UNIVERSITÄT
DES
SAARLANDES


# Alignment Statistics

## Algorithms for Sequence Analysis

Sven Rahmann

Summer 2021

# Overview

## Previously: Scoring Pairwise Sequence Alignments

- Score maximization with general scoring schemes,
- Four variants: global, semiglobal, overlapping, local
- Derivation and estimation of score matrices

# Overview

## Previously: Scoring Pairwise Sequence Alignments

- Score maximization with general scoring schemes,
- Four variants: global, semiglobal, overlapping, local
- Derivation and estimation of score matrices

## Today's Lecture: Alignment Statistics

- Scores of local alignments of random sequences
- E-values and P-values of local alignment scores
- Functional form of score distributions
- Estimating parameters (ideas)

# Typical and Rare Scores of Local Alignments

## Setting

- We have locally aligned sequences of lengths $m$, $n$.
  Observed score is some $s \geq 0$.
- Is this **unusually** high, i.e., can it be explained by **random chance** or not?

# Typical and Rare Scores of Local Alignments

## Setting

- We have locally aligned sequences of lengths $m$, $n$.
  Observed score is some $s \geq 0$.
- Is this **unusually** high, i.e., can it be explained by **random chance** or not?

## Approach

- Compute local alignment score distribution on random sequences.
- Depends on parameters $\theta$
  - lengths $m, n$
  - scoring scheme (score matrix, gap costs)
  - random text model (uniform iid, iid, Markov, etc.)

# Typical and Rare Scores of Local Alignments

## Setting

- We have locally aligned sequences of lengths $m$, $n$.
  Observed score is some $s \geq 0$.
- Is this **unusually** high, i.e., can it be explained by **random chance** or not?

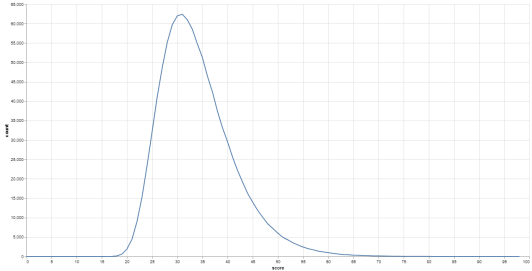
## Approach

- Compute local alignment score distribution on random sequences.
- Depends on parameters $\theta$
  - lengths $m, n$
  - scoring scheme (score matrix, gap costs)
  - random text model (uniform iid, iid, Markov, etc.)
- For **fixed parameters** $\theta$, $\mathbf{P}_\theta(S \geq s)$ is called the **p-value** of score $s$:
  probability that a local alignment of two random sequences
  achieves a score $S$ at least as high as the observed $s$.

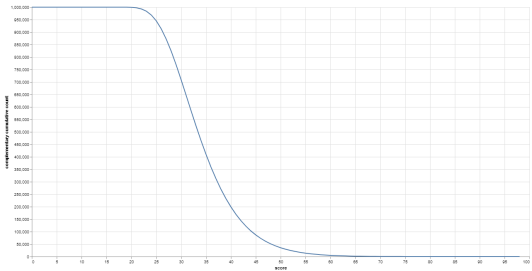
# Example: Simulation of Score Distribution

## Parameters

$T = 1\,000\,000$ random sequence pairs with $m = n = 100$,
BLOSUM62 score matrix, gaps $-5$, i.i.d. uniform amino acid frequencies.
Our interest is in the **far right tail** of the distribution (hard to simulate: rare events).



probability mass function (pmf)



complementary cumulative distribution function (ccdf)

# Theory

## Definitions

- $\theta$: parameters $m, n$, score matrix, gap penalties, text model
- $S$: random variable, optimal local alignment score of two random sequences
- $\mathbf{P}_\theta(S \geq s)$: **p-value** of observed score $s$ ($< 0.05$ is called significant).

