Molecular Biophysics & Biochemistry 447b3 / 747b3

Bioinformatics

Mark Gerstein

Class 5, 1/26/98

Yale University

Aligning Text Strings

Raw Data ???

```
 T  C  A  T  G
    C  A  T  T  G
```

2 matches, 0 gaps

```
 T  C  A  T  G
          |  |
 C  A  T  T  G
```

3 matches (2 end gaps)

```
 T  C  A  T  G  .
    |  |  |
 .  C  A  T  T  G
```

4 matches, 1 insertion

```
 T  C  A  -  T  G
    |  |     |  |
 .  C  A  T  T  G
```

4 matches, 1 insertion

```
 T  C  A  T  -  G
    |  |  |     |
 .  C  A  T  T  G
```

Step 5 -- Traceback

Find Best Score (8) and Trace Back

A B C N Y - R Q C L C R - P M

A Y C - Y N R - C K C R B P

	A	B	C	N	Y	R	Q	C	L	C	R	P	M
A	**8**	7	6	6	5	4	4	3	3	2	1	0	0
Y	7	**7**	6	6	6	4	4	3	3	2	1	0	0
C	6	6	**7**	6	5	4	4	4	3	3	1	0	0
Y	6	6	6	5	**6**	4	4	3	3	2	1	0	0
N	5	5	5	6	5	4	4	3	3	2	1	0	0
R	4	4	4	4	4	**5**	4	3	3	2	2	0	0
C	3	3	4	3	3	3	3	**4**	3	3	1	0	0
K	3	3	3	3	3	3	3	3	**3**	2	1	0	0
C	2	2	3	2	2	2	2	3	2	**3**	1	0	0
R	2	1	1	1	1	2	1	1	1	1	**2**	0	0
B	1	2	1	1	1	1	1	1	1	1	1	0	0
P	0	0	0	0	0	0	0	0	0	0	0	**1**	0

The Score

S = Total Score

S(i,j) = similarity matrix score for aligning i and j

Sum is carried out over all aligned i and j

n = number of gaps (assuming no gap ext. penalty)

G = gap penalty

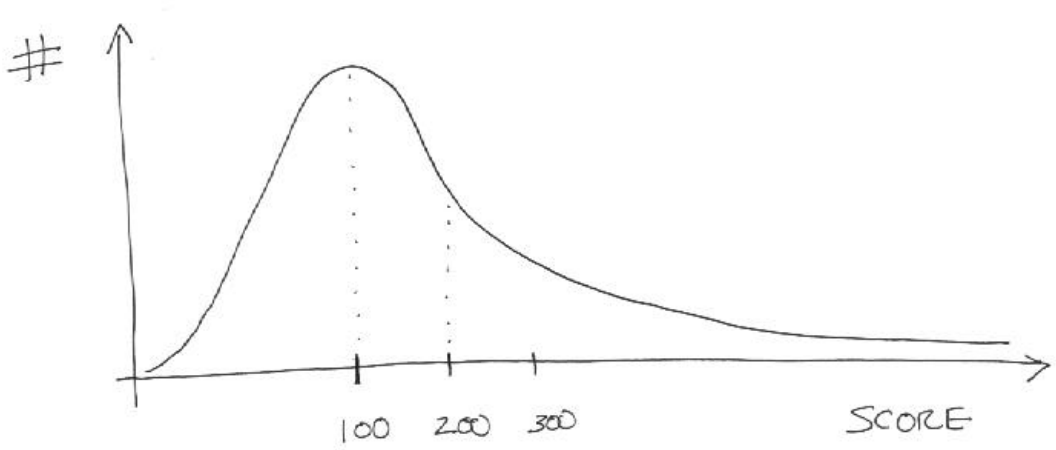
$$S = \sum_{i,j} S(i, j) - nG$$

What does a Score of 10 mean?

- What is the Right Cutoff?

