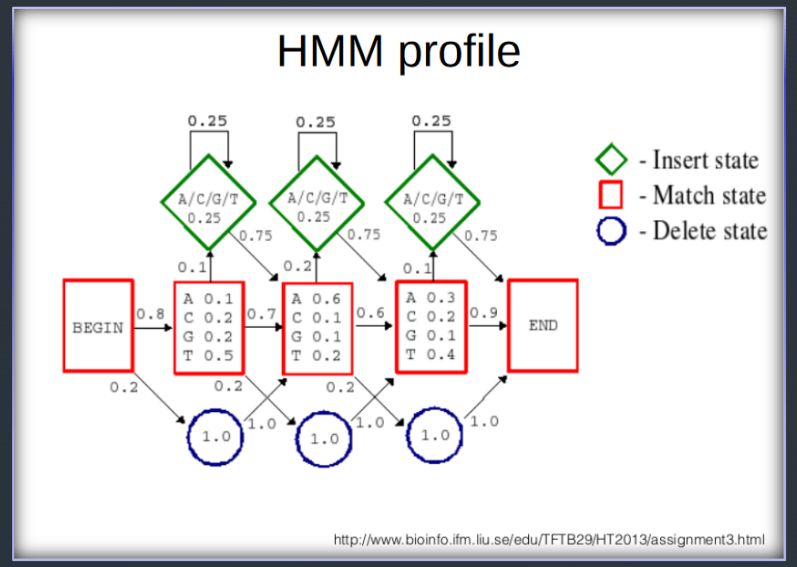
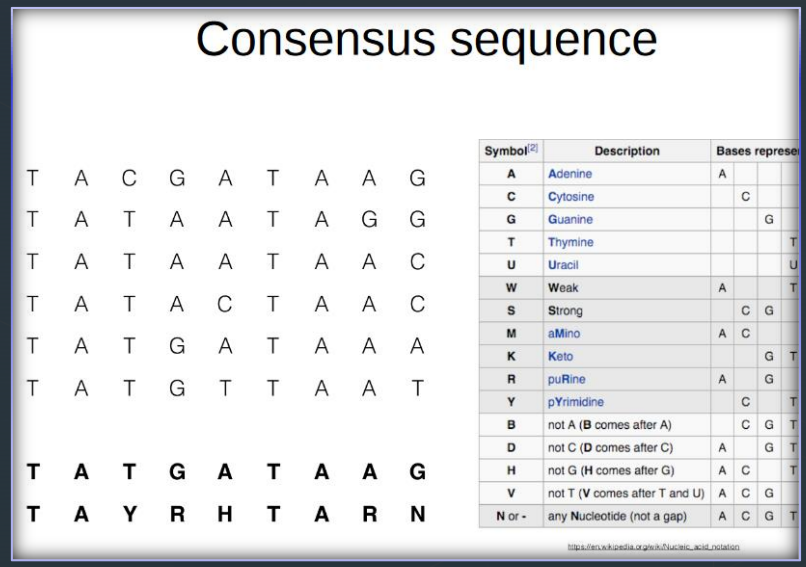
Keishin Nishida, Martin C. Frith and Kenta Nakai

Pseudocounts for transcription factor binding sites

DNA Motifs

- A pattern of nucleotide sequences.
- Standard, palindromes and gapped.
- Associated to DNA-protein binding sites.





Positional Matrices

Set of sequences

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

Position Frequency Matrix (PFM)

	1	2	3	4	5	6	7	8	9
A	0	6	0	3	4	0	6	5	1
C	0	0	1	0	1	0	0	0	2
G	0	0	0	3	0	0	0	1	2
T	6	0	5	0	1	6	0	0	1

Position Probability Matrix (PPM)

	1	2	3	4	5	6	7	8	9
A	0	1	0	0.5	0.67	0	1	0.83	0.17
C	0	0	0.17	0	0.17	0	0	0	0.34
G	0	0	0	0.5	0	0	0	0.17	0.34
T	1	0	0.83	0	0.17	1	0	0	0.17

$$M_{k,j} = \frac{1}{N}\sum_{i=1}^{N} I(X_{i,j} = k)$$

Position Weight Matrix (PWM)

