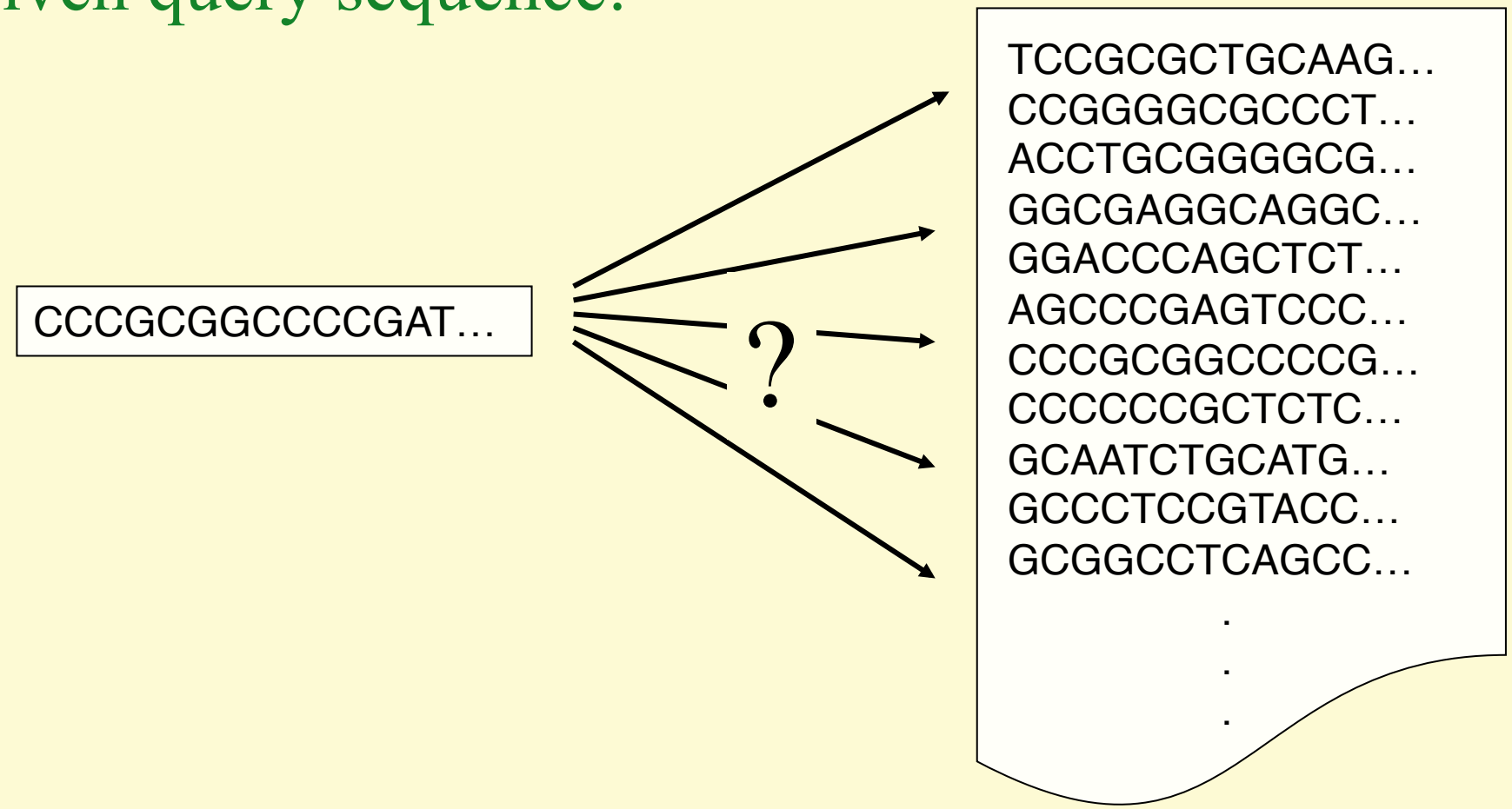
Alignment-free sequence comparisons using *k*-word matches

Sue Wilson
Hilary Booth
Ruth Kantorovitz
Conrad Burden
Sylvain Forêt
Junmei Jing

A common problem in biology is sequence matching: finding a DNA sequence or a protein sequence in a data base that is a ‘close’ match to a given query sequence:



- Used, for example, to identify homologous genes or proteins in one species or genes related by a common ancestor in different species
- Don’t just want to know whether sequences are related, but need a measure of similarity (or dissimilarity)
- Assign a p-value based on a null hypothesis that two sequences being compared are unrelated
- Simplest null hypothesis is that the sequences are strings of independently and identically distributed letters from a given alphabet

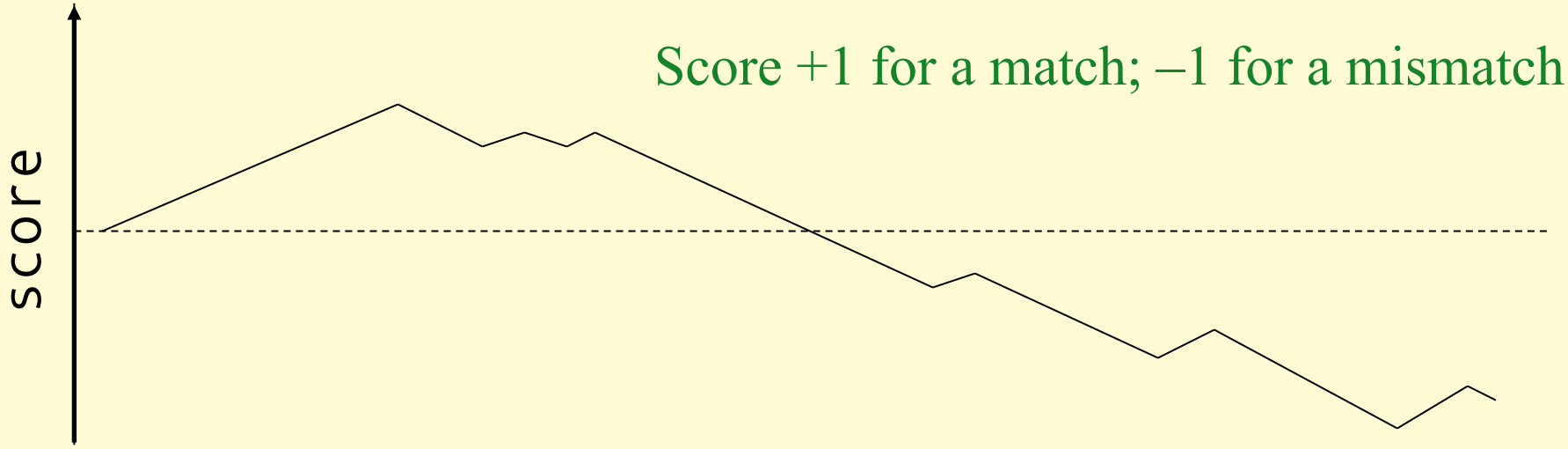
In biology, the most commonly used algorithms for sequence comparison are alignment-based algorithms, e.g. BLAST.

(Basic Local Alignment Search Tool)

BLAST looks for long alignments and relies on the theory of random walks:

ATGCTTTGCTAGCGCTAGCATGCTTTCGCAAACTCAT

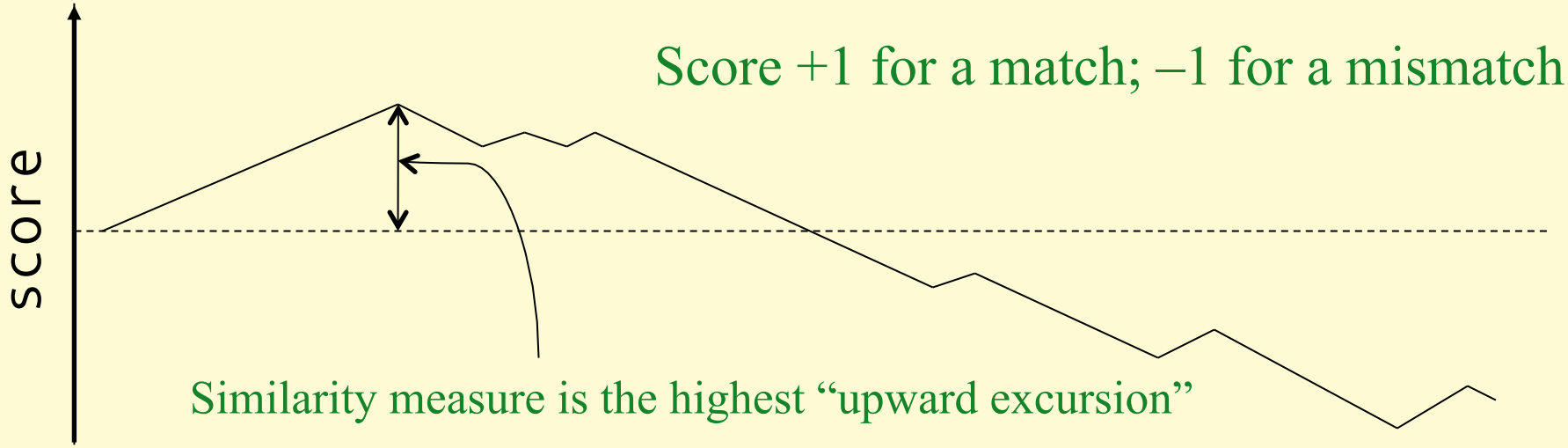
ATGCTTTTTAAAACCGAGCTGGTCACAAGCGCTAACAA



BLAST looks for long alignments and relies on the theory of random walks:

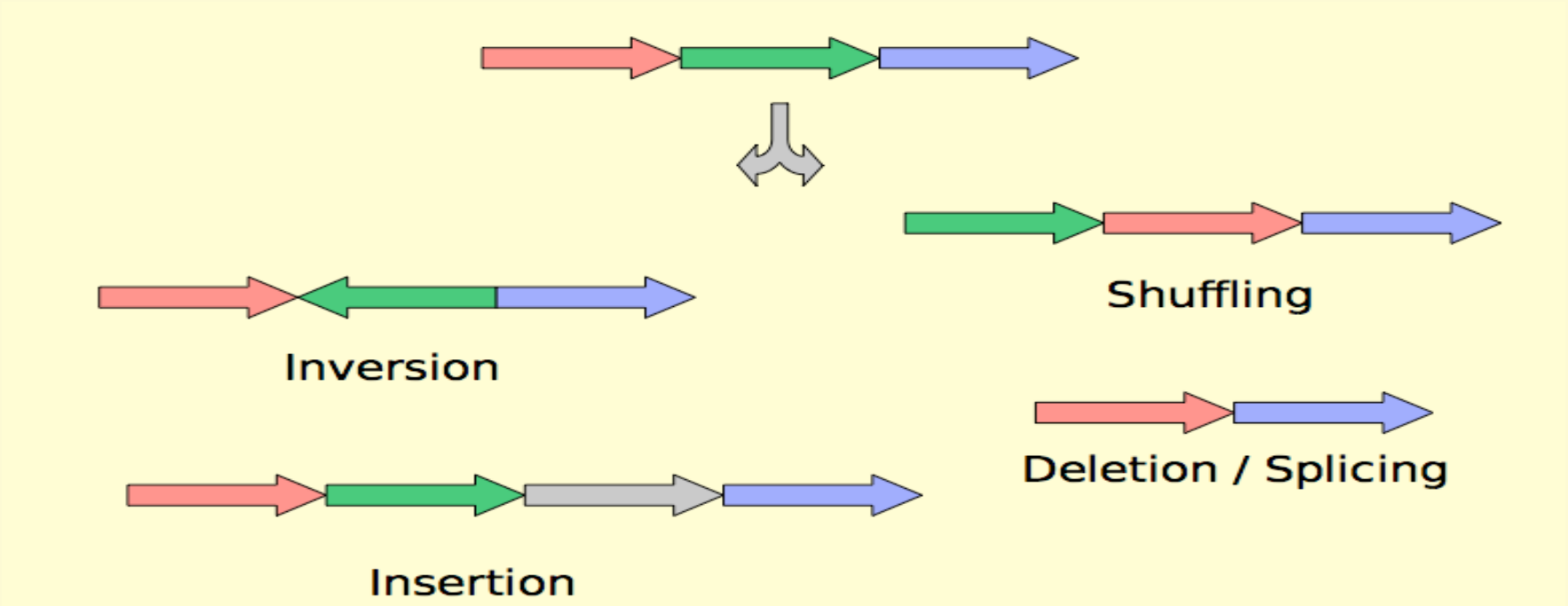
ATGCTTTGCTAGCGCTAGCATGCTTTCGCAAACTCAT

ATGCTTTTAAAACCGAGCTGGTCACAAGCGCTAACAA



Large upward excursion = ‘good’ alignment

Alignment-based sequence comparisons assume contiguity between related sequences. But the process of evolution may involve rearrangements of sections of the genome, and the process of translating genes to proteins may involve alternate deletions and splicings



The assumption of contiguity is not always appropriate!

Alignment-free methods:

There are many (distance methods, covariance methods, information theory based measures, angle metrics, …)

We have been studying

k-word matches and the D_2 statistic

Definition: Given two sequences

$\mathbf{A} = (A_1, A_2, \ldots, A_m)$ and $\mathbf{B} = (B_1, B_2, \ldots, B_n)$,

D_2 is the number of matches of words (including overlaps) of prespecified length k between two given sequences

Definition: Given two sequences

$\mathbf{A} = (A_1, A_2, \ldots, A_m)$ and $\mathbf{B} = (B_1, B_2, \ldots, B_n)$,

D_2 is the number of matches of words (including overlaps) of prespecified length k between two given sequences

Example: consider these two sequences and $k = 7$ …

A: ATGCTTTGCTAGCGCTAGCATGCTTTCGCAAACTCAT

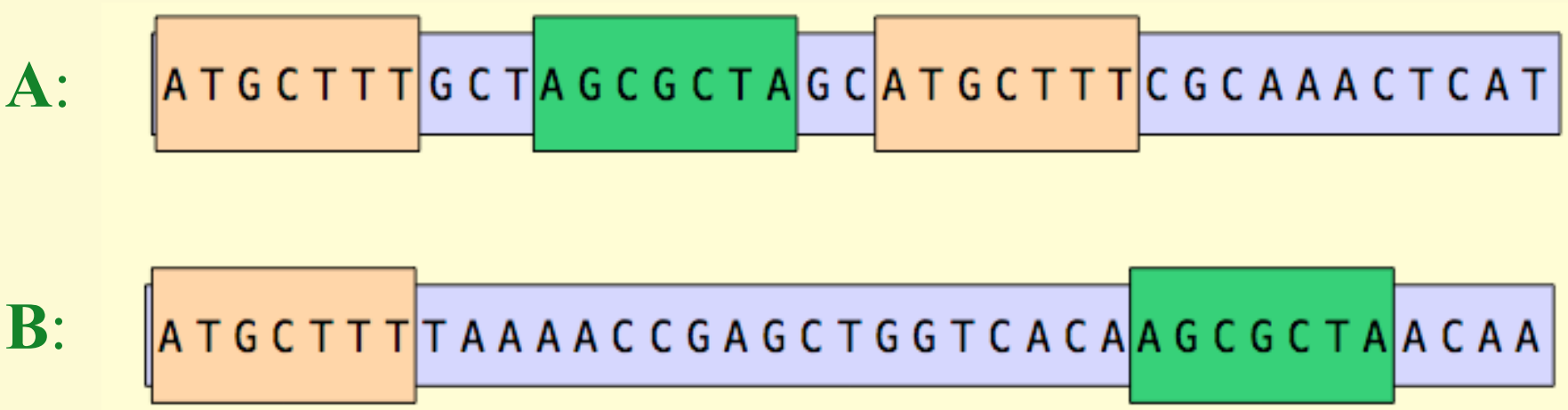
B: ATGCTTTTAAAACCGAGCTGGTCACAAGCGCTAACAA

Definition: Given two sequences

$\mathbf{A} = (A_1, A_2, \ldots, A_m)$ and $\mathbf{B} = (B_1, B_2, \ldots, B_n)$,

D_2 is the number of matches of words (including overlaps) of prespecified length k between two given sequences

Example: consider these two sequences and $k = 7$ …

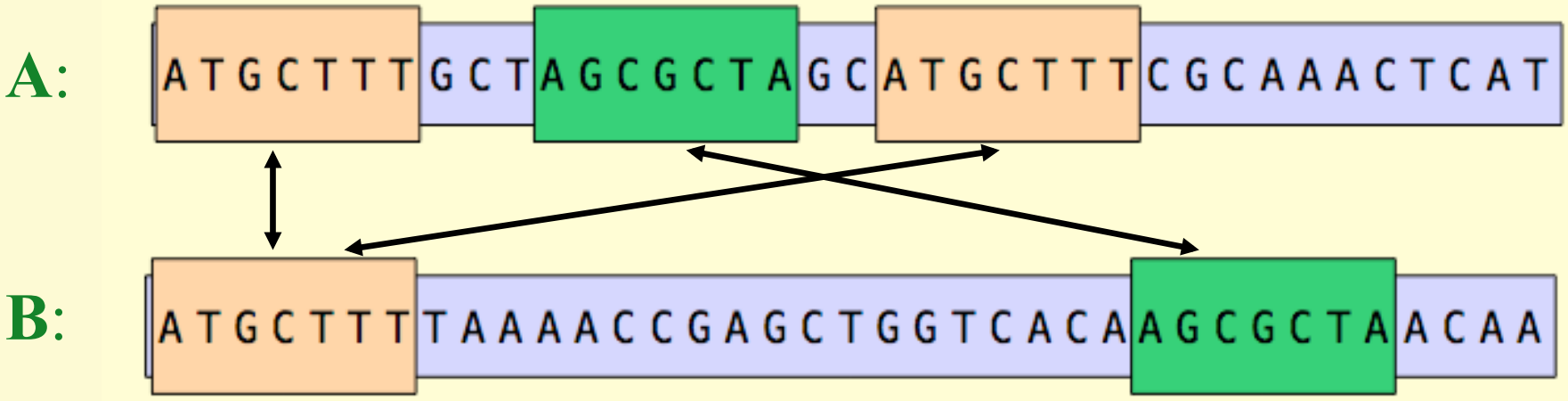


Definition: Given two sequences

$\mathbf{A} = (A_1, A_2, \ldots, A_m)$ and $\mathbf{B} = (B_1, B_2, \ldots, B_n)$,

