- What do we expect to find at all, if the sequence is long but finite?
- If we do not find an exact match of the word s, do we find at least a similar string?
- Do we find a high scoring local alignment of s and the database?
- How high scoring do we expect it to be?
- This obviously depends on the word s.
- It also depends on the length of the random sequence.
- The longer we can search, the more chances we have for finding something.

Some more notations

- To ensure local alignments, we now assume:

  $$\sum_{i,j} \mu_i \, \mu_j \, S_{ij} < 0$$

  and let

- Sequence1 $X_1, X_2, \ldots$
  Sequence2 $Y_1, Y_2, \ldots$
  be both i.i.d. sequences.

- We align the sequences (trivially and non optimally) by matching X_1 with Y_1, X_2 with Y_2, and so on ...

- Let $Z_i = (X_i, Y_i)$. We get a new i.i.d. sequence of real numbers $S(Z_1), S(Z_2), \ldots$, where $S(Z_i)$ is the score of matching X_i and Y_I.

- Let μ be the distribution of Z_i, we have

  $$E_\mu[S(Z_i)] = m_0 < 0$$

Large Deviations

By the Law of large numbers we have:

$$\frac{1}{n}\sum_{k=1}^{n} S(Z_k) \longrightarrow m_0$$

For $m > m_0$ let

$$A_n(m) := \left\{ \frac{1}{n}\sum_{k=1}^{n} S(Z_k) > m \right\}.$$

Again by the law of large numbers we have

$$\lim_{n} P\left[A_n(m)\right] = 0.$$

Events of this type are called *large deviations*.

High scoring alignments

- Let us assume that the first n positions of the two random sequences are very similar...
  ... more precisely let us assume they have a positive score.
- This event is a large deviation and hence, it becomes more and more unlikely the longer the sequences are.
- With which rate does the probability of this event converge to zero ?

Cramer's theorem

- **Theorem:** (Cramér 37)

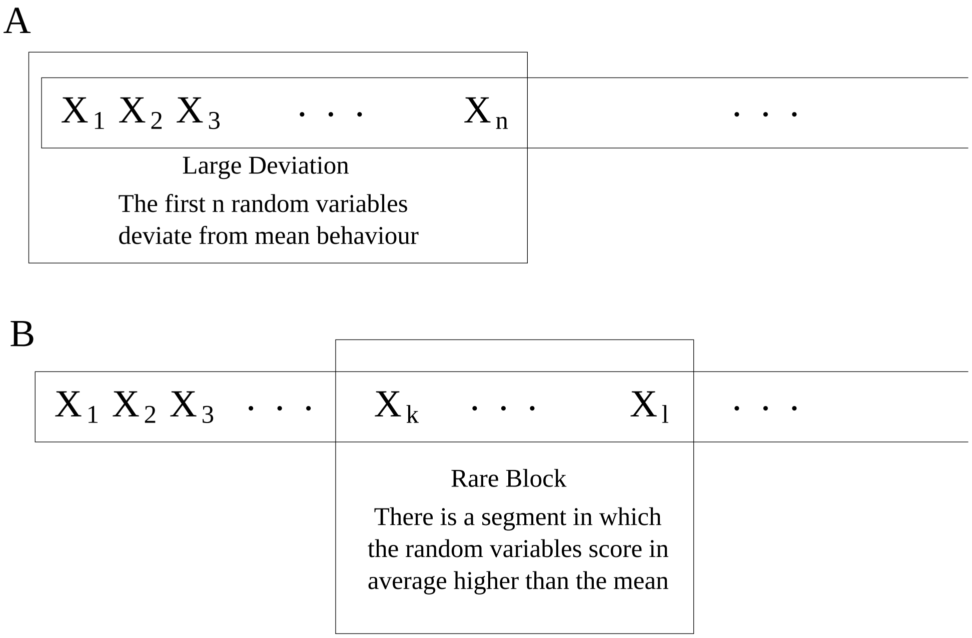
$$\lim_{n\to\infty} \frac{1}{n} \log \left(P \left[A_n(m) \right] \right) = -I(m, S, \mu) \quad a.s.\,.$$

- Probabilities of large deviations decrease exponentially: $P\,[A_n(m)] \approx \exp\,(-n\,I(m, S, \mu))$.