# Theory

## Definitions

- $\theta$: parameters $m, n$, score matrix, gap penalties, text model
- $S$: random variable, optimal local alignment score of two random sequences
- $\mathbf{P}_\theta(S \geq s)$: **p-value** of observed score $s$ ($< 0.05$ is called significant).
- $N(s)$: (random) number of independent local alignments with score $\geq s$
- $\mathbf{E}_\theta[N(s)]$: **E-value** of $s$: expected number of local alignments with score $\geq s$

# Theory

## Definitions

- $\theta$: parameters $m, n$, score matrix, gap penalties, text model
- $S$: random variable, optimal local alignment score of two random sequences
- $\mathbf{P}_\theta(S \geq s)$: **p-value** of observed score $s$ ($< 0.05$ is called significant).
- $N(s)$: (random) number of independent local alignments with score $\geq s$
- $\mathbf{E}_\theta[N(s)]$: **E-value** of $s$: expected number of local alignments with score $\geq s$

## Assumption and Observation

If $s$ is sufficiently large ($\mathbf{E}[N(s)] \ll 1$, $\mathbf{P}(S \geq s) \leq 0.01$), we have a rare event.
Then $N(s)$ approximately has a Poisson distribution.

# The Poisson Distribution

## Intuition

Poisson distribution counts number of successes $X$ when

- there are many attempts $n \to \infty$,
- each has a small probability of success $p \to 0$,
- such that the expected number of successes $\lambda := np > 0$ is constant;

Limit of Binomial distribution $\mathbf{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$

# The Poisson Distribution

## Intuition

Poisson distribution counts number of successes $X$ when

- there are many attempts $n \to \infty$,
- each has a small probability of success $p \to 0$,
- such that the expected number of successes $\lambda := np > 0$ is constant;

Limit of Binomial distribution $\mathbf{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$

## Poisson Distribution

The entire distribution $\mathbf{P}(X = k)$ depends only on its expected value $\lambda > 0$:

$$P(X = k) = \mathrm{e}^{-\lambda} \cdot \lambda^k / k! \qquad (k = 0, 1, 2, \dots)$$

# The Poisson Distribution

## Intuition

Poisson distribution counts number of successes $X$ when

- there are many attempts $n \to \infty$,
- each has a small probability of success $p \to 0$,
- such that the expected number of successes $\lambda := np > 0$ is constant;

Limit of Binomial distribution $\mathbf{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$

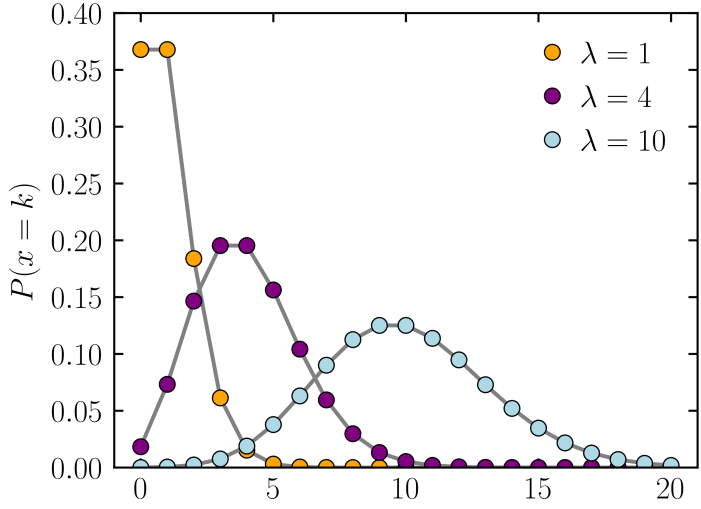
## Poisson Distribution

The entire distribution $\mathbf{P}(X = k)$ depends only on its expected value $\lambda > 0$:

$$P(X = k) = \mathrm{e}^{-\lambda} \cdot \lambda^k / k! \qquad (k = 0, 1, 2, \dots)$$

**Task:** Verify that $\sum_{k=0}^{\infty} \mathbf{P}(X = k) = 1$ and $\mathbf{E}[X] = \sum_{k=0}^{\infty} k\,\mathbf{P}(X = k) = \lambda$.