D_2 is the number of matches of words (including overlaps) of prespecified length k between two given sequences

Example: consider these two sequences and $k = 7$ …



In this example, for $k = 7$, $D_2 = 3$

Also of interest is the approximate word count:

Definition: Given two sequences

$\mathbf{A} = (A_1, A_2, \ldots, A_m)$ and $\mathbf{B} = (B_1, B_2, \ldots, B_n)$,

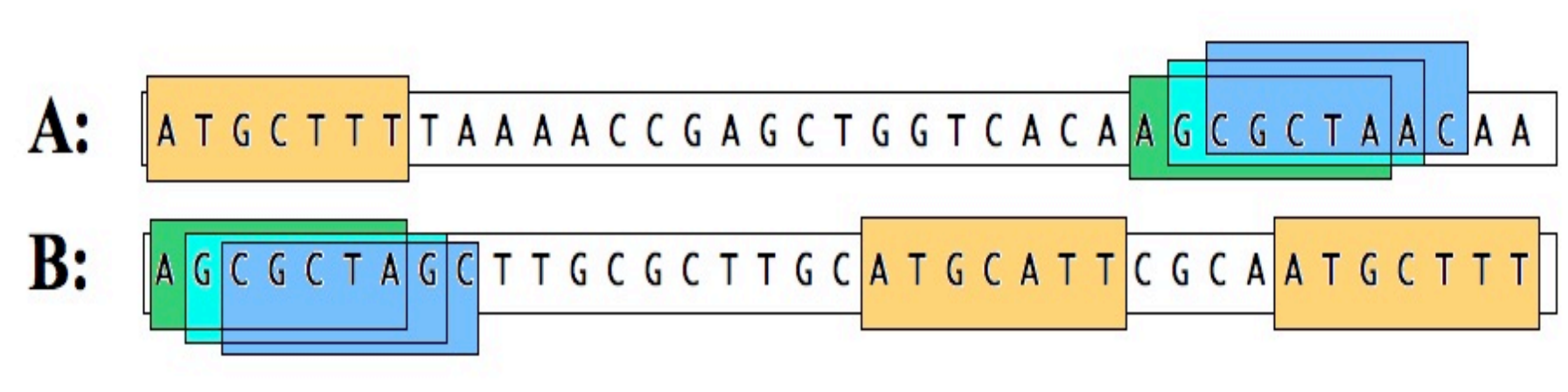
$D_2^{(t)}$ is the number of matches of words of prespecified length k with up to t mismatches

Also of interest is the approximate word count:

Definition: Given two sequences

$\mathbf{A} = (A_1, A_2, \ldots, A_m)$ and $\mathbf{B} = (B_1, B_2, \ldots, B_n)$,

$D_2^{(t)}$ is the number of matches of words of prespecified length k with up to t mismatches



In this example, for $k = 7$, $t = 1$, $D_2 = 5$

Performance:

For sequences of length m and n,

- D_2 has algorithmic complexity $O(m + n)$ …fast!
- $D_2^{(t)}$ is at worst $O(m*n)$ … somewhat slower

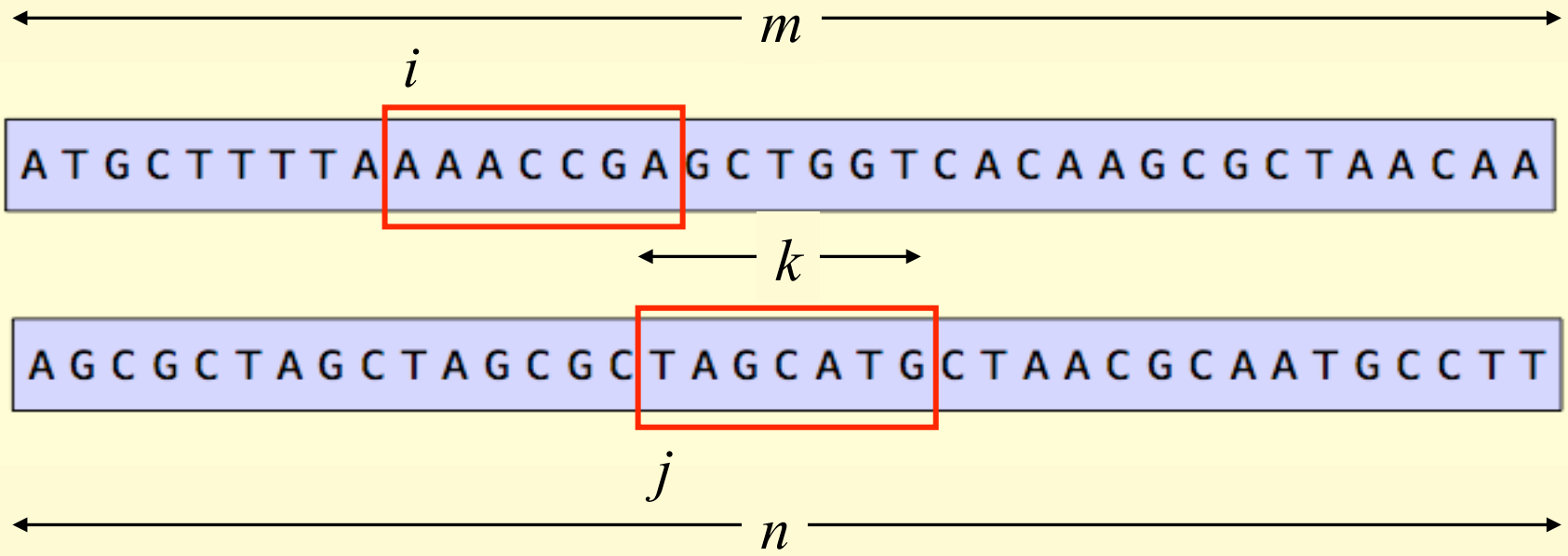
But to assess whether a match is significant, we need knowledge of the distribution of these measures under a suitable null hypothesis

What do we know about the distribution of D_2 and $D_2^{(t)}$?

Although we do not have an exact formula for the distribution of D_2, we are able to derive for sequences of i.i.d. letters

- The means $E(D_2)$ and $E(D_2^{(t)})$
- The variance $Var(D_2)$ and, for a uniform letter distribution $Var(D_2^{(t)})$
- A fast, accurate numerical algorithm for $Var(D_2^{(t)})$ for a non-uniform letter distribution
- The limiting distribution of D_2 as the sequence length $n \to \infty$ for $k < 1/2 \log n$ or $k > 2 \log n$
- The limiting distribution of $D_2^{(t)}$ as the sequence length $n \to \infty$ for $k < 1/2 \log n$
- An accurate empirical fit to the distribution for bilogically relevant values of n, k and t.

Mean of D_2 for i.i.d. sequences (Waterman, 1995):



Let probability of letter at given site be f_a, $a \in \{C, A, G, T\}$

Set indicator variable $Y_{ij} = 1$ if k-word at i matches k-word at j,
0 otherwise

Then $E(Y_{ij}) = \text{Prob}(Y_{ij} = 1) = (\sum_a f_a^2)^k$, so

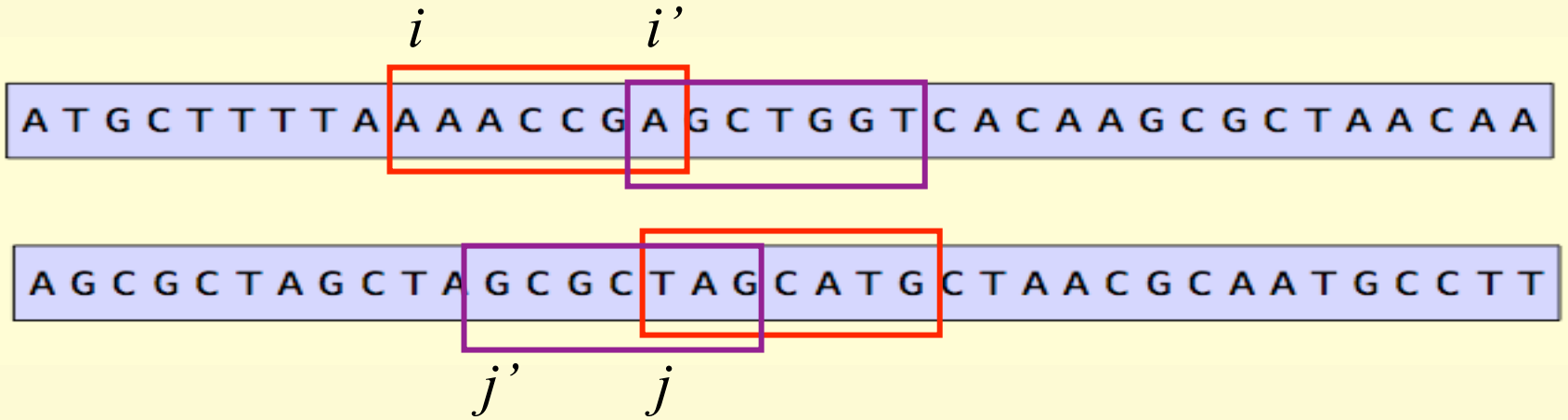
$E(D_2) = E(\sum_{i,j} Y_{ij}) = \sum_{i,j} E(Y_{ij}) = (m - k + 1)(n - k + 1) (\sum_a f_a^2)^k$