	1	2	3	4	5	6	7	8	9
A	-inf	1.38	-inf	0.69	0.99	-inf	1.38	1.20	-0.39
C	-inf	-inf	-0.39	-inf	-0.39	-inf	-inf	-inf	0.31
G	-inf	-inf	-inf	0.69	-inf	-inf	-inf	-0.39	0.31
T	1.38	-inf	1.20	-inf	-0.39	1.38	-inf	-inf	-0.39

$$M_{k,j} = \ln\ (M_{k,j}/b_k)$$

S: TAT**A**ATAAT Score = 1*1*0.83*0.5*0.67*1*1*0.83*0.17= 0.03

S: TAT**C**ATAAT Score = 0

Motif matching with pseudocounts

	1	2	3	4	5	6	7	8	9
A	1	7	1	4	5	1	7	6	2
C	1	1	2	1	2	1	1	1	3
G	1	1	1	4	1	1	1	2	3
T	7	1	6	1	2	7	1	1	2

S: TAT**A**ATAAT Score = 0.7*0.7*0.6*0.4*0.5*0.7*0.7*0.6*0.2= 0.00345

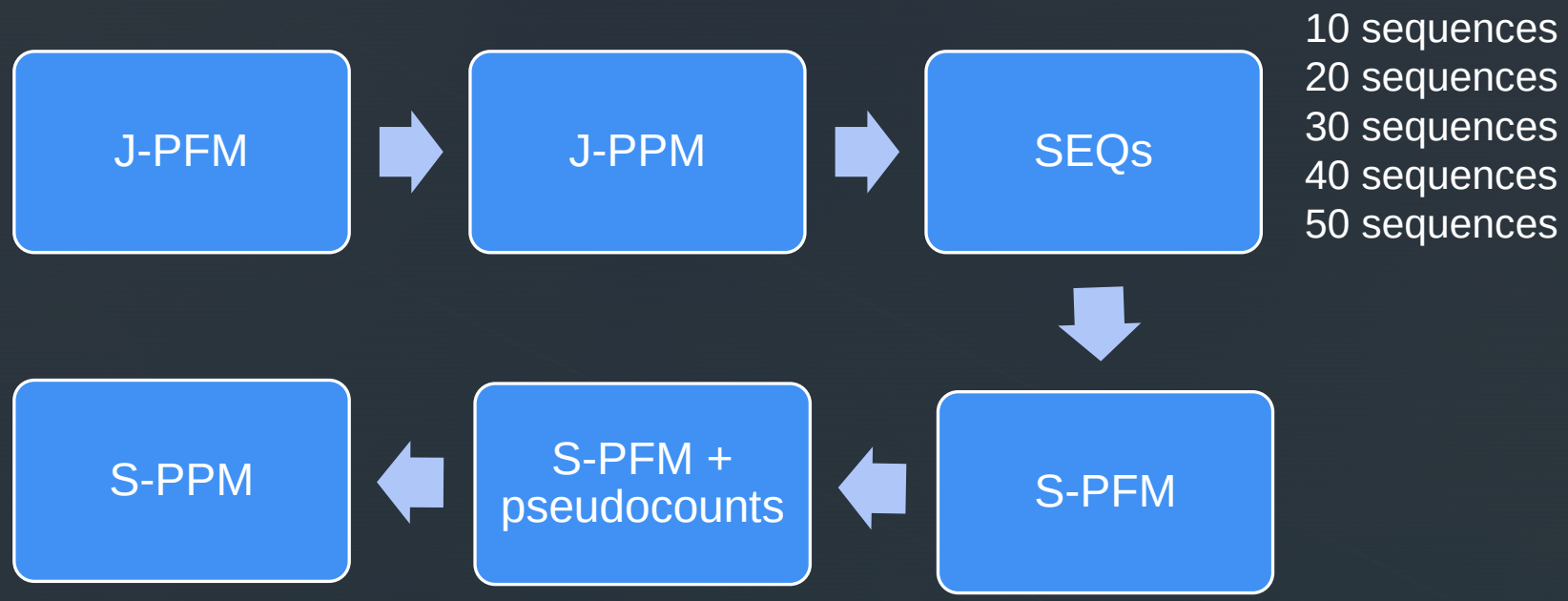
S: TAT**C**ATAAT Score = 0.7*0.7*0.6*0.1*0.5*0.7*0.7*0.6*0.2= 0.00086