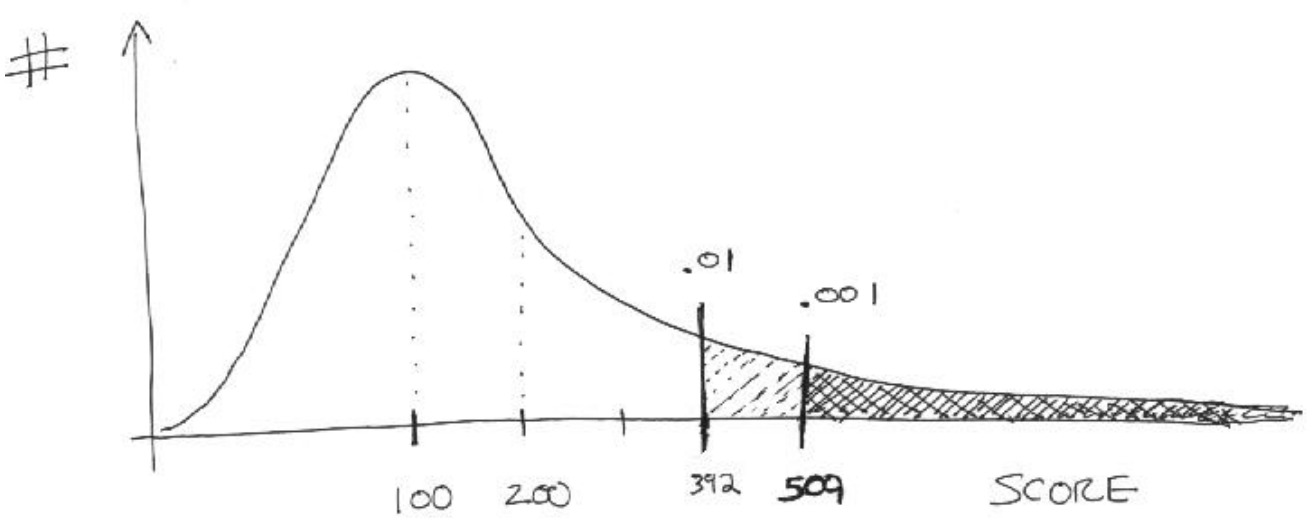
Score in Context of Other Scores

- How does Score Rank Relative to all the Other Possible Scores
  - ◊ P-value
  - ◊ Percentile Test Score Rank
- All-vs-All comparison of the Database (100K x 100K)
  - ◊ Graph Distribution of Scores
  - ◊ ~10^{10} scores much smaller number of true positives
  - ◊ N dependence