The variance of D_2 is much harder:

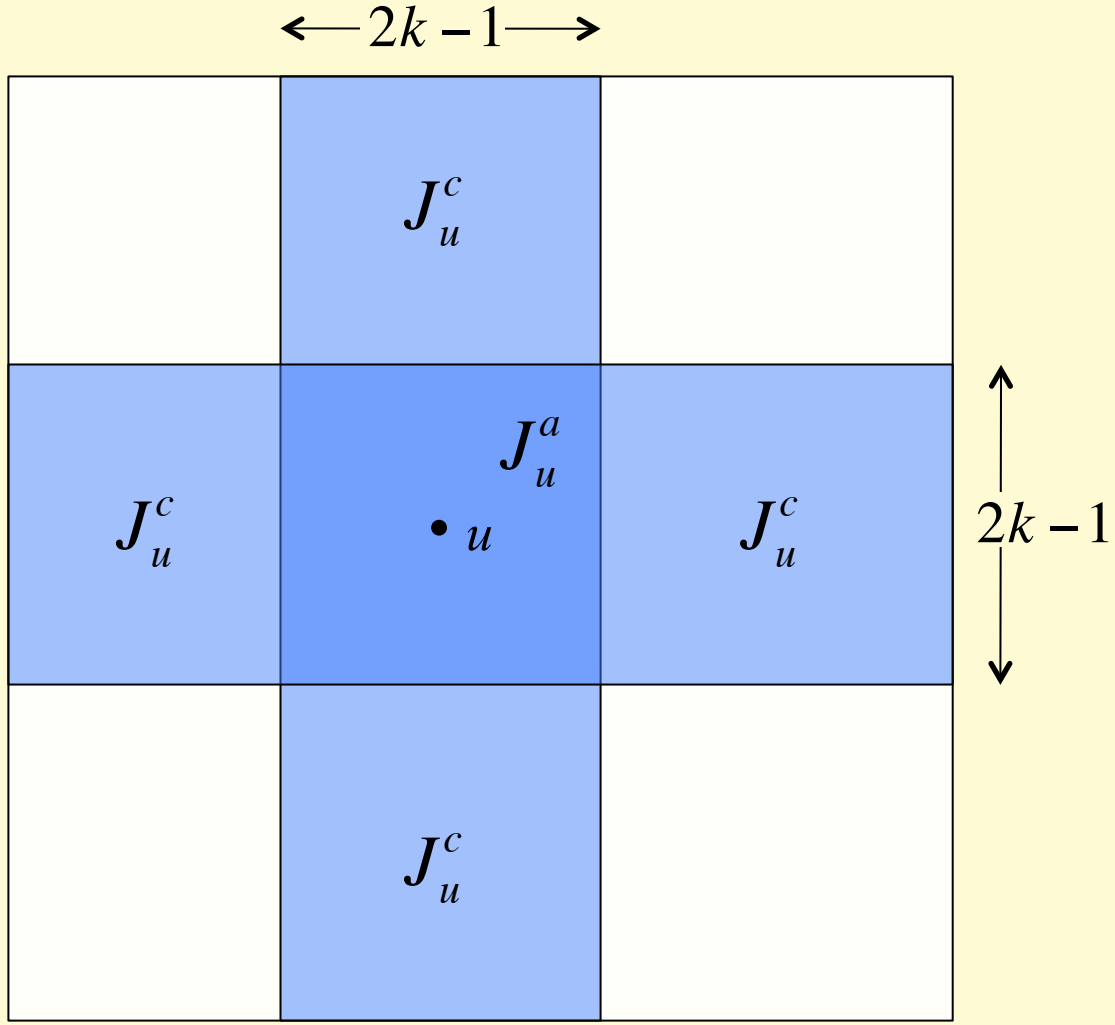
$$\mathrm{Var}(D_2) = \mathrm{Var}\left(\sum_{i,j} Y_{ij}\right) = \mathrm{E}\left(\left(\sum_{i,j} Y_{ij}\right)^2\right) - \left(\mathrm{E}\left(\sum_{i,j} Y_{ij}\right)\right)^2$$

$$= \sum_{i,j;i',j'} \mathrm{Cov}\left(Y_{ij}, Y_{i'j'}\right) - {\mu_{D_2}}^2$$

but $\mathrm{Cov}(Y_{ij}, Y_{i'j'})$ is difficult to calculate when there are overlaps

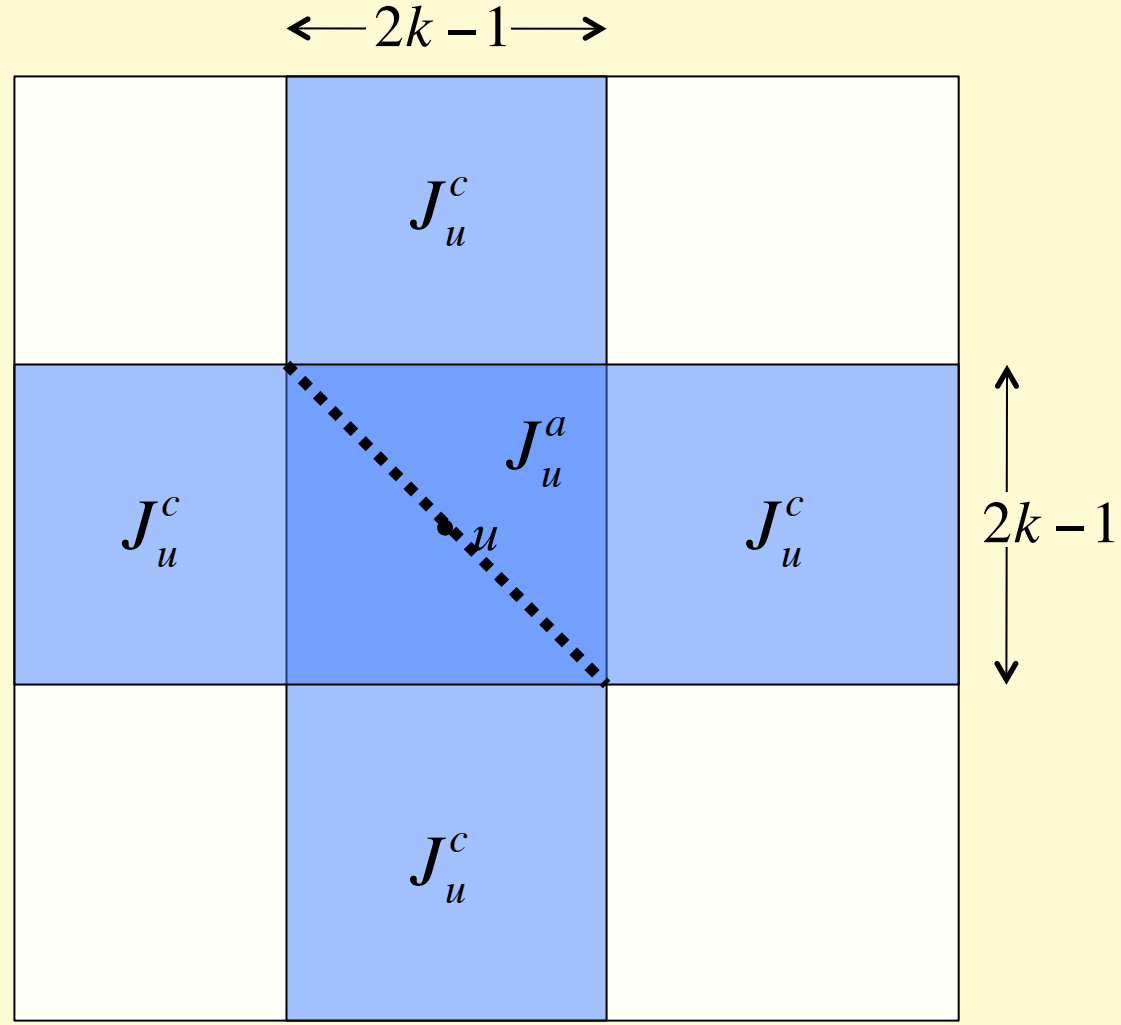


To calculate $\mathrm{Cov}\left(Y_{ij}, Y_{i'j'}\right) = \mathrm{Cov}\left(Y_u, Y_v\right)$, where $u = (i,j)$, $v = (i',j')$, write the dependency neighbourhood as $J_u = J_u^a + J_u^c$



("Accordion"
+
"Crabgrass")