# Example: Poisson Distribution for Different Values of $\lambda$




Source: by Skbkekas - own work, CC BY 3.0, https://commons.wikimedia.org/w/index.php?curid=9447142

# Back to Theory

## Definitions

- $S$: random variable, optimal local alignment score of two random sequences
- $\mathbf{P}_\theta(S \geq s)$: **p-value** of observed score $s$
- $N(s)$: (random) number of independent local alignments with score $\geq s$
- $\mathbf{E}_\theta[N(s)]$: **E-value** of $s$: expected number of local alignments with score $\geq s$

## Assumption and Observation

- If $s$ is sufficiently large ($\mathbf{E}[N(s)] \ll 1$, $\mathbf{P}(S \geq s) \leq 0.01$), we have a rare event. Then $N(s)$ approximately has a Poisson distribution.
- Equivalent events: $N(s) \geq 1 \iff S \geq s$

# Back to Theory

## Definitions

- $S$: random variable, optimal local alignment score of two random sequences
- $\mathbf{P}_\theta(S \geq s)$: **p-value** of observed score $s$
- $N(s)$: (random) number of independent local alignments with score $\geq s$
- $\mathbf{E}_\theta[N(s)]$: **E-value** of $s$: expected number of local alignments with score $\geq s$

## Assumption and Observation

- If $s$ is sufficiently large ($\mathbf{E}[N(s)] \ll 1$, $\mathbf{P}(S \geq s) \leq 0.01$), we have a rare event. Then $N(s)$ approximately has a Poisson distribution.
- Equivalent events: $N(s) \geq 1 \iff S \geq s$

$$p = \mathbf{P}(S \geq s) = \mathbf{P}(N(s) \geq 1) = 1 - \mathbf{P}(N(s) = 0) = 1 - \mathrm{e}^{-\lambda} \approx \lambda = E \qquad (0 < \lambda \ll 1)$$

"For small E-values $E \ll 1$ and p-values $p \ll 1$, we have $p \approx E$."

# Computing the E-value (and p-value)

## Definitions

- $N(s)$: (random) number of independent local alignments with score $\geq s$
- $E_s = \mathbf{E}_\theta[N(s)]$: **E-value** of $s$: expected number of local alignments with score $\geq s$

# Computing the E-value (and p-value)

## Definitions

- $N(s)$: (random) number of independent local alignments with score $\geq s$
- $E_s = \mathbf{E}_\theta[N(s)]$: **E-value** of $s$: expected number of local alignments with score $\geq s$

## Assumptions and Observation

- $E_s$ decreases exponentially with $s$:
  Essentially, the only way of increasing the score is more matches in a row.

# Computing the E-value (and p-value)

## Definitions

- $N(s)$: (random) number of independent local alignments with score $\geq s$
- $E_s = \mathbf{E}_\theta[N(s)]$: **E-value** of $s$: expected number of local alignments with score $\geq s$

## Assumptions and Observation

- $E_s$ decreases exponentially with $s$:
  Essentially, the only way of increasing the score is more matches in a row.
- $E_s$ increases linearly with $m, n$:
  Longer sequence offer more locations where a high-scoring alignment could be.

# Computing the E-value (and p-value)

## Definitions

- $N(s)$: (random) number of independent local alignments with score $\geq s$
- $E_s = \mathbf{E}_\theta[N(s)]$: **E-value** of $s$: expected number of local alignments with score $\geq s$

## Assumptions and Observation

- $E_s$ decreases exponentially with $s$:
  Essentially, the only way of increasing the score is more matches in a row.
- $E_s$ increases linearly with $m, n$:
  Longer sequence offer more locations where a high-scoring alignment could be.
- For small $E_s$ (i.e., large enough $s$),

$$p_s \approx E_s \approx C \cdot mn \cdot q^s \qquad (C > 0 \text{ and } 0 < q < 1)$$

  with constants $C, q$ depending on scoring scheme and text model.