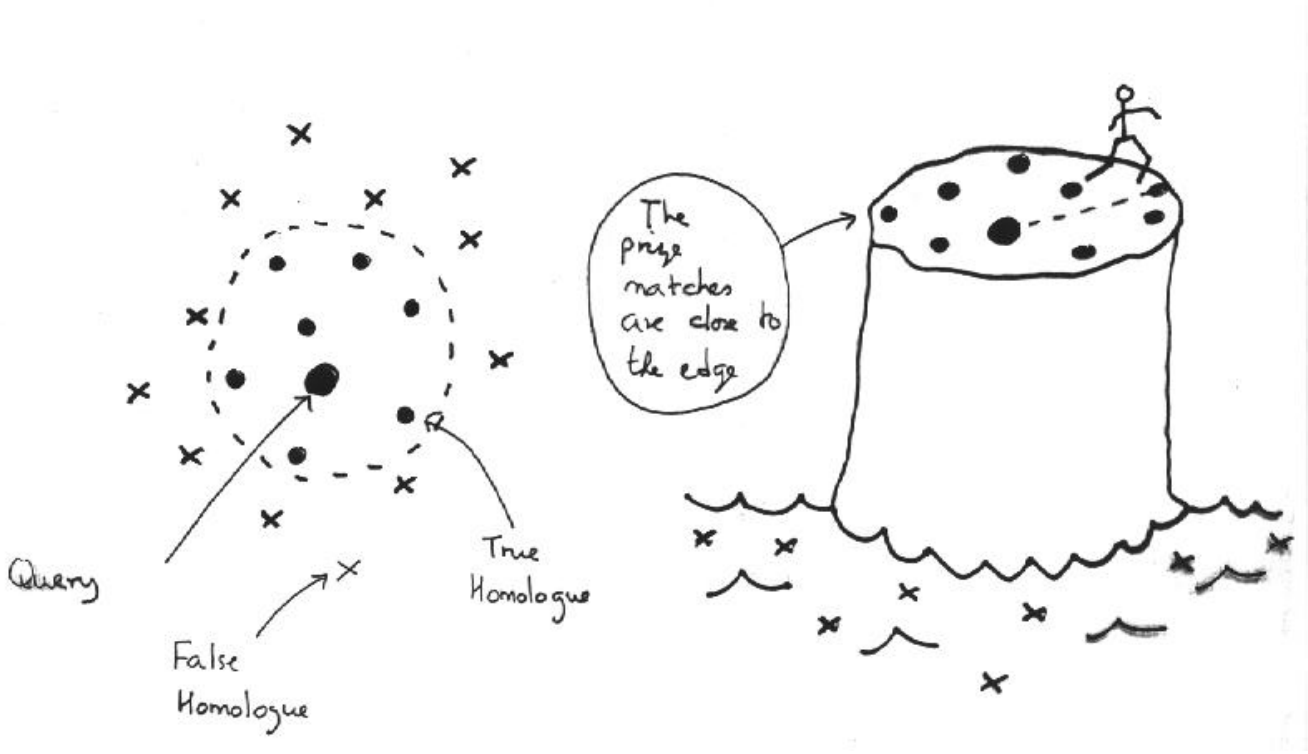
P-value in Sequence Matching

- P(s > S) = .01
  - ◊ P-value of .01 occurs at score threshold S (392 below) where score s from random comparison is greater than this threshold 1% of the time
- Likewise for P=.001 and so on.