To calculate $\mathrm{Cov}\left(Y_{ij}, Y_{i'j'}\right) = \mathrm{Cov}\left(Y_u, Y_v\right)$, where $u = (i,j)$, $v = (i',j')$, write the dependency neighbourhood as $J_u = J_u^a + J_u^c$



("Accordion"
+
"Crabgrass")

And further divide the accordion into diagonal and off-diagonal pieces

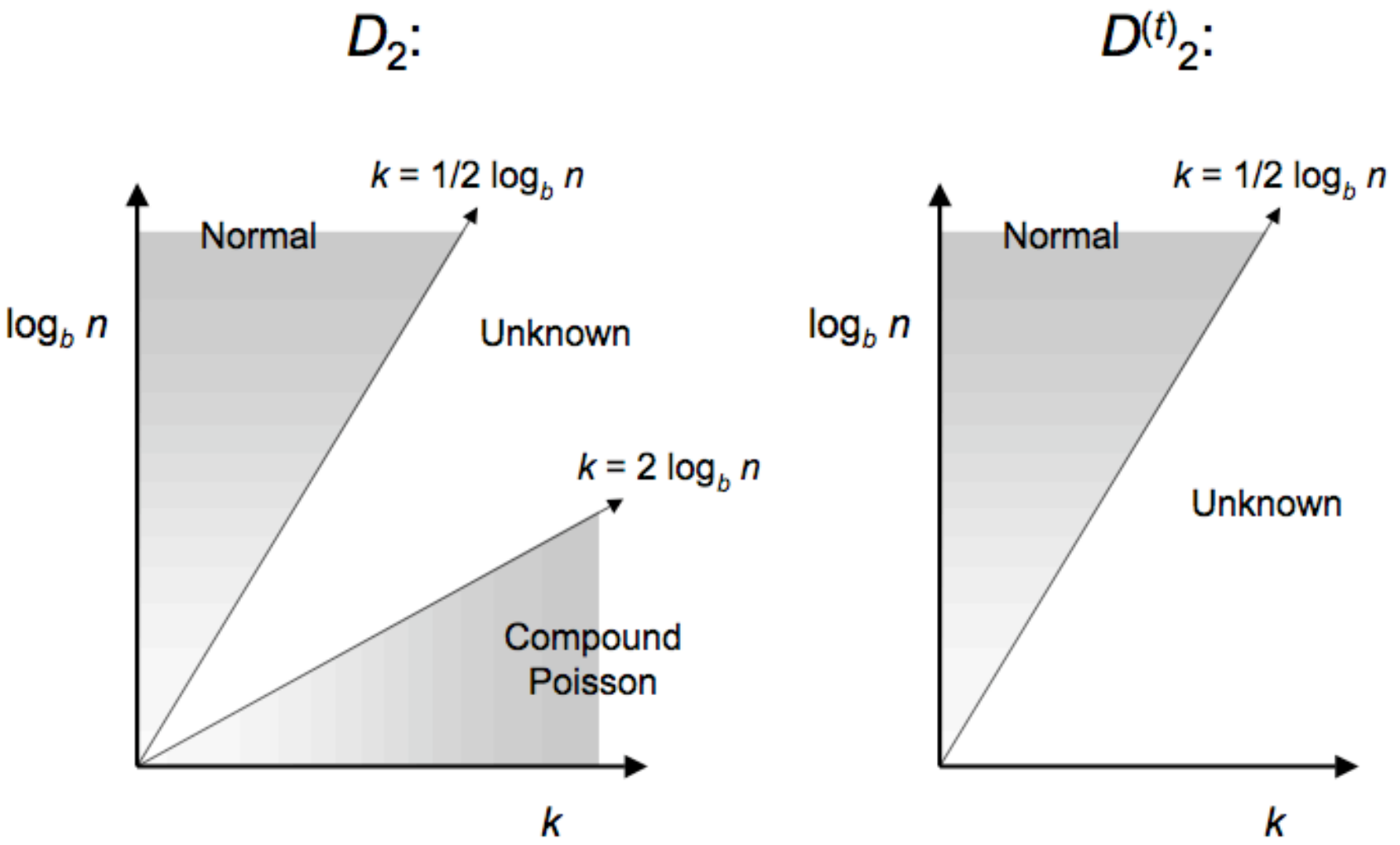
For practical purposes, we can calculate all contributions to the covariance:

	Crabgrass	**Diagonal Accordion**	**Off-diag Accordion**
Exact matches, Uniform letter distribution	0	Analytic formula	0
Exact matches, Non-uniform distribution	Analytic formula	Analytic formula	Analytic formula
Approx. matches, Uniform letter distribution	0	Analytic formula	0
Approx. matches, Non-uniform distribution	Analytic formula	Numerical look-up table in parameters k, t and f_a	

So we can calculate the mean and variance of D_2 and $D_2^{(t)}$ for any set of parameters n, m, k, t and f_a

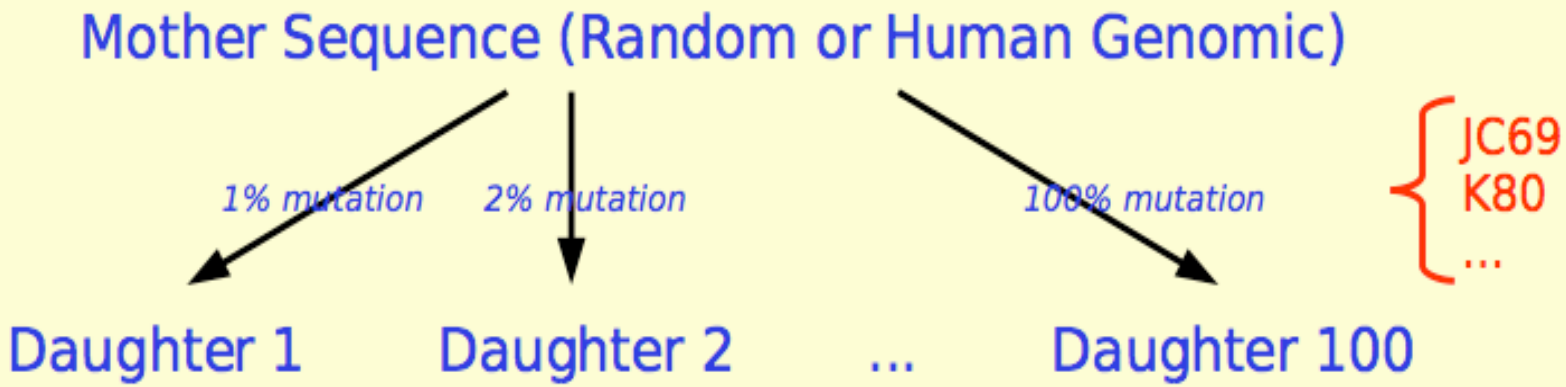
But what about the shape of the distribution

Have proved theorems for the limiting distributions as n or $k \to \infty$:



(Limits taken along lines $k = \text{const.} \times \log_b n$, where $b = (\sum_a f_a^2)^{-1}$)

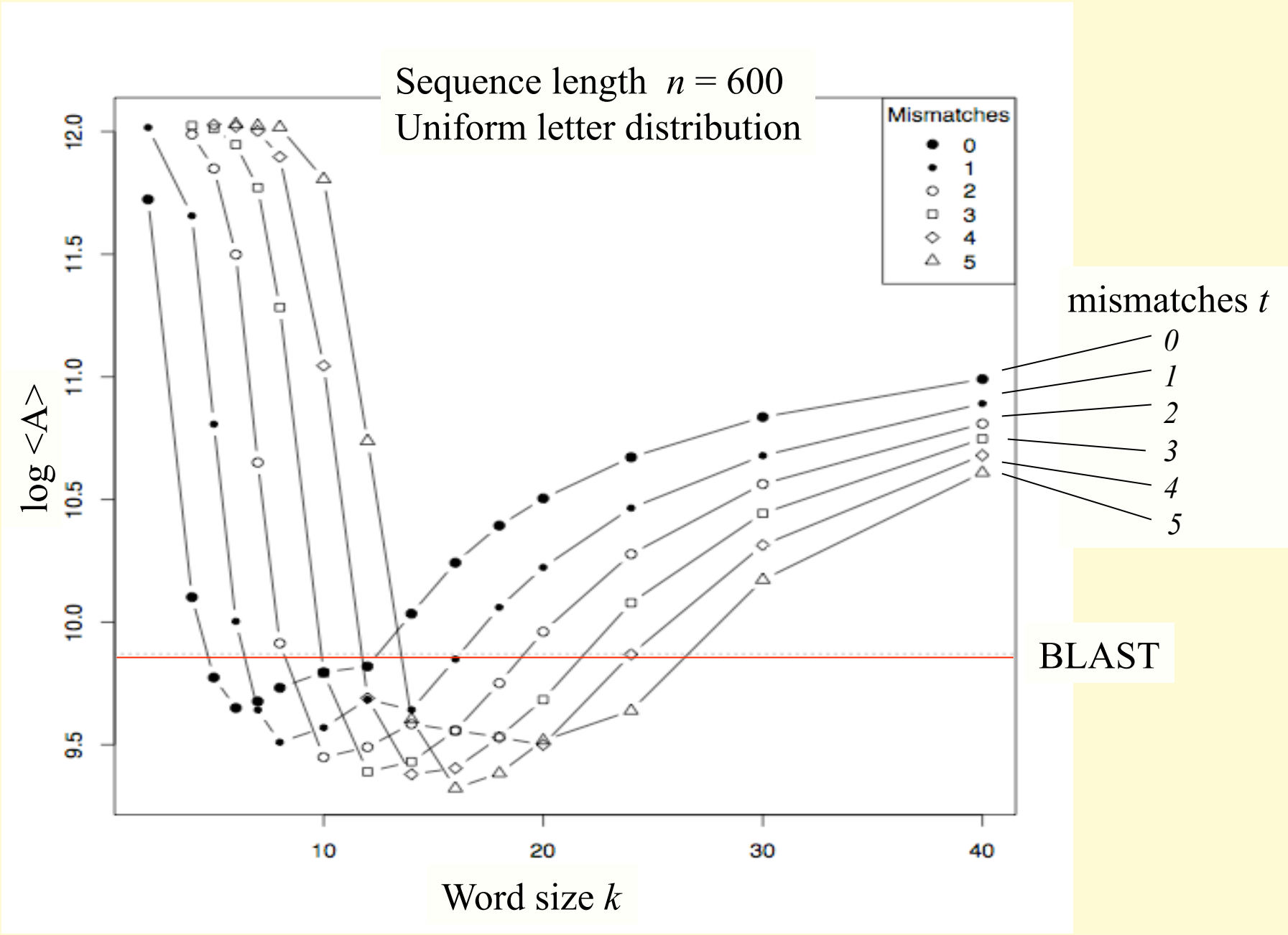
Numerical experiments to determine optimal word size k