- There is an explicit formula for $I(m, S, \mu)$ ... see notes on the web.

Local Scores

- The above alignments start at the first position of the sequences.
- However, local alignments can start every where inside the sequences.
- For our problem, we are not interested in the probability that the start segment of the alignment is high scoring, but in the probability that there is at least one high scoring segment inside the alignment.
- We search the optimal scoring segment in the fixed alignment.
- The corresponding score is called the maximal local score of the sequence $S(Z_1), S(Z_2), \ldots$.


A
X_1 X_2 X_3 . . . X_n . . .
Large Deviation
The first n random variables
deviate from mean behaviour
B
X_1 X_2 X_3 . . . X_k . . . X_l . . .
Rare Block
There is a segment in which
the random variables score in
average higher than the mean

The Problem:

- For an i.i.d. sequence of real valued random variables we have the following problem:
  What is the probability that the sequence contains a segment with an average well above the expectation?

Rare Blocks

- Consider a long segment of the sequence $Z_1, Z_2, \ldots$ given by a start point k and an end point $l > k$.
- If k and l are selected randomly and if the length of the segment $l-k$ is long enough we expect by the law of large numbers:

$$\frac{1}{l-k} \sum_{i=k}^{l} S(Z_i) \approx m_0.$$

- Suppose the average score inside the segment is $m > m_0$. This can be described by the event

$$B_{k,l}(m) = \left\{ \frac{1}{l-k} \sum_{i=k}^{l} S(Z_i) > m \right\}.$$

Such events are called rare blocks. For fixed (k,l) Cramér's theorem yields

$$P\left[B_{k,l}(m)\right] \approx e^{-|l-k|\, I(m,S,\mu)} \quad .$$

Long rare blocks

- We will often observe short segments in the fixed alignment with a positive score. However, long stretches of similarity are rare.
- We are interested in the length R_n of the longest rare block inside a sequences of length n.
- Long sequences provide more possible starting points for a rare block than short ones.
- Therefore, R_n grows with the length of the i.i.d. sequence under consideration.
- What is the relationship between R_n and the length of $S(Z_1), S(Z_2), \ldots$?

Erdoes-Renyi Law

- The answer is given by

  Theorem: (Erdös Rényi 1970) Let $S(Z_1), S(Z_2), \ldots$ be i.i.d. with common mean m_0. For $m > m_o$ let R_n denote the length of the longest m-block in $(S(Z_1), \ldots, S(Z_n))$. We have

  $$\lim_{n\to\infty} \frac{R_n}{\log(n)} = \frac{1}{I(m,S,\mu)} \quad \text{a. s.}$$

- The key observation is that the maximal local score is significantly higher than the score of the sequence beginning at position 1. Consequently, it can be interpreted as a large deviation in the context of Cramer's theorem.

- See the notes for more details. The notes also come back to the problem of waiting times for rare blocks.

- Rare blocks grow on a logarithmic rate with the length of the underlying sequence.

Head Runs

- We toss a fair coin n times.
- The possible outcomes are Head (H) and Tail (T).

  HTHHTHTHTTHTHHHHHTTTTTTHTHTHTTHTHTTHTHHHHHHH
- How long is the longest head run?
- Erdoes Renyi: It depends on the length n of the sequence.

  The length of the longest head run is proportional to $log(n)$.

Exact matches

- We have a fixed alignment of two random sequences:
- Match is head, mismatch is tail:

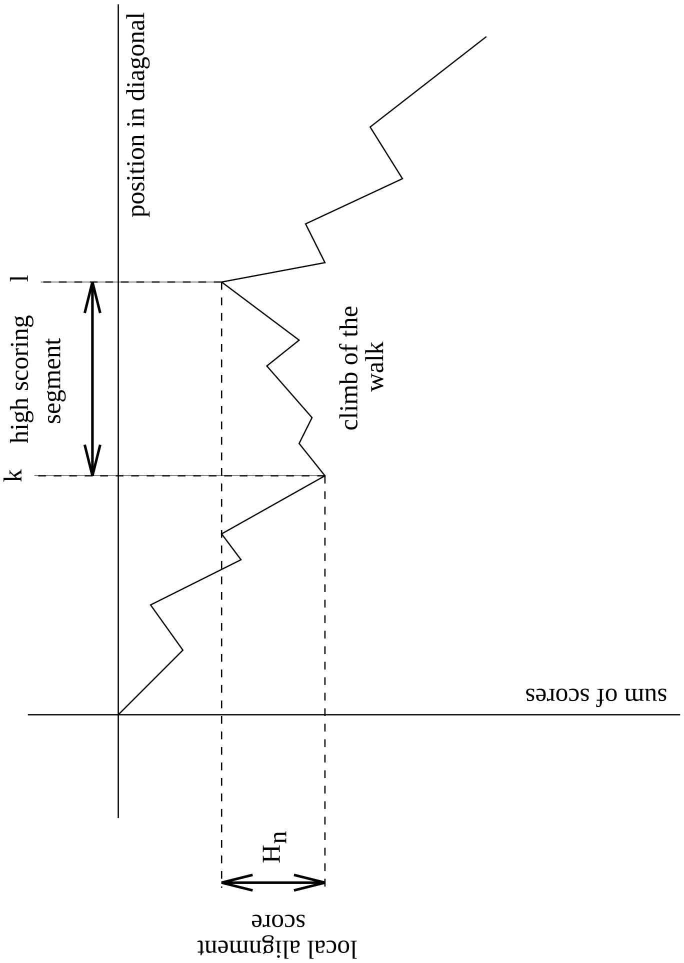
  TATAGCATCATTACT
  CATGGATCCATTCGA
  THHTHTTTHHHHTTT
- The length of the longest match is proportional to the log of the length of the alignment.

Climbs of a random walk with negative drift

- Consider the partial sums

$$W_n = \sum_{i=1}^{n} S(Z_i).$$

- Since the expectation of the C_i is negative, the sequence of partial sums $W_1, W_2, \ldots,$ yields a random walk with a negative drift.


position in diagonal
sum of scores
k
l
high scoring segment
climb of the walk
H_n
local alignment score

- With probability one we have

$$\lim_{n\to\infty} W_n = -\infty,$$

  and therefore for $l \gg k$ we have $W_l - W_k < 0$ with high probability.

- Now suppose we have a segment with $H = W_l - W_k = h_0 > 0$. Note that this segment corresponds to a gap-free local alignment with score h_0.

- The setup is very similar as in the Erdös-Rènyi law: ''typical'' behavior is a decreasing path, but since the walk is random there are also some short excursions of the walk, in which it is climbing against its drift.

- The main difference in the concept is, that the deviation is not determined by the length of the climb, but by its height.

Large deviations for local scores

The two large deviation problems of high alignment scores and rare blocks are closely related, and it is not surprising that they have similar answers.

Theorem:(Dembo and Karlin 1991) Let $W_1, W_2, \ldots$ be a random walk with discrete i.i.d. increments $S(Z_i)$ that have a negative expectation. If H_n denotes the height of the highest climb in $W_1, \ldots, W_n$, i.e.

