A Fast and Simple Method for Mining Subsequences with Surprising Event Counts

Jefrey Lijffijt
Helsinki Institute for Information Technology (HIIT)
Department of Information and Computer Science
Aalto University, Finland

Summary

- Situation: We study event sequences
  - Sequence of labels, e.g., words or amino acids
  - (*A, C, T, G, G, C, G, G, A, T, T, A*)

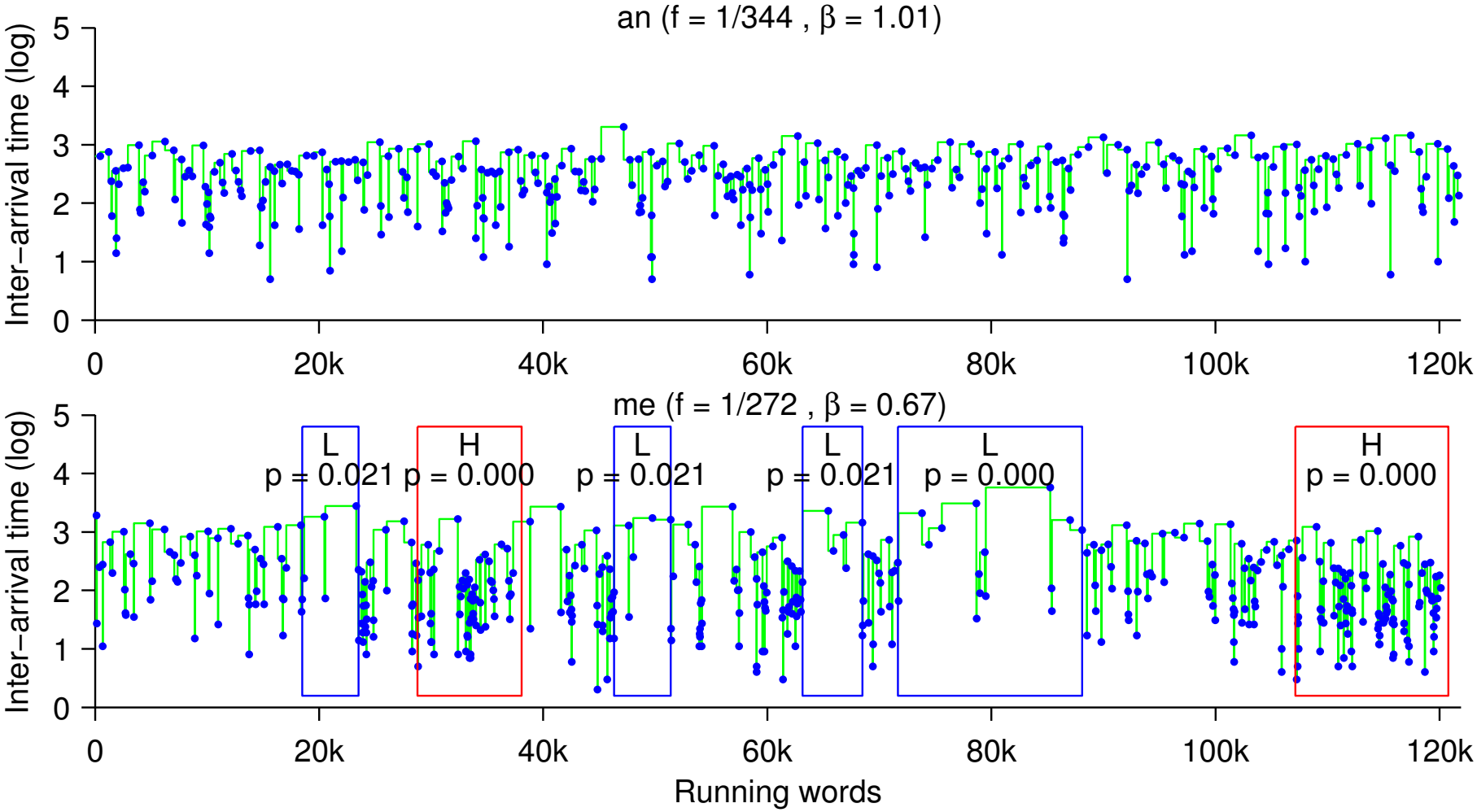
- Aim: Find subsequences where a given event is surprisingly frequent or infrequent