Ranking of the daughters according to their D_2 relative to the mother sequence to get an inferred ranking $r(\gamma)$, $\gamma = 1, \ldots, 100$

Spearman's Rank Statistic

$$A = \sum_{\gamma=1}^{100} \left(r(\gamma) - \gamma \right)^2$$


Sequence length $n = 600$
Uniform letter distribution
Mismatches
0
1
2
3
4
5
log <A>
12.0
11.5
11.0
10.5
10.0
9.5
10
20
30
40
Word size k
mismatches t
0
1
2
3
4
5
BLAST

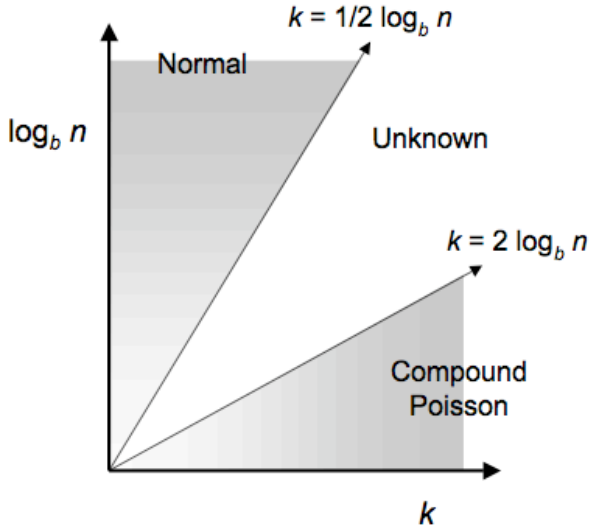
Optimum word sizes

Mismatches	Sequence Lengths				
	200	400	800	1600	3200
0	6	7	7	7	7
1	8	10	10	10	10
2	10	12	12	12	12
3	12	14	14	14	14
4	14	16	16	16	16
5	16	18	18	18	18

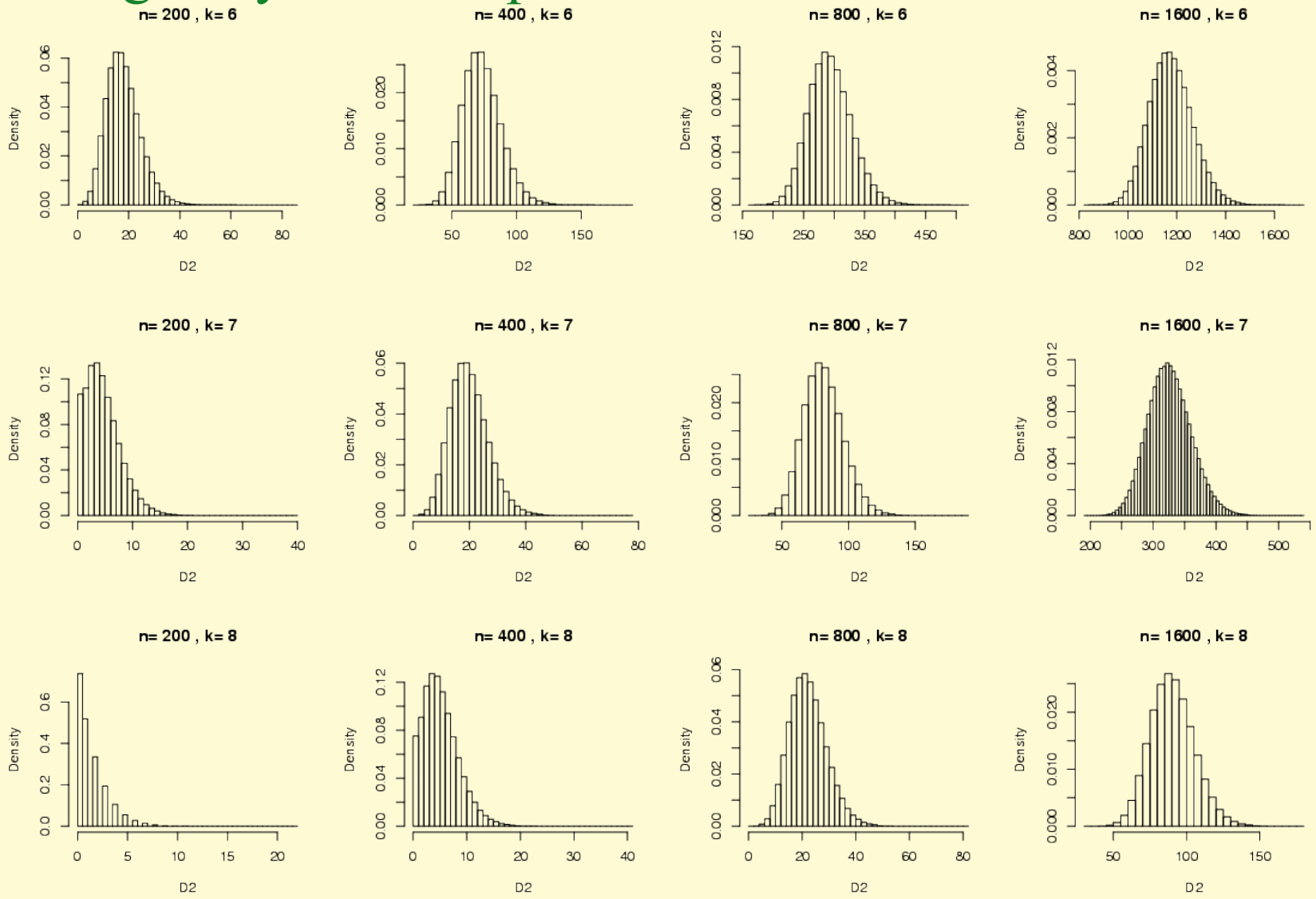
Optimum word sizes

Mismatches	Sequence Lengths				
	200	400	800	1600	3200
0	6	7	7	7	7
1	8	10	10	10	10
2	10	12	12	12	12
3	12	14	14	14	14
4	14	16	16	16	16
5	16	18	18	18	18

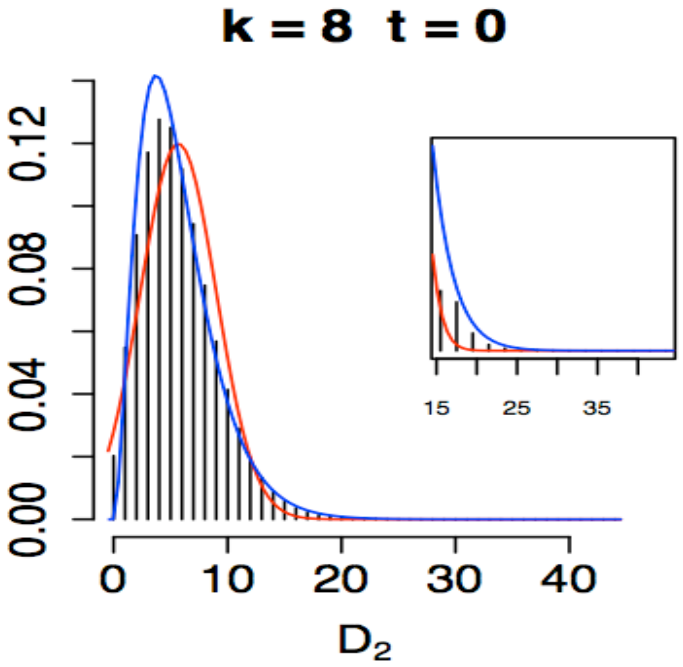
For biologically relevent parameter values, optimal word sizes fall outside known limiting cases