$$H_n = \max_{l,k} W_l - W_k,$$

we have

$$\lim_{n\to\infty} \frac{H_n}{\log(n)} = \theta \quad \text{a. s. } ,$$

where θ is the unique positive solution of the equation

$$E\left[\exp(S(Z_1)/\theta)\right] = 1.$$

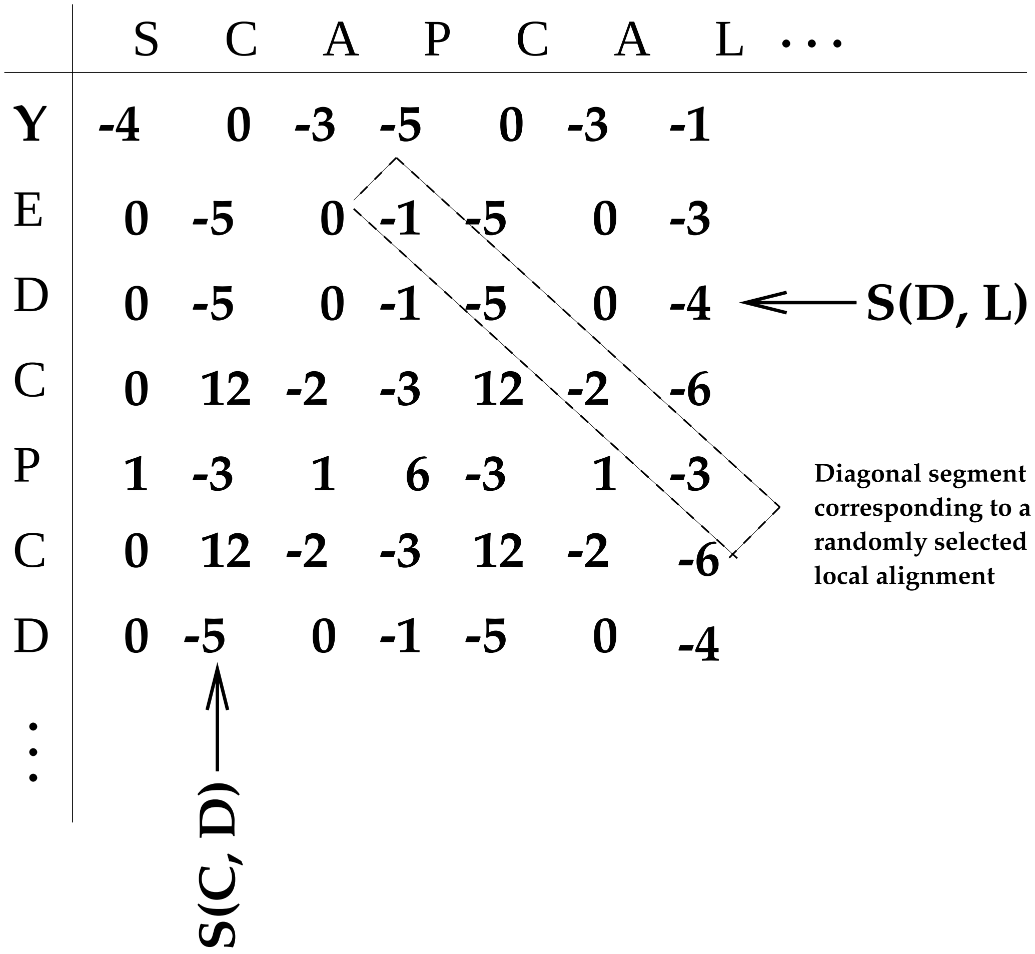
Optimal alignment

- Up to know, we have fixed the alignment
  X1 X2 X3 X4 ...
  Y1 Y2 Y3 Y4 ...

- That is not how real alignments are made.
- Let us now proceed to local alignments without gaps.
- We choose among all possible pairs of segments the one that optimizes the score.

Calculate a table C with

$$C_{i,j} = S(X_i, Y_j).$$



	S	C	A	P	C	A	L ⋯
Y	**-4**	**0**	**-3**	**-5**	**0**	**-3**	**-1**
E	**0**	**-5**	**0**	**-1**	**-5**	**0**	**-3**
D	**0**	**-5**	**0**	**-1**	**-5**	**0**	**-4**
C	**0**	**12**	**-2**	**-3**	**12**	**-2**	**-6**
P	**1**	**-3**	**1**	**6**	**-3**	**1**	**-3**
C	**0**	**12**	**-2**	**-3**	**12**	**-2**	**-6**
D	**0**	**-5**	**0**	**-1**	**-5**	**0**	**-4**
⋮							

S(D, L)

Diagonal segment corresponding to a randomly selected local alignment

S(C, D)

- There is a one to one correspondence of gap-free local alignments and segments of diagonals in this table.
- To each pair of equally long segments in the sequences there exists a corresponding segment of a diagonal, such that the sum over all entries in this diagonal segment yields the corresponding local alignment score.
- If the local alignment starts at positions X_i and Y_j and has length L its score is given by the sum over the diagonal segment $(C_{i,j}\ ,\ C_{i+1,j+1}\ ,\ \ldots\ ,\ C_{i+L-1,j+L-1})$

