,7
- Compute a Distribution
  - 1-2      XXX
  - 3-4      XXXX
  - 5-6      XXXX
  - 7-8      XX
  - 9-10
  - 11-13    X
  - 15-17
  - Are

# P-value computation

- A simple empirical method:
  - Compute a distribution of scores against a random database.
  - Use an estimate of the area under the curve to get the probability.
  - OR, fit the distribution to one of the standard distributions.

# Z-scores for alignment

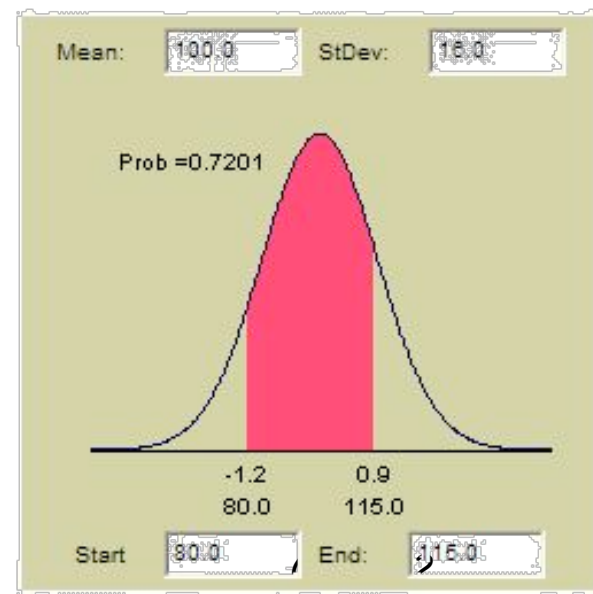
- Initial assumption was that the scores followed a normal distribution.
- Z-score computation:
  - For any alignment, score  $S$ , shuffle one of the sequences many times, and recompute alignment. Get mean and standard deviation

$$Z_S = \frac{S - \mu}{\sigma}$$

- Look up a table to

# Blast E-value

- Initial (and natural) assumption was that scores followed a Normal distribution
- 1990, Karlin and Altschul showed that ungapped local alignment scores follow an exponential distribution
- Practical consequence:
  - Longer tail.
  - Previously significant hits now not so significant



October 09

# Altschul Karlin statistics

- For simplicity, assume that the database is a binary string, and so is the query.
  - Let match-score=1,
  - mismatch score=-  $\infty$ ,
  - indel=- $\infty$  (No gaps allowed)
- What does it mean to get a score k?

# Exponential distribution

- Random Database,  $\Pr(1) = p$
- What is the expected number of hits to a sequence of  $k$  1's

$$(n - k)p^k \cong ne^{k \ln p} = ne^{-k \ln\left(\frac{1}{p}\right)}$$

- Instead, consider a random binary Matrix. Expected # of diagonals of  $k$  1s

$$\Lambda = (n - k)(m - k)p^k \cong nme^{k \ln p} = nme^{-k \ln\left(\frac{1}{p}\right)}$$

- As you increase  $k$ , the number decreases exponentially.
- The number of diagonals of  $k$  runs can be approximated by a Poisson process

$$\Pr[u] = \frac{\Lambda^u e^{-\Lambda}}{u!}$$

$$\Pr[u > 0] = 1 - e^{-\Lambda}$$

- In ungapped alignments, we replace the coin tosses by column scores, but the behaviour does not change (Karlin & Altschul).
- As the score increases, the number of alignments that achieve the score decreases exponentially

# Blast E-value

- Choose a score such that the expected score between a pair of residues  $< 0$
- Expected number of alignments with a score  $S$

$$E = Kmne^{-\lambda S} = mn2^{-\left(\frac{\lambda S - \ln K}{\ln 2}\right)}$$
$$\Pr(S \geq x) = 1 - e^{-Kmne^{-\lambda x}}$$

- For small values, E-value and P-value are the same

# The last step in Blast

- We have discussed
  - Alignments
  - Db filtering using keywords
  - Scoring matrices
  - E-values and P-values
- The last step: Database filtering requires us to scan a large sequence fast for matching keywords

# Keyword search

- Recall: In BLAST, we get a collection of keywords from the query sequence, and identify all db locations with an exact match to the keyword.
- Question: Given a collection of strings (keywords), find all occurrences in a database string where they keyword might match.

- End of lecture 6