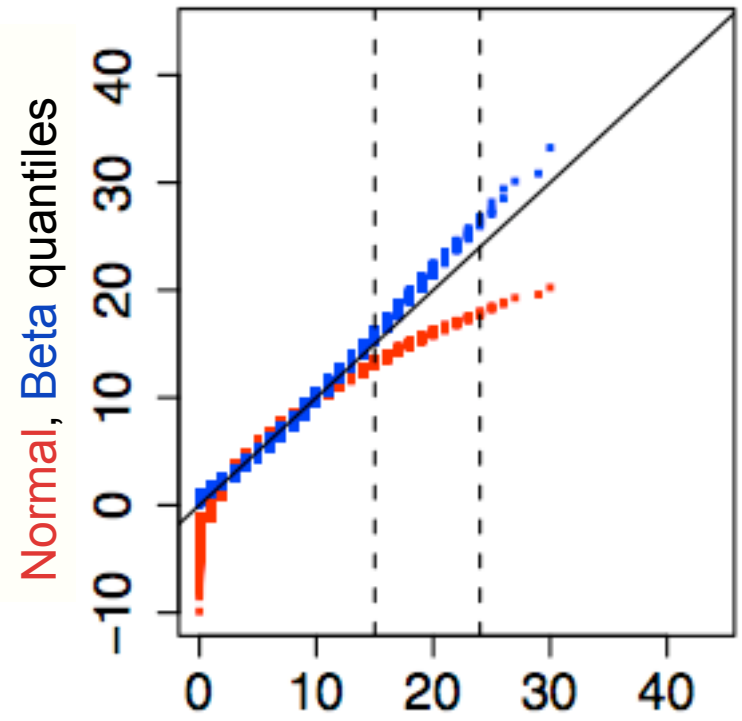


→ Numerical simulations to examine the distribution of D_2 under the assumption of an i.i.d. letter distribution for biologically relevant parameter values:



Empirical distribution fits well to a Beta distribution with analytically determined E(D_2) and Var(D_2)…





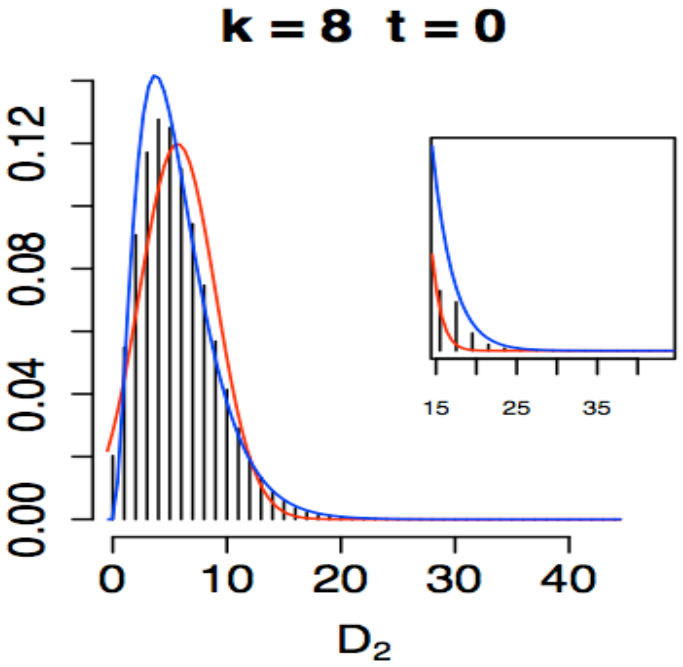
D_2 quantiles, n = 400, k = 8

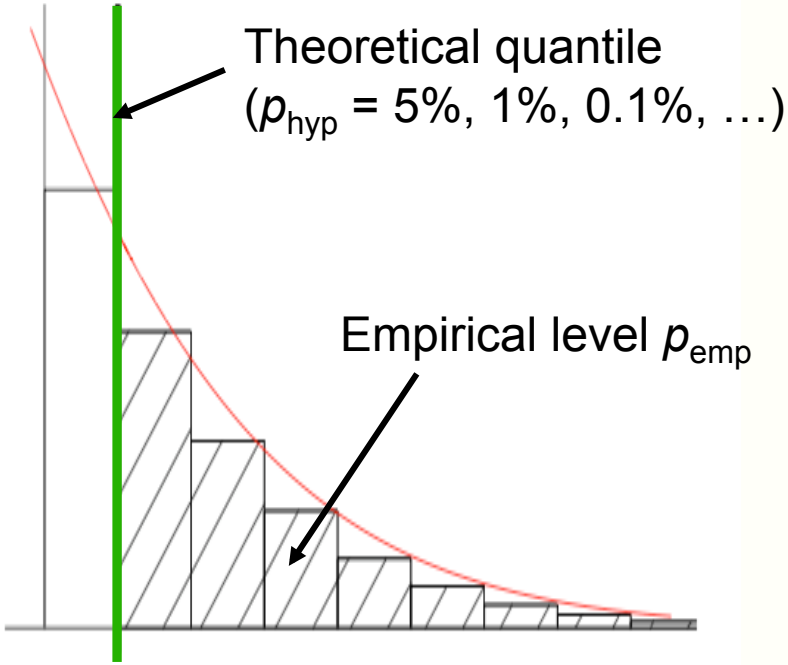
Empirical distribution fits well to a Beta distribution with analytically determined E(D_2) and Var(D_2)…

…but for accurate p-values the tail of the distribution is important.

→ Measure discrepancy between hypothesised (assuming e.g. Normal or Gamma) and empirical p-values:

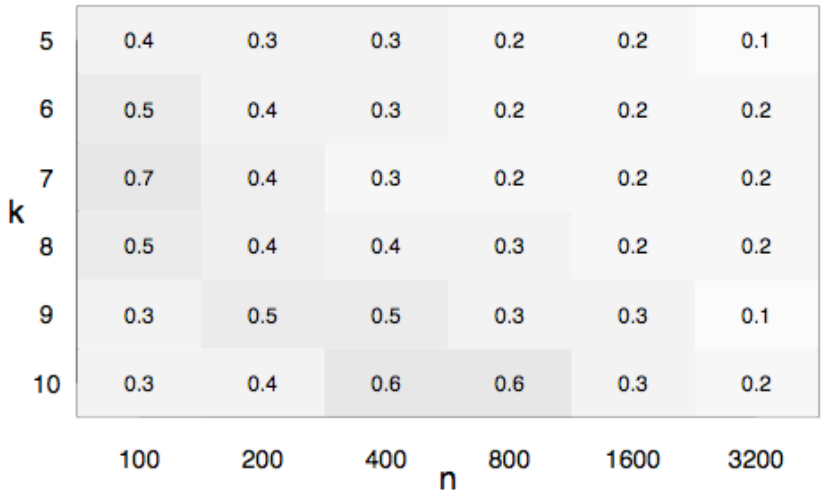
$$\delta = \log_{10}\left(p_{\text{emp}} / p_{\text{hyp}}\right)$$