- Now, we are talking optimal local alignments without gaps.
- Rare blocks correspond to long alignments with a score above the average and climbs of a random walk correspond to high scoring alignments
- The main difference is that there are $n\,m$ possible starting points for a rare block or a climb of a random walk.
- The diagonals are dependent ... a technical problem that can be overcome.
  Dembo Karlin (1994)

- Erdoes Renyi: The length of the longest common word of two sequences of length n and m grows proportional to $log(n\,m)$
- Dembo Karlin: The optimal local ungapped alignment score grows proportional to $log(n\,m)$.
- The rates are different.

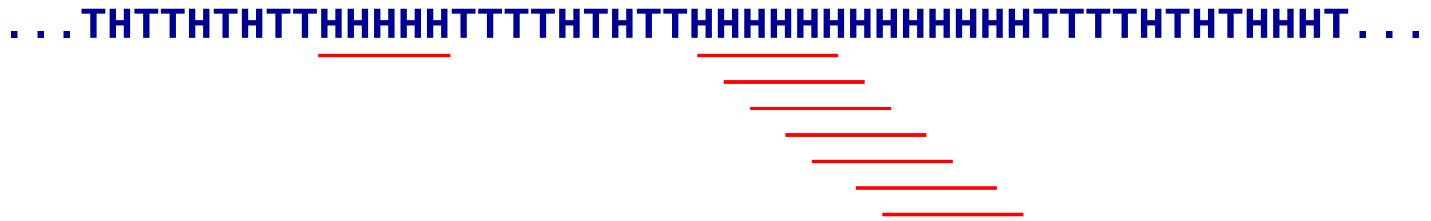
Some more Head Runs

- Match is head, mismatch is tail:
  TATAGCATCATTACT
  CATGGATCCATTCGA
  THHTHTTTHHHHTTT

- Let us fix a head run length t such that head runs of this length are rare.
- Let $W(t)$ be the number of head runs that are longer or equal to t.
- $W(t)$ is a count of rare events ...
- ... does it follow a Poisson distribution?

Clumping Head runs

t=5



...THTTHTHTTHHHHHTTTTHTHTTHHHHHHHHHHHHHTTTTHTHTHHHT...

Clump of head runs
that have length 5

Declumping Head runs

t=5

```
             ...TTHHHHHHHTTTTHTTHHHHHHHHHHHHHTTTTHTTTTHHHT...
head runs       001110000000000011111111100000000000000000
   clumps       001000000000000010000000000000000000000000
```

head runs =12

# Clumps = 2

A clump is a 1 which is preceeded by a 0

Poisson approximation

- The number of long head runs is not Poisson ... because of clumps
- We declump
- The number of clumps is Poisson distributed
- Not trivial because clumps are not independent
- Chen-Stein Theorem ... see Waterman's book for more details.

- Probability of a head is p and of a tail $(1-p)$.
- The length of the sequence is L
- Fix a minimal length t for a head run and a position n in the sequence.
- What is the probability P that a clump starts at n?

  1. $n = 1: \quad P = p^t$
  2. $1 < n \leq L - t: \quad P = (1-p)p^t$
  3. $n > L - t \qquad P = 0$

- The expected number of clumps in the sequence

$$\lambda(t) := E[W(t)] = p^t\{(n-t)(1-p)+1\}.$$

- Let $Z(\lambda)$ be a Poisson variable with intensity λ

$$P[Z(\lambda)=k] = \frac{e^{-\lambda}\,\lambda^k}{k!},$$

- $$E[Z(\lambda)] = \lambda$$

- Approximate $W(t)$ by $Z(\lambda(t))$.