# Estimating Constants $C, q$

## Logarithmic view: Affine function

Examine high-scoring tail of score distribution of random local alignments

$$
\begin{aligned}
p_s \approx E_s &\approx C \cdot mn \cdot q^s && (C > 0 \text{ and } 0 < q < 1) \\
\log p_s &\approx \log C + \log(mn) + s \cdot \log q \\
&= K + \log(mn) - \lambda s && (K = \log C \text{ and } \lambda > 0)
\end{aligned}
$$

$\Rightarrow$ log p-value is a falling **affine function** of $s$ with **slope** $-\lambda$, **offset** $K + \log(mn)$.
(Here $\lambda : -\log q > 0$ and $K := \log C$.)

# Estimating Constants $C, q$

## Logarithmic view: Affine function

Examine high-scoring tail of score distribution of random local alignments

$$\begin{aligned} p_s \approx E_s &\approx C \cdot mn \cdot q^s && (C > 0 \text{ and } 0 < q < 1) \\ \log p_s &\approx \log C + \log(mn) + s \cdot \log q \\ &= K + \log(mn) - \lambda s && (K = \log C \text{ and } \lambda > 0) \end{aligned}$$

$\Rightarrow$ log p-value is a falling **affine function** of $s$ with **slope** $-\lambda$, **offset** $K + \log(mn)$.
(Here $\lambda : -\log q > 0$ and $K := \log C$.)

## Naïve simulation

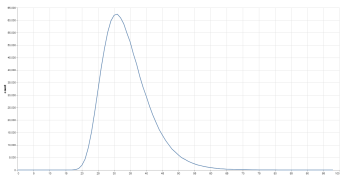
- Create $T$ random sequence pairs of length $m, n$ according to text model
- Compute $T$ optimal local alignment scores $S_1, \ldots S_T$ and empirical p-values $\hat{p}_s := |\{i : S_i \geq s\}|/T$ for all sufficiently large $s$

# Estimating Constants

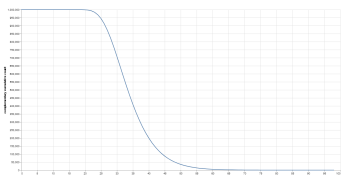
## Fit an affine function from empirical p-values

We have a functional form and empirical observations for $\log p_s$:

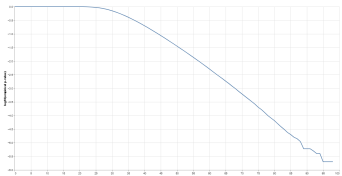
$$\log p_s \approx K + \log(mn) - \lambda s$$
$$\log \hat{p}_s = \log(|\{i : S_i \geq s\}|/T)$$