- Subsequence = part of a long event sequence

- Surprising = improbable, assuming no structure

Example: text analysis

Burstiness measure
[Altmann et al. 2009]



Some basics

- Approach is based on statistical significance testing
- The null hypothesis is that the event probability is *p*
- Given a random subsequence of length *m* with count *k*

$$p_H = \sum_{i=k}^{m} Bin(i; m, p) = \sum_{i=k}^{m} \binom{m}{i} p^i (1-p)^{m-i}$$

Problem setting

- Basic procedure:
  low p-value → significant structure
- However, we analyse all subsequences of a given length *m*
- Account for multiple hypotheses to prevent spurious results
- Family-wise error rate control:
  $Pr(FP > 0) \leq \alpha$

		Declared significant	
		No	Yes
Null hypothesis	True	TN	FP
	False	FN	TP

Traditional solutions

- Apply post-hoc correction (Hochberg's procedure)
  → Low power, does not account for dependencies
- Or, use randomisation
  → Computationally demanding
- Alternative proposed in paper: analytical upper-bound

The dependency structure

- Full sequence: $(X_1,\ldots,X_n): X_i \in \{0,1\},\ \Pr(X_i = 1) = p$

- Sequence 1: $(X_1,\ldots,X_m)$
- Sequence 2: $(X_2,\ldots,X_{m+1})$
- …
- Sequence n-m+1: $(X_{n-m+1},\ldots,X_n)$

- Test statistic $Z_{i,m} = X_i + \ldots + X_{i+m-1}$

- FWER adjusted p-value: $p_H^{(k)} = \Pr\left(\bigcup_{i=1}^{n-m+1} \{Z_{i,m} \geq k\}\right)$

Approximation

- Computing this exactly is computationally costly

- Approximation:

$$\Pr\left(\bigcup_{i=1}^{n-m+1}\{Z_{i,m} \geq k\}\right)$$

$$= \Pr\left(\{Z_{1,m} \geq k\}\right) + \Pr\left(\{Z_{2,m} \geq k\} \cap \{Z_{1,m} < k\}\right)$$

$$+ \Pr\left(\{Z_{3,m} \geq k\} \cap \{Z_{2,m} < k\} \cap \{Z_{1,m} < k\}\right) + \ldots$$

$$\leq \Pr\left(\{Z_{1,m} \geq k\}\right) + (n-m) \cdot \Pr\left(\{Z_{2,m} \geq k\} \cap \{Z_{1,m} < k\}\right)$$