Extreme value distribution

- If the longest head run is shorter than t, the count of head runs longer or equal to t is zero.

$$\{R_n < t\} = \{W(t) = 0\}$$

  and therefore

$$P[R_n \geq t] \approx 1 - e^{-\lambda(t)}.$$

- ... and therefore

$$P[R_n \geq t] \approx 1 - e^{-\lambda_n(t)}.$$

- For $\theta = -1/\log(p)$, large t, and $(n-t)/n \approx 1$ we have

$$\begin{aligned} P[R_n \geq t] &\approx 1 - \exp\left(-p^t\{(n-t)(1-p)+1\}\right) \\ &\approx 1 - \exp\left(-\gamma\, n\, e^{-\frac{t}{\theta}}\right), \qquad (1) \end{aligned}$$

- Let H_{nm} be the optimal ungapped local alignment score of two i.i.d. sequences with length n and m.
- Along the same lines as above, we can get

$$P[H_{nm} > t] \approx 1 - \exp(-\gamma\, m\, n\, e^{-\frac{t}{\theta}})$$

- Of course the parameters γ and θ are different.

- A standard extreme value distributed random variable G is defined by

$$P[G < t] = \exp(-e^{-t}).$$

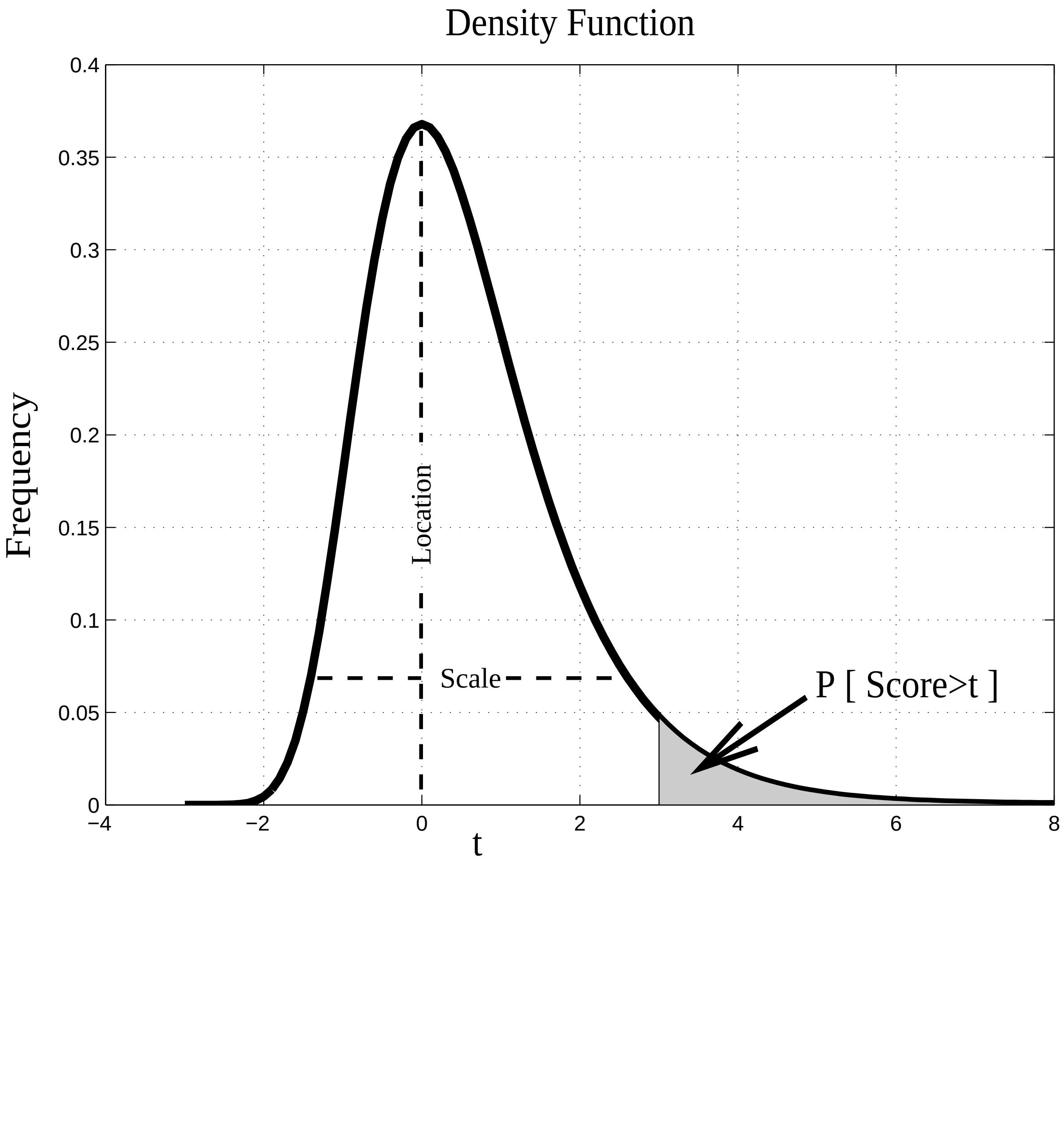
- Shifting and rescaling yields the two parameter family of variables