pmf

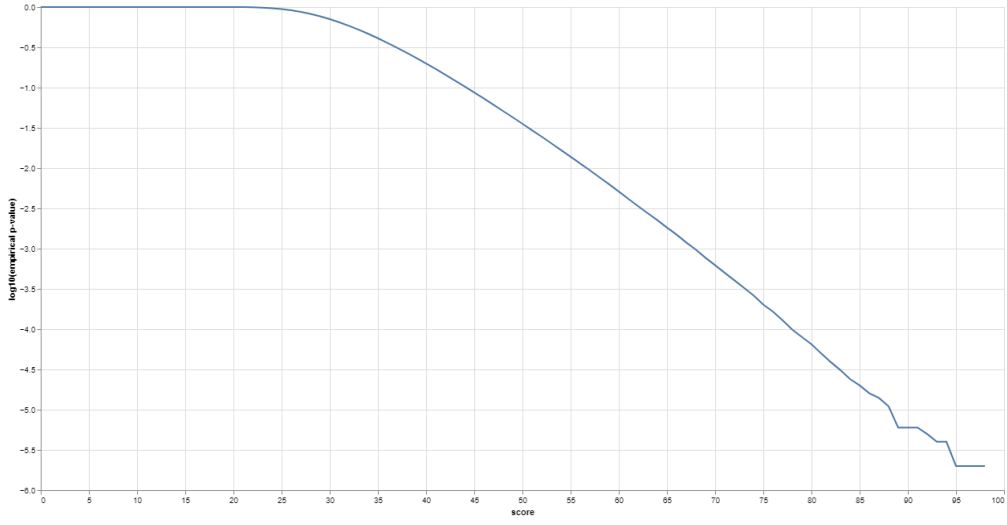


ccdf

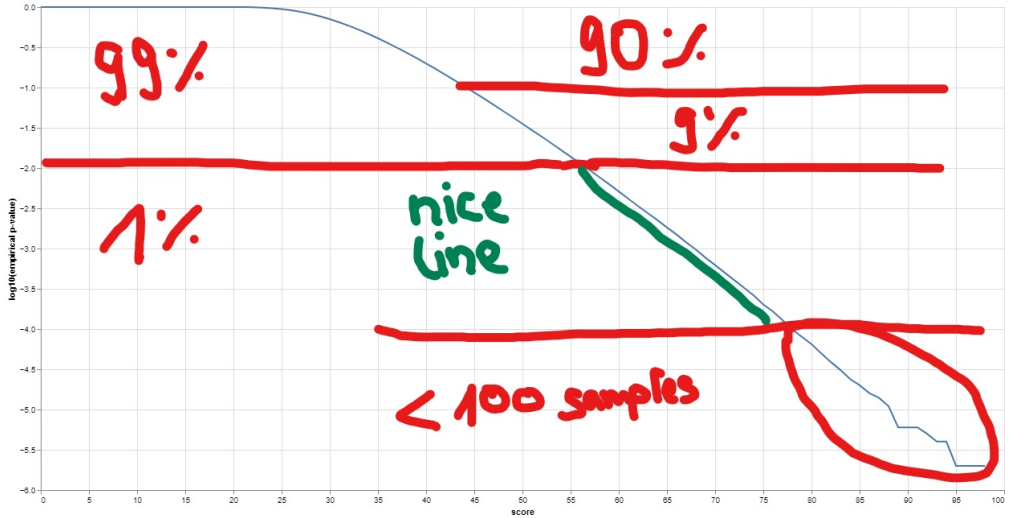


log ccdf

# Estimating Constants: Fit affine function



# Estimating Constants: Fit affine function



# Challenges and Possible Solutions

## Rare events (upper 1% scores) are hard to simulate

- Need 1M ($10^6$) samples to have 10 000 ($10^4$) samples in upper 1%.
- Cannot fit well for very rare events (too few samples, say $< 100$).
- Must fit the (theoretical) functional shape on a limited range.
- High computational load for limited effectiveness (99% of simulation useless).

# Challenges and Possible Solutions

## Rare events (upper 1% scores) are hard to simulate

- Need 1M ($10^6$) samples to have 10 000 ($10^4$) samples in upper 1%.
- Cannot fit well for very rare events (too few samples, say $< 100$).
- Must fit the (theoretical) functional shape on a limited range.
- High computational load for limited effectiveness (99% of simulation useless).

## More effective simulations

- Use more than one score per random sequence pair:
  many **independent local maxima** in local alignment matrix
  (islands in a sea of zeros)

# Challenges and Possible Solutions

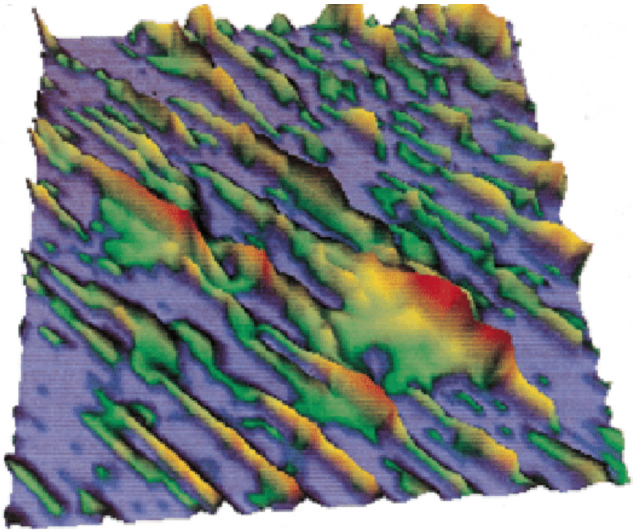
## Rare events (upper 1% scores) are hard to simulate

- Need 1M ($10^6$) samples to have 10 000 ($10^4$) samples in upper 1%.
- Cannot fit well for very rare events (too few samples, say $< 100$).
- Must fit the (theoretical) functional shape on a limited range.
- High computational load for limited effectiveness (99% of simulation useless).