MATERIALS AND METHODS

JASPAR dataset

- JASPAR 2008
- Motifs for multicellular Eukaryotes
- 122 PFM

Sampled PFM from original PFM

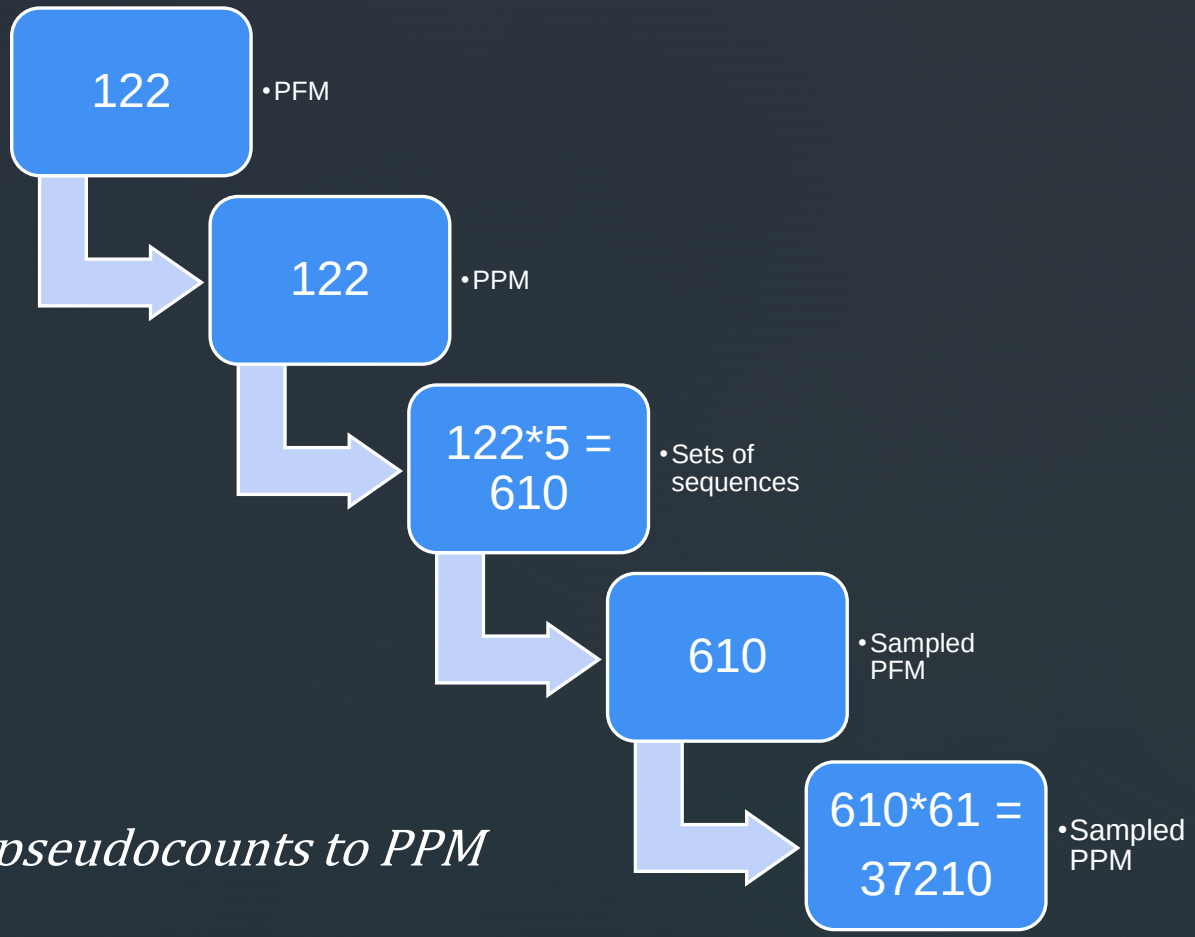


Pseudocount addition

- $p'_{a,i} = \frac{c'_{a,i}+B/4}{m+B}$
- For every sample PFM built up to 61 new “sample” PPM with pseudocount B have been created.
- Pseudocounts exponentially bigger. $B = 10^{\left(\frac{x}{y}\right)-2}$

Operations recap

- *From PFM to PPM*
- *From PPM to set of sequences*
- *From set of sequences to PFM*
- *From PFM with the addition of pseudocounts to PPM*



Comparison procedures

Seven methods divided in two categories

Matrix-based comparison:

- Compare JASPAR PPM matrices and Sampled PPM matrices.