Normal

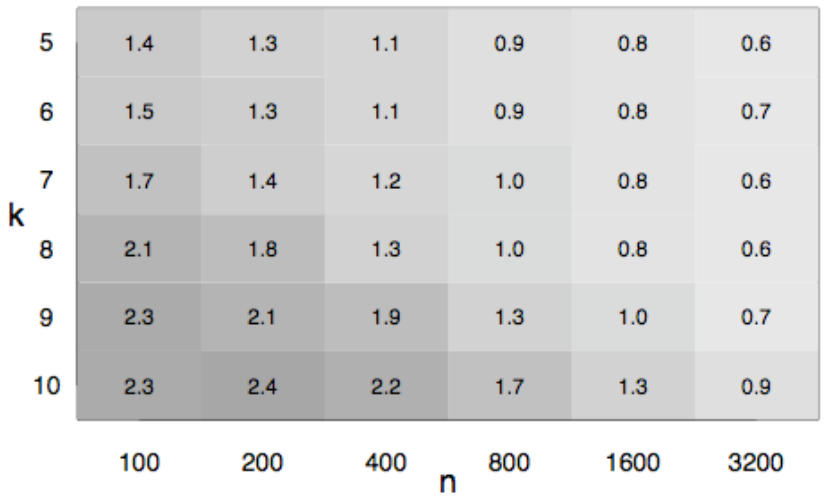
$p_{\text{emp}} = 0.01$



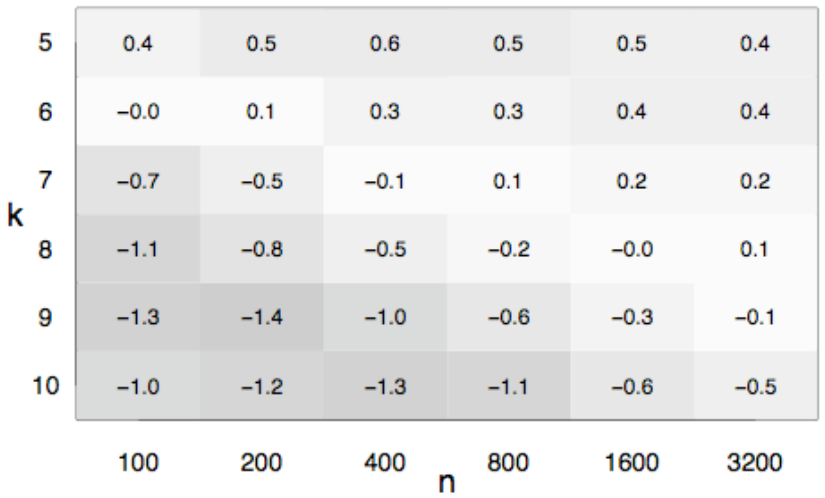
Beta

$p_{\text{emp}} = 0.01$

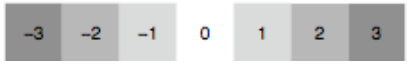
$p_{\text{emp}} = 0.0001$



$p_{\text{emp}} = 0.0001$

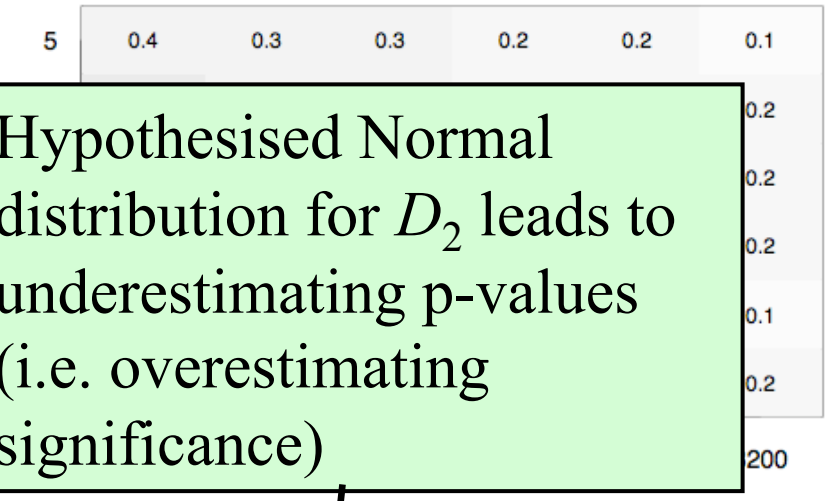


$\delta = \log_{10}\left(p_{\text{emp}} / p_{\text{hyp}}\right)$: -3 -2 -1 0 1 2 3



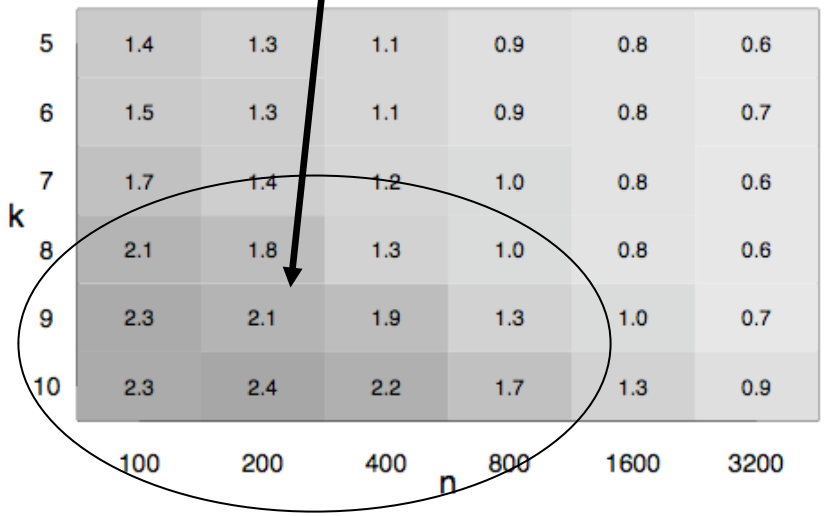
Normal

$p_{\text{emp}} = 0.01$



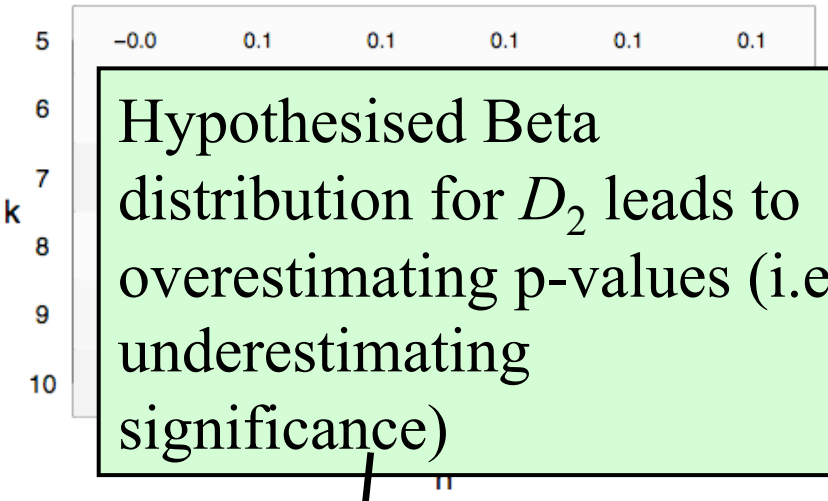
Hypothesised Normal distribution for D_2 leads to underestimating p-values (i.e. overestimating significance)

$p_{\text{emp}} = 0.0001$



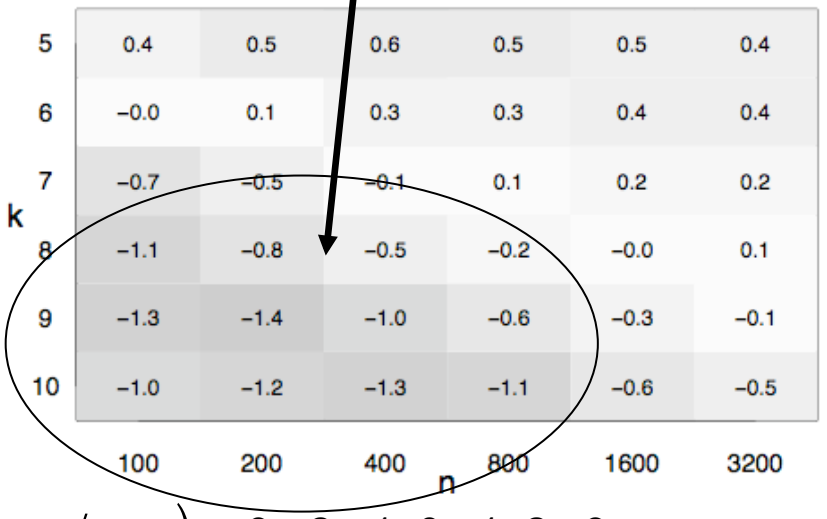
Beta

$p_{\text{emp}} = 0.01$



Hypothesised Beta distribution for D_2 leads to overestimating p-values (i.e. underestimating significance)

$p_{\text{emp}} = 0.0001$



$\delta = \log_{10}\left(p_{\text{emp}} / p_{\text{hyp}}\right)$: -3 -2 -1 0 1 2 3

Database test (using database from Kantorowitz, et al., *Bioinf*. 23 249 (2007)):

2 sets of sequences:

- positive control – a set of known cis- regulatory modules (mouse or human)

CACAAGATGAGAAGTTGTGTACTTG
GCAAACTTAGAGCTGACCTTTGCTGATTTG
GAAGTTGAAGATTACCCAACCATTGCA
GGTTTATCAGTTCTTTCTTGTTTAT
AGGTTGAGTTAATCATA
AGAAACAAAACCTACATGACCCTT
CTCTTGTTTTTTTATTCATTC
ACTGCCAAGAAGC
ATGCCAAAGTTAATCATTGG
CCCTGCTGAGTACATGGCCGATCAGGC
TGTTTTTGTGTGCCTGTTTTTCTATTTTAC
GTAAATCACCCTGAACATGTTTGCATCAAC
CTACTGGTGATGCACCTTTGATCAA
…
…

Database test (using database from Kantorowitz, et al., *Bioinf.* 23 249 (2007)):

2 sets of sequences:

- positive control – a set of known cis- regulatory modules (mouse or human)

CACAAGATGAGAAGTTGTGTACTTG
GCAAACTTAGAGCTGACCTTTGCTGATTTG
GAAGTTGAAGATTACCCAACCATTGCA
GGTTTATCAGTTCTTTCTTGTTTAT
AGGTTGAGTTAATCATA
AGAAACAAAACCTACATGACCCTT
CTCTTGTTTTTTTATTCATTC
ACTGCCAAGAAGC
ATGCCAAAGTTAATCATTGG
CCCTGCTGAGTACATGGCCGATCAGGC
TGTTTTTGTGTGCCTGTTTTTCTATTTTAC
GTAAATCACCCTGAACATGTTTGCATCAAC
CTACTGGTGATGCACCTTTGATCAA
…
…