## More effective simulations

- Use more than one score per random sequence pair:
  many **independent local maxima** in local alignment matrix
  (islands in a sea of zeros)
- Use **importance sampling**: Sample rare events more frequently,
  apply correction factor for computing empirical p-values.
  (Details can be difficult; ongoing research)

# Islands in a Sea of Zeros



Color-coded 3D visualization of a local alignment matrix.

The peak score of every independent island can be considered of the histogram, not only the highest peak.

**Source:**
SF Altschul, R Bundschuh, RM Olsen, T Hwa: The estimation of statistical parameters for local alignment score distributions. Nucleic acids research 29:2(2001) 351–361.

# Importance Sampling

Create pairs of more related sequences by a random walk that locally modifies each sequence. Compute correction factors (probability of independent random pair vs. probability of pair created by random walk of certain length).

**Source:**
S Wolfsheimer, I Herms, S Rahmann et al. Accurate statistics for local sequence alignment with position-dependent scoring by rare-event sampling. BMC Bioinformatics 12, 47 (2011).

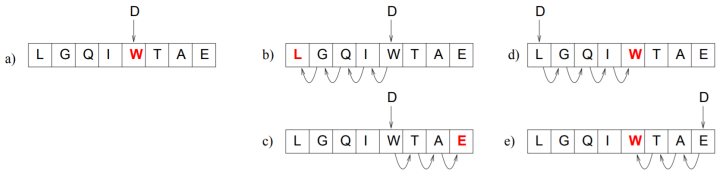



**Figure 2 Monte Carlo moves used in the simulation**. (a) substitution, (b) insertion with left shift, (c) insertion with right shift,(d) deletion with right shift and (e) deletion with left shift.

# Key Points

## Significance depends on random model and parameters

- Random text models: i.i.d. uniform, i.i.d., Markov, etc.
- Sequence lengths $m, n$, score matrix, gap costs
- **p-value hacking:** tuning model + parameters until results become significant. **Scientific fraud**, but unfortunately relatively widespread behavior.

# Key Points

## Significance depends on random model and parameters

- Random text models: i.i.d. uniform, i.i.d., Markov, etc.
- Sequence lengths $m, n$, score matrix, gap costs
- **p-value hacking:** tuning model + parameters until results become significant. **Scientific fraud**, but unfortunately relatively widespread behavior.

## Functional form is robust against model changes

- Same behavior $p(s) = C \cdot \exp(-\lambda s)$ holds for many variations of the assumptions:
  - Both sequences random, same composition (today)
  - Both sequences random, different compositions
  - Only one sequence random, other sequence fixed
- Generalization: Locally aligning a query sequence to a **pangenome graph**; same parametric form apparently also holds (current research).

# Summary

## Alignment Statistics

- Simple random sequence models
- Definition: E-value, p-value of an observed score
- Score ccdf (complementary cumulative distribution function) $\rightarrow$ p-values
- Exponential decrease of p-value with increasing score (line in log-plot)
- Parametric form $p(s) = C \cdot \exp(-\lambda s)$ is robust against model changes
- Challenge: estimating $C, \lambda$ efficiently (e.g., importance sampling)

# Possible Exam Questions

- Define the p-value and E-value of an observed local alignment score.
- Why is p-value $\approx$ E-value when both are very small?
- Explain why the parametric form $p(s) = C \cdot \exp(-\lambda s)$ holds.
- How do the sequence lengths $m, n$ enter the parametric form?
- How can the parameters $C, \lambda$ be estimated?