Sequence-based comparison:

- Enumeration of all possible *w-mers**.
- Compute the probability *s* that a PPM generates a *w*-mers.
- $s = \prod_i p_{a_i,i}$
- Compare the probabilities of the sequences

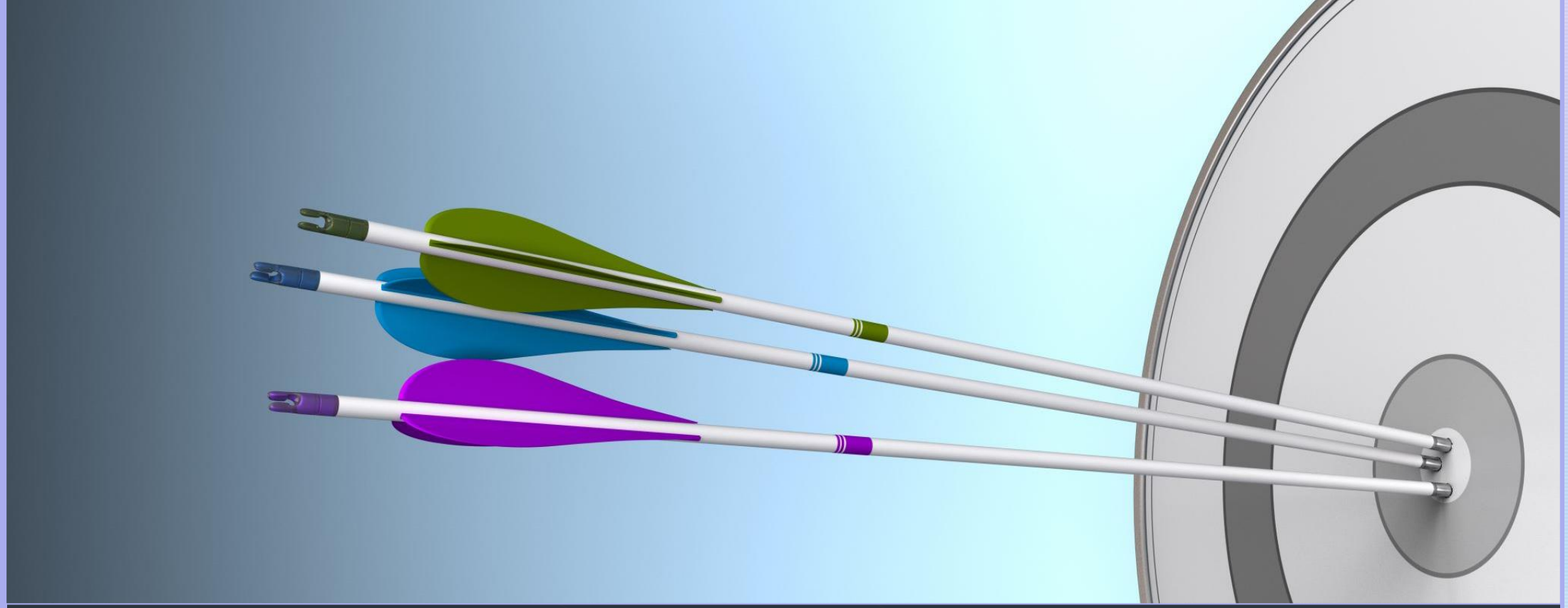
Comparison functions

Matrix-based comparison:

- Euclidian distance (ED)
- Cosine distance (COS)
- Total variation (TVD)

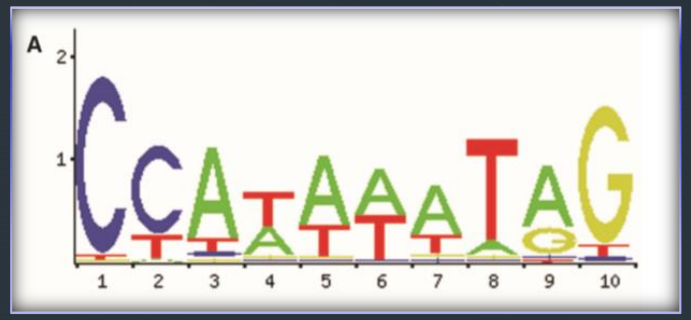
Sequence-based comparison:

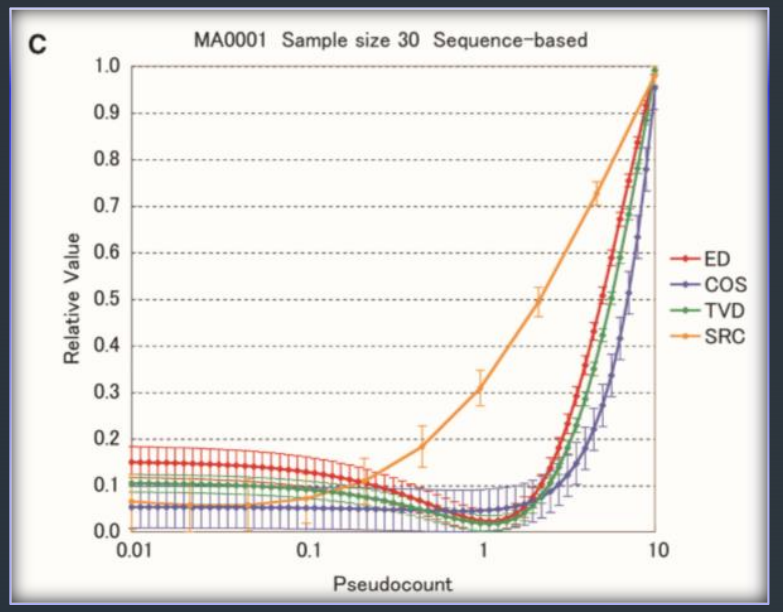
- Euclidian distance (ED)
- Cosine distance (COS)
- Total variation (TVD)
- Spearman’s Rank Correlation (SRC)*