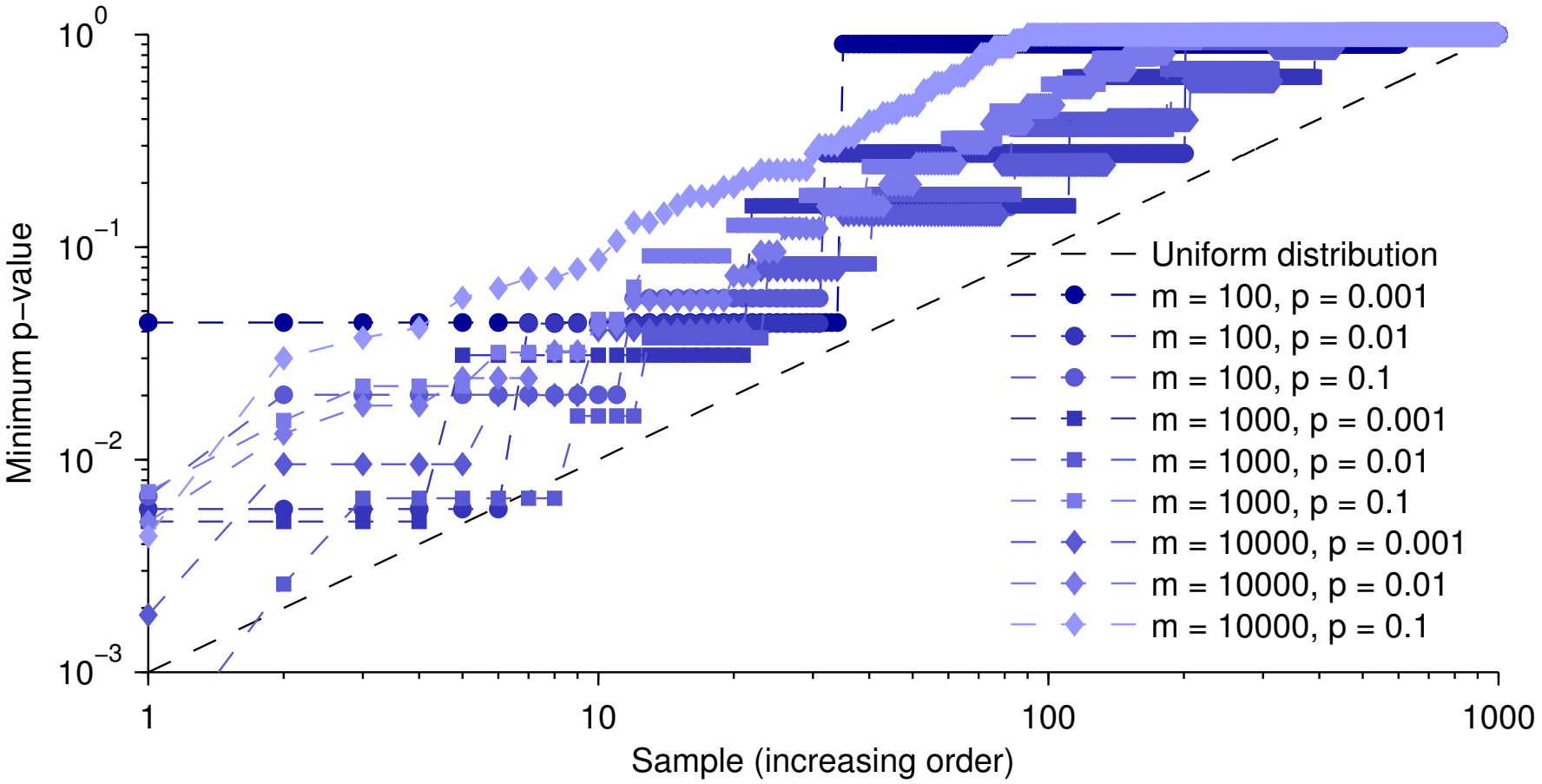
Upper bound

- $$\tilde{p}_H = (n-m)\cdot \Pr\left(\{Z_{2,m} \geq k\} \cap \{Z_{1,m} < k\}\right) + \Pr\left(\{Z_{1,m} \geq k\}\right)$$
$$= (n-m)\cdot(1-p)\cdot p \cdot Bin(k-1; m-1, p) + \sum_{i=k}^{m} Bin(i; m, p)$$