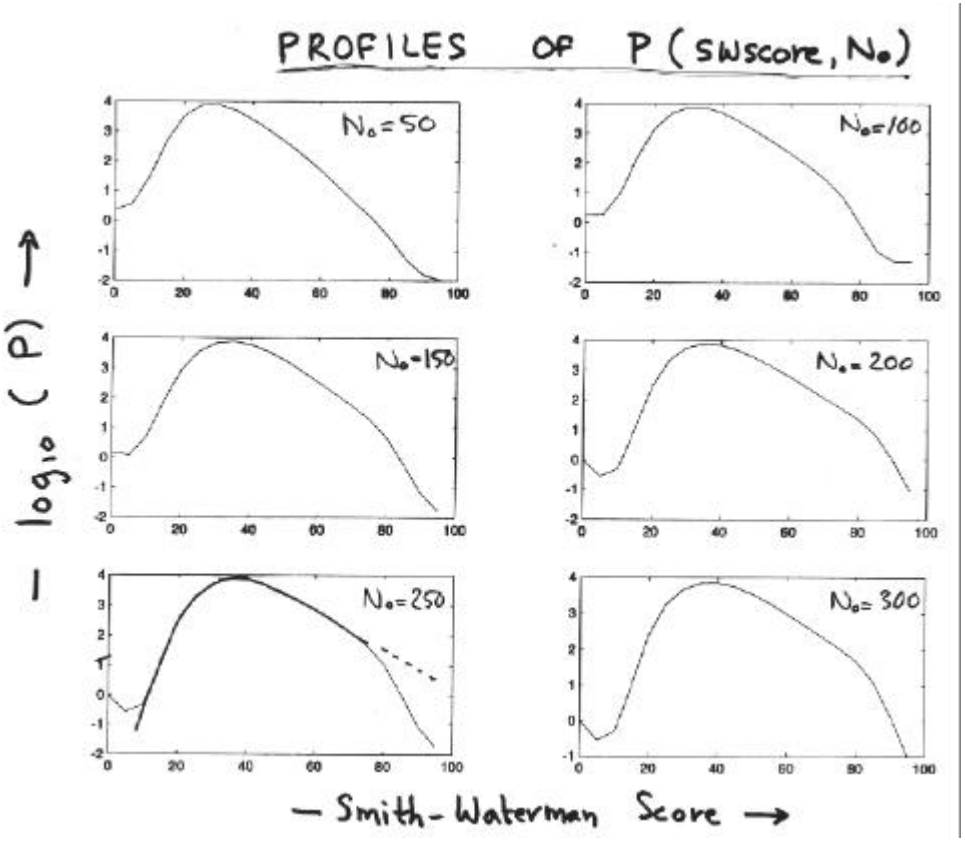
Objective is to Find Distant Homologues

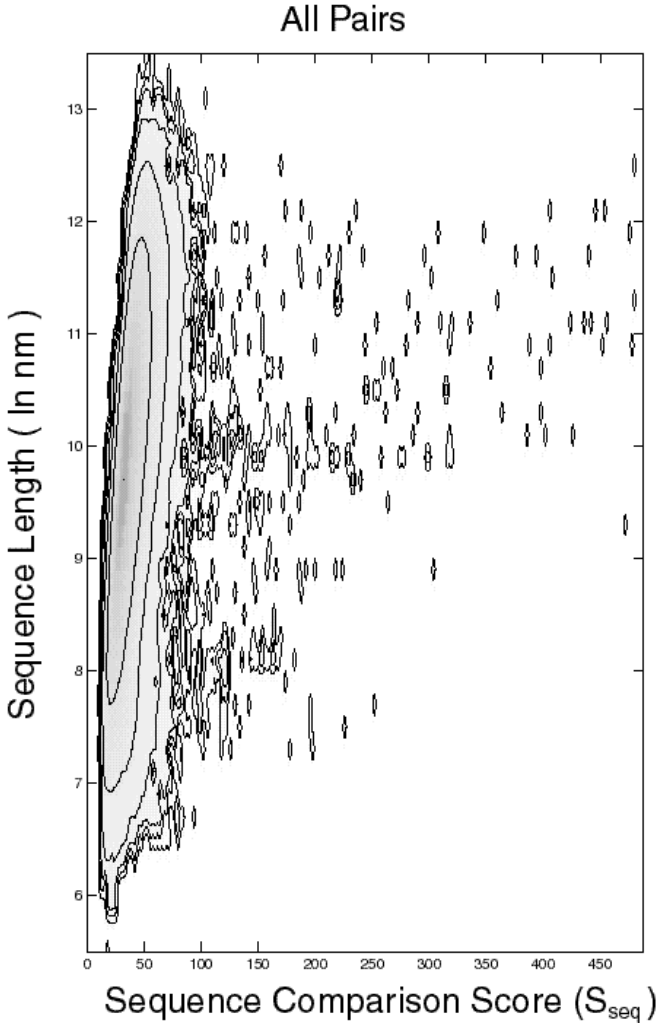
- Score (Significance) Threshold
- Maximize Coverage with an Acceptable Error Rate