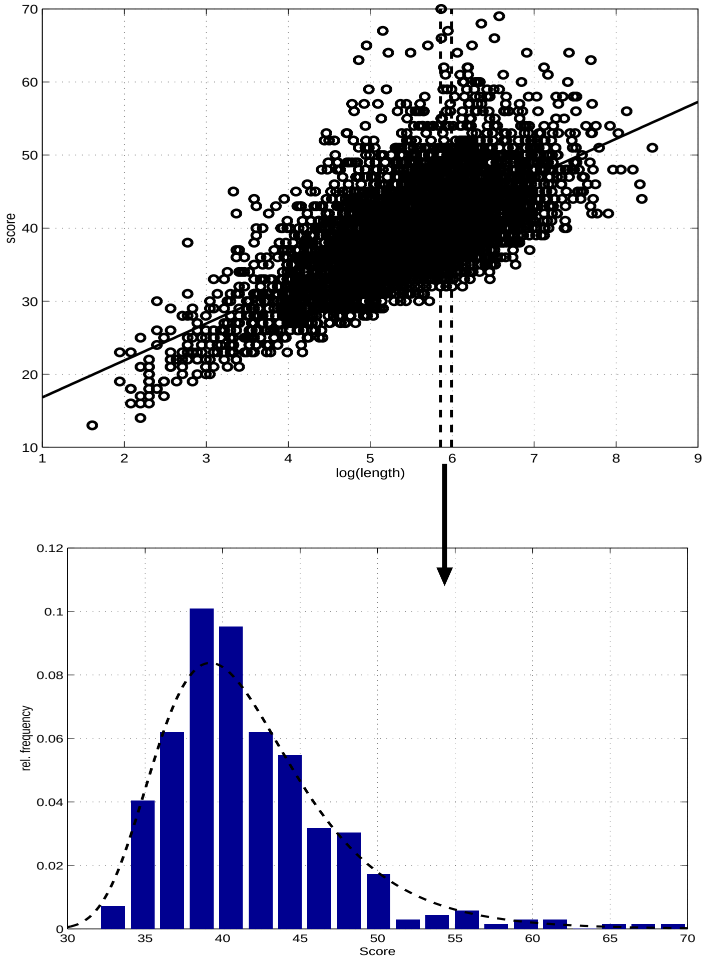
$$H = \theta\, G + \xi$$

We call ξ the location and θ the scale of H.

- With this notation we have the regression equation

$$H_{nm} \sim \alpha + \theta\, \log(n\, m) + \theta\, G.$$


Density Function
Frequency
t
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
−4
−2
0
2
4
6
8
Location
Scale
P [Score>t]


score
70
60
50
40
30
20
10
1
2
3
4
5
6
7
8
9
log(length)
rel. frequency
0.12
0.1
0.08
0.06
0.04
0.02
0
30
35
40
45
50
55
60
65
70
Score

- Note, that the location of the score distribution, depends on the length of sequences
- In the comparisons of a database search, the length of the query is fixed but the lengths of the database entries vary.
- Long sequences are more likely to produce a high score, than short ones
- Long sequences tend to show up as false positives in the twilight zone.
- We should compensate for this effect.

Length corrections

1. Rank sequences according to the p-values

$$P[H_{nm} > t] \approx 1 - \exp(-\gamma\, m\, n\, e^{-\frac{t}{\theta}})$$

2. ... or rank them according to length adjusted scores

$$A = H_{nm} - log(nm)\theta.$$