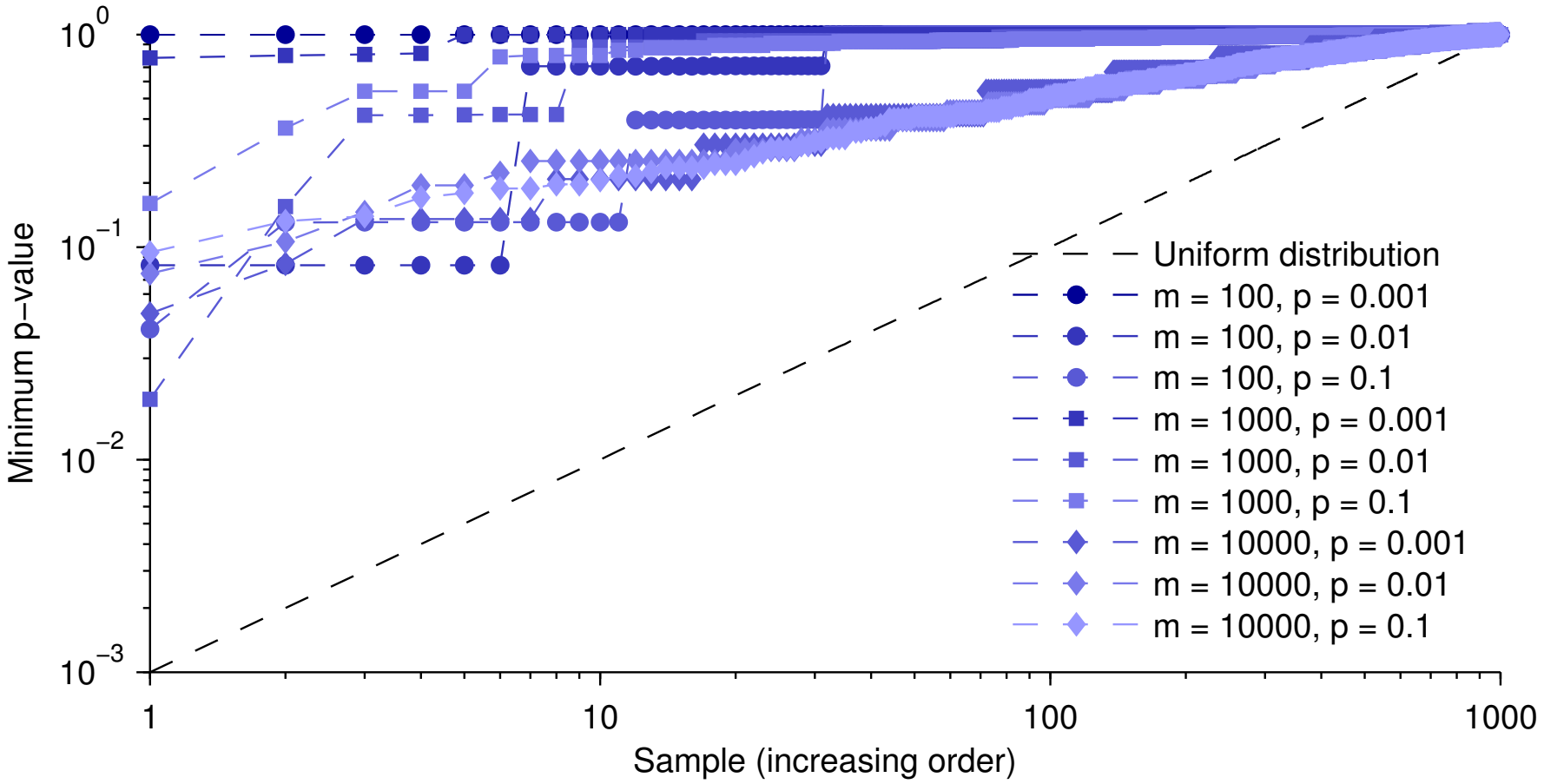
- Binomial and cumulative binomial can be computed in O(1) time [Loader, 2000]

- See paper for upper bound in low-frequency direction and for subsets of all subsequences