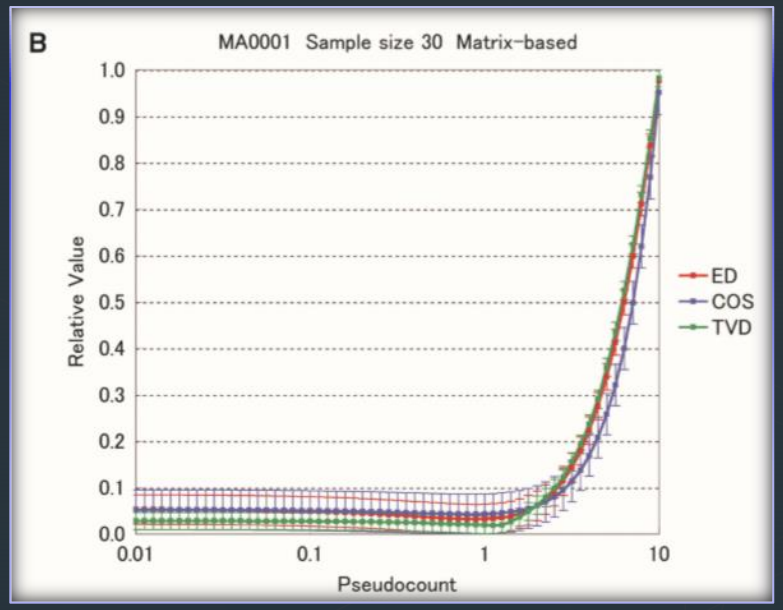


RESULTS

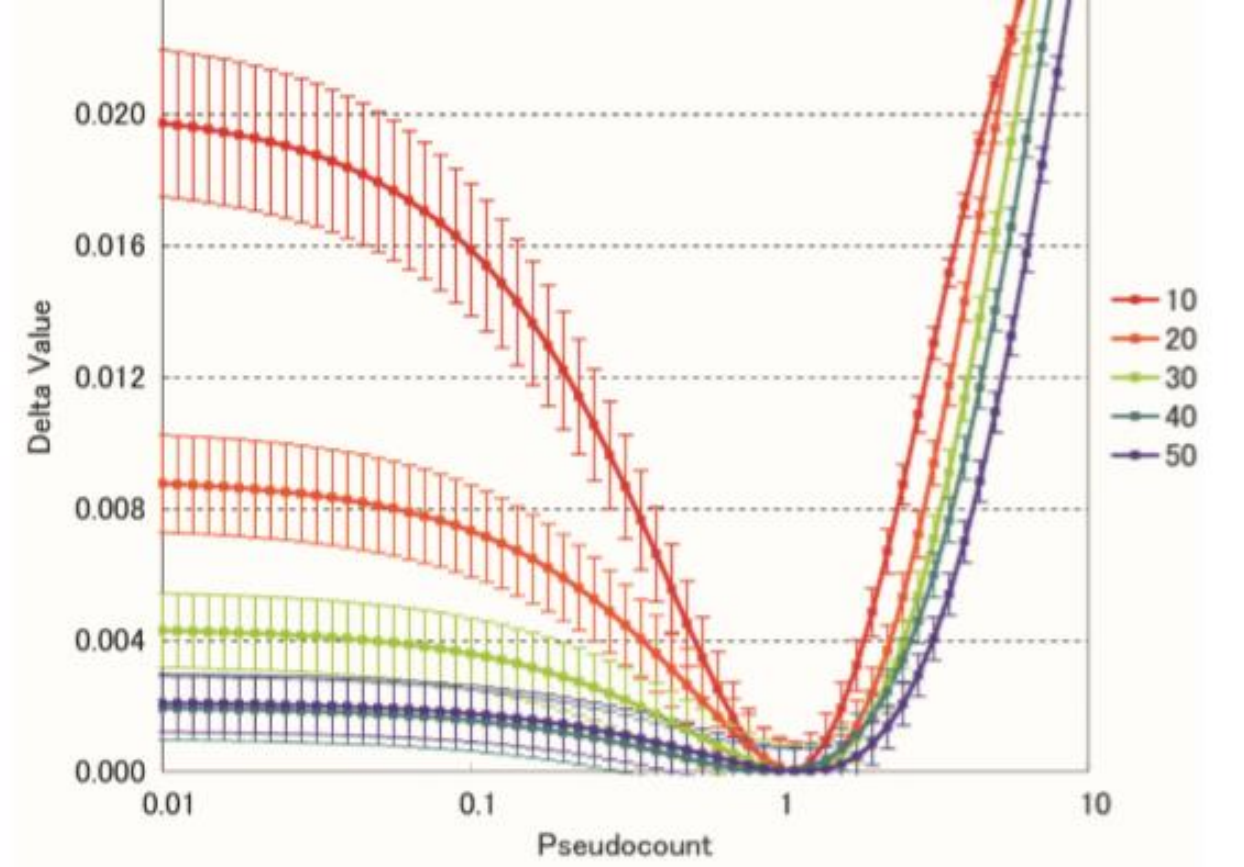
Differences between the seven comparison methods