What Distribution Really Looks Like

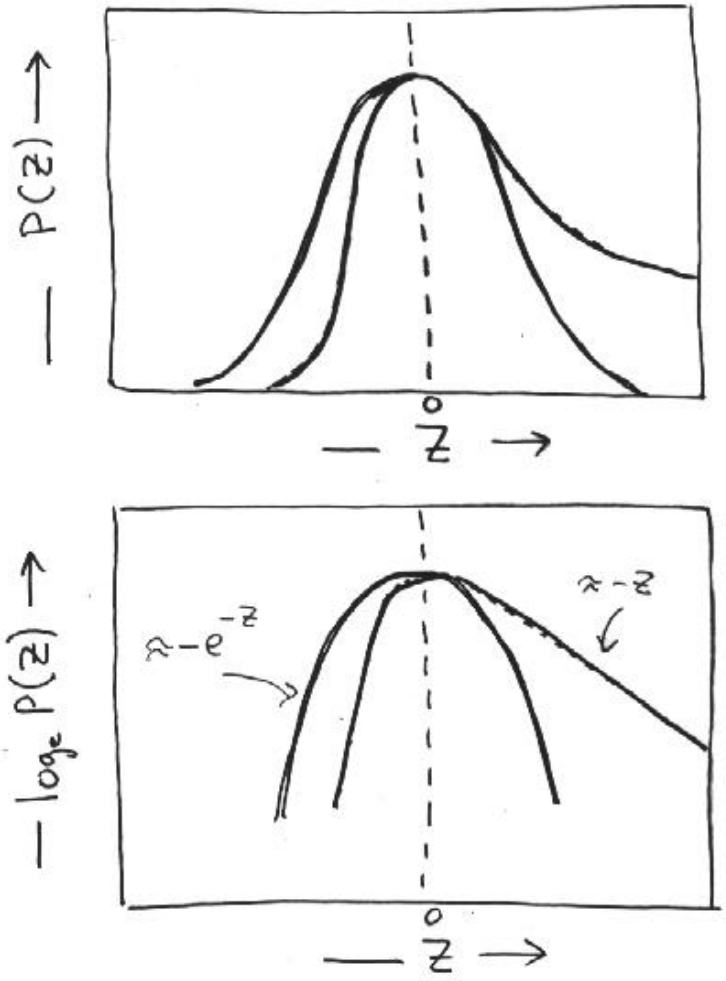
- N Dependence
- True Positives





Extreme Value Distribution Fits

- Normal
  $P(z) = \exp(-z^2)$
  $\ln P(z) = -z^2$
- Extreme Value
  (long tail)
  $P(z) = \exp(-z-\exp(-z))$
  $\ln P(z) = -z - e^{-z}$
- Good Fit Empirically for FASTA
  Analytic Formula For Blast



Explicit Form of the P-value in terms of Extreme Value Distribution

F(s) = E.V.D of scores

F(s) = exp(-Z(s) - exp(-Z(s)))

Z(s) = s/A + ln(NM) + B

= (s' - L)/W

s = Score from random S-W Alignment

L = most common one (mode)

W = width parameter (like SD)

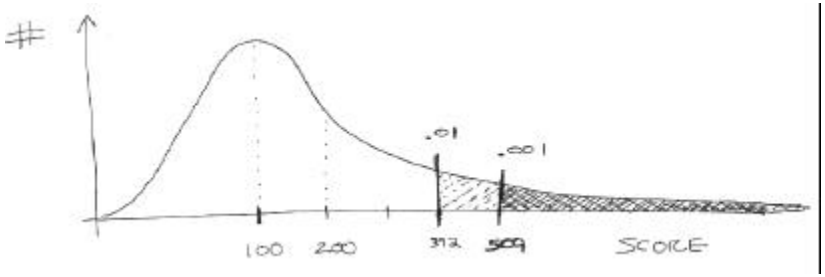
N & M are lengths of 2 seq.

A & B are fit parameters

P(s>S) = CDF = integral[F(s)]

P(s>S) = 1 - exp(-exp(-Z(s)))

Given Score Threshold S (1%), P (s > S) is the chance that a given random score **s** is greater than the threshold



Significance Depends on Database Size

- The Significance of Similarity Scores Decreases with Database Growth
  - ◊ The score between any pair of sequence pair is constant
  - ◊ The number of database entries grows exponentially
  - ◊ The number of nonhomologous entries >> homologous entries
  - ◊ Greater sensitivity is required to detect homologies
    Greater s
- Score of 100 might rank as best in database of 1000 but only in top-100 of database of 1000000

Low-Complexity Regions

- Low Complexity Regions
  - ◊ Different Statistics for matching
    AAATTTAAATTTAAATTTAAATTTAAATTT
    than
    ACSQRPLRVSHRSENCVASNKPQLVKLMTHVKDFCV
  - ◊ Automatic Programs Screen These Out (SEG)
- Also, Compositional Bias
  - ◊ Matching A-rich query to A-rich DB vs. A-poor DB

Practical Issues

- Local vs. Global Alignment
- Speed of Program (Query Hashing to make it FASTA)
- using BLAST, FASTA
- General Points