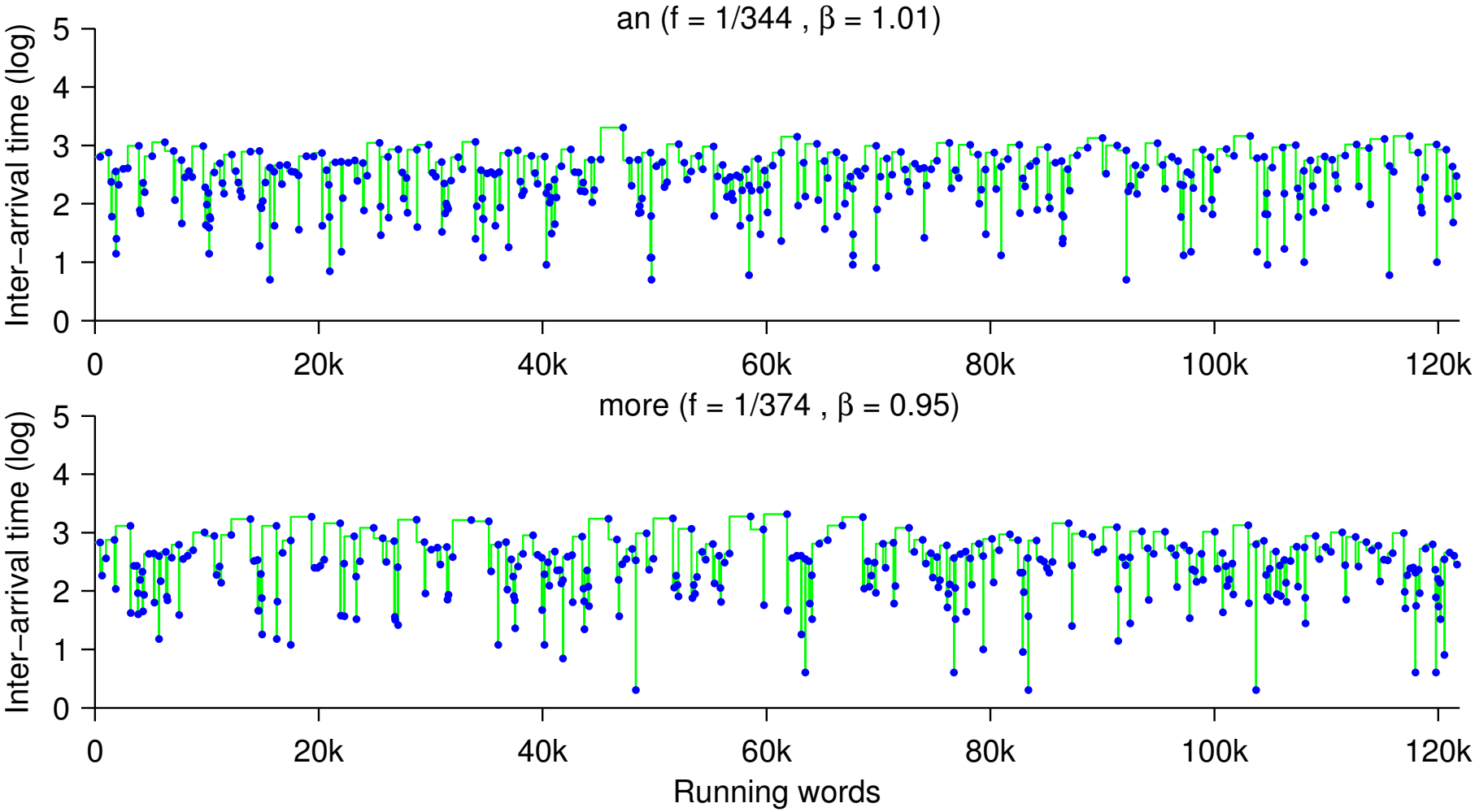
Uniformity/Power



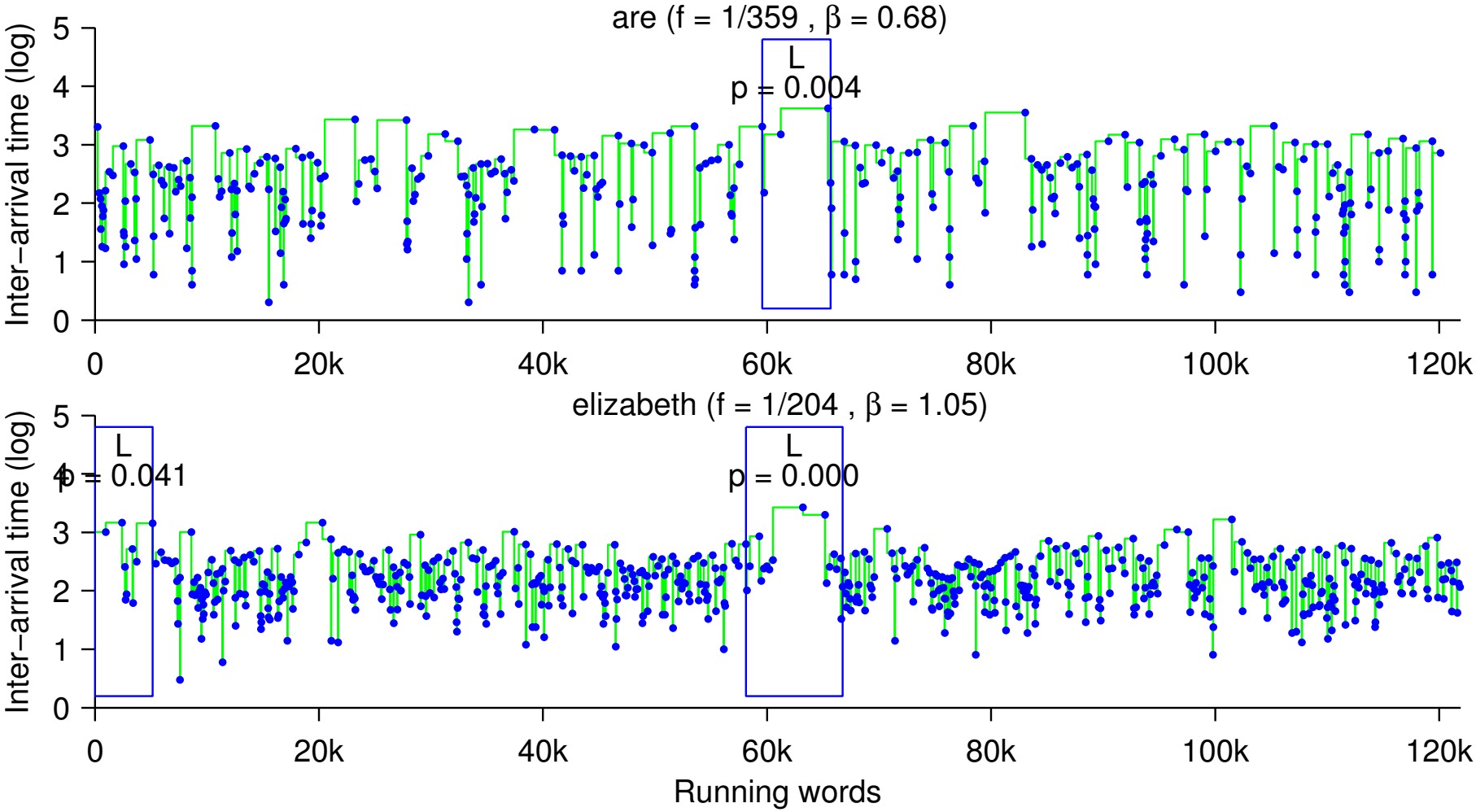
Uniformity Hochberg's procedure



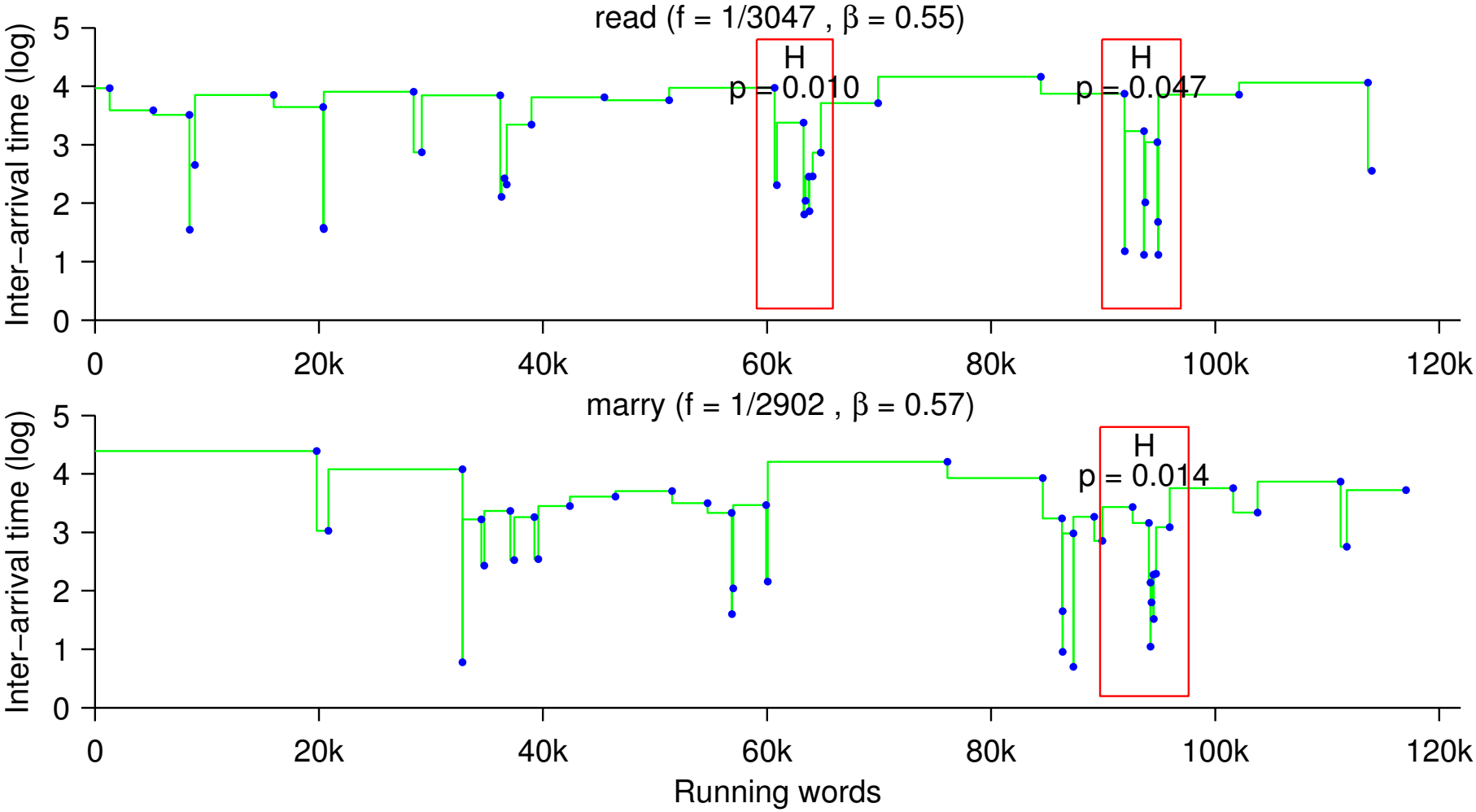
Some words occur uniformly



We find more than β measure indicates



You can find bursts that co-occur



Conclusion

- Aim:
  - Find subsequences where a given event is surprisingly frequent or infrequent
- Method:
  - Find all subsequences of a given length
  - Control family-wise error rate
  - Analytical approximation
  - O(1) complexity per subsequence