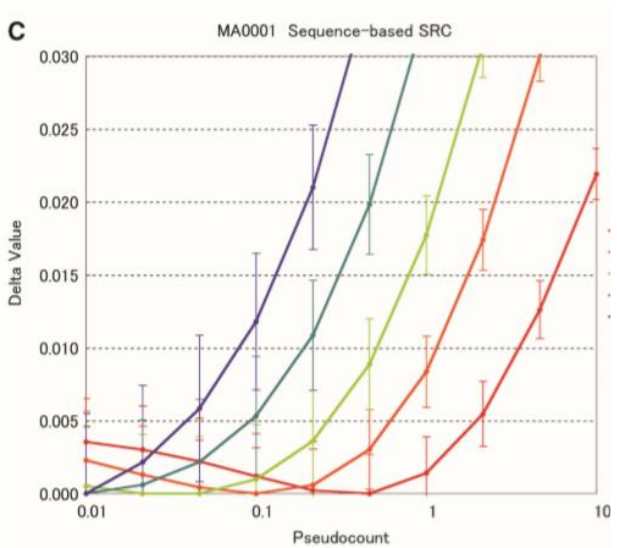






Effects on sample size





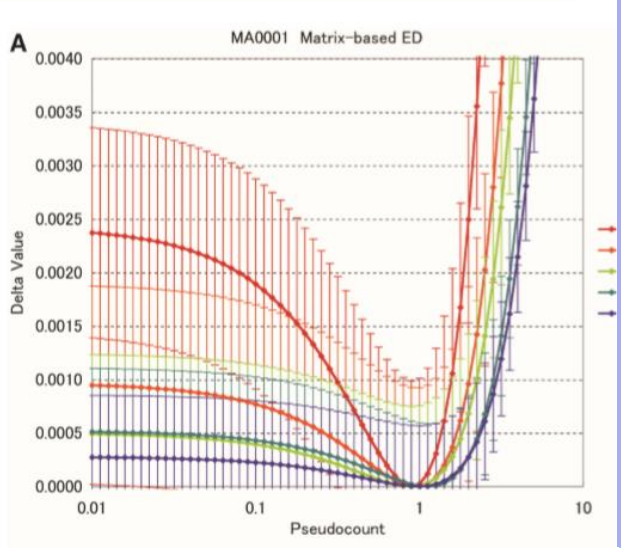
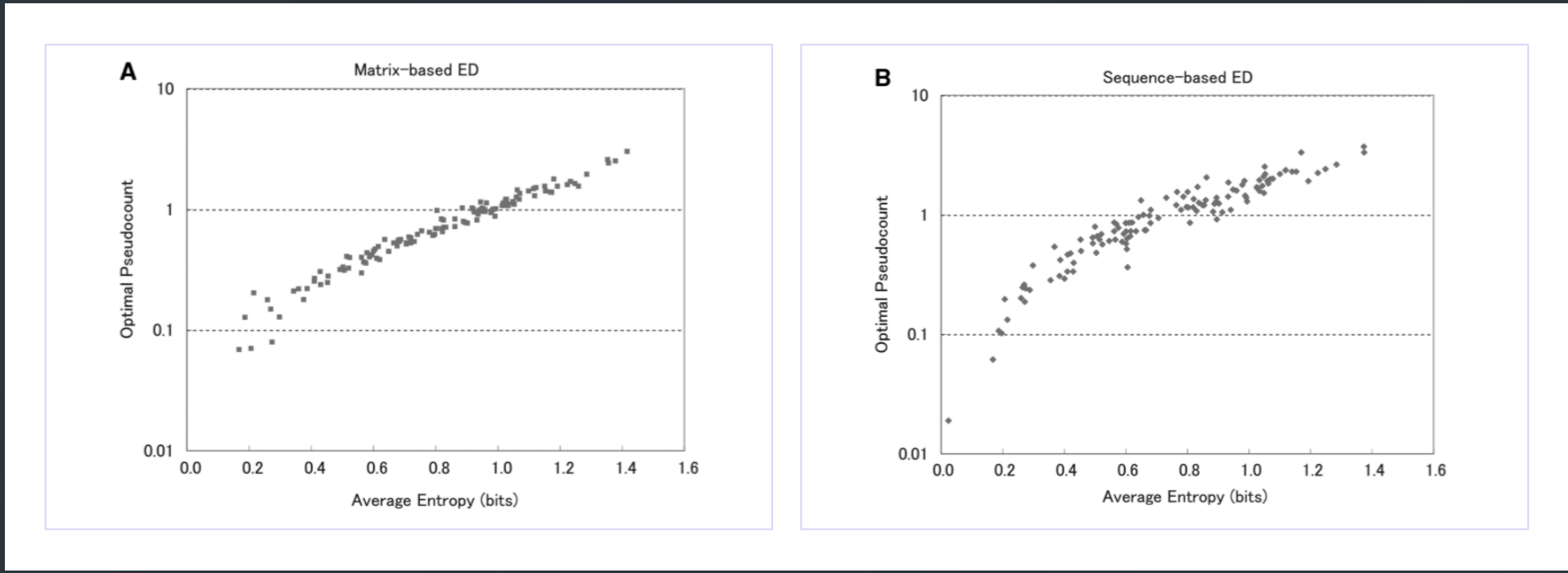


Table 1. Percentage of optimal pseudocount existence

Sample size	Matrix-based			Sequence-based			
	ED (%)	COS (%)	TVD (%)	ED (%)	COS (%)	TVD (%)	SRC (%)
10	99.2	100.0	28.7	100.0	95.9	54.1	45.9
20	99.2	100.0	16.4	99.2	95.9	26.2	34.4
30	100.0	100.0	13.1	100.0	94.3	13.1	20.5
40	100.0	100.0	9.0	99.2	95.1	11.5	14.8
50	100.0	99.2	10.7	99.2	95.1	9.0	5.7

Existence of an optimal pseudocount



Dependence of optimal pseudocount on sample size and entropy

CONCLUSIONS

Conclusions

- All pseudocounts much above 1 are a poor choice
- ED and COS comparison functions suggest values close to 1
- SRC and TVD comparison functions suggest values much smaller than 1
- Depending on the comparison method, pseudocounts are either around 1 or very low.

Thank you for the attention

Francesco Penasa

Comparison functions

Euclidian distance (ED)

$$ED = \frac{1}{w}\sqrt{\sum_{a}\sum_{i}(p_{a,i} - p'_{a,i})^2}$$

Cosine distance (COS)

Total variation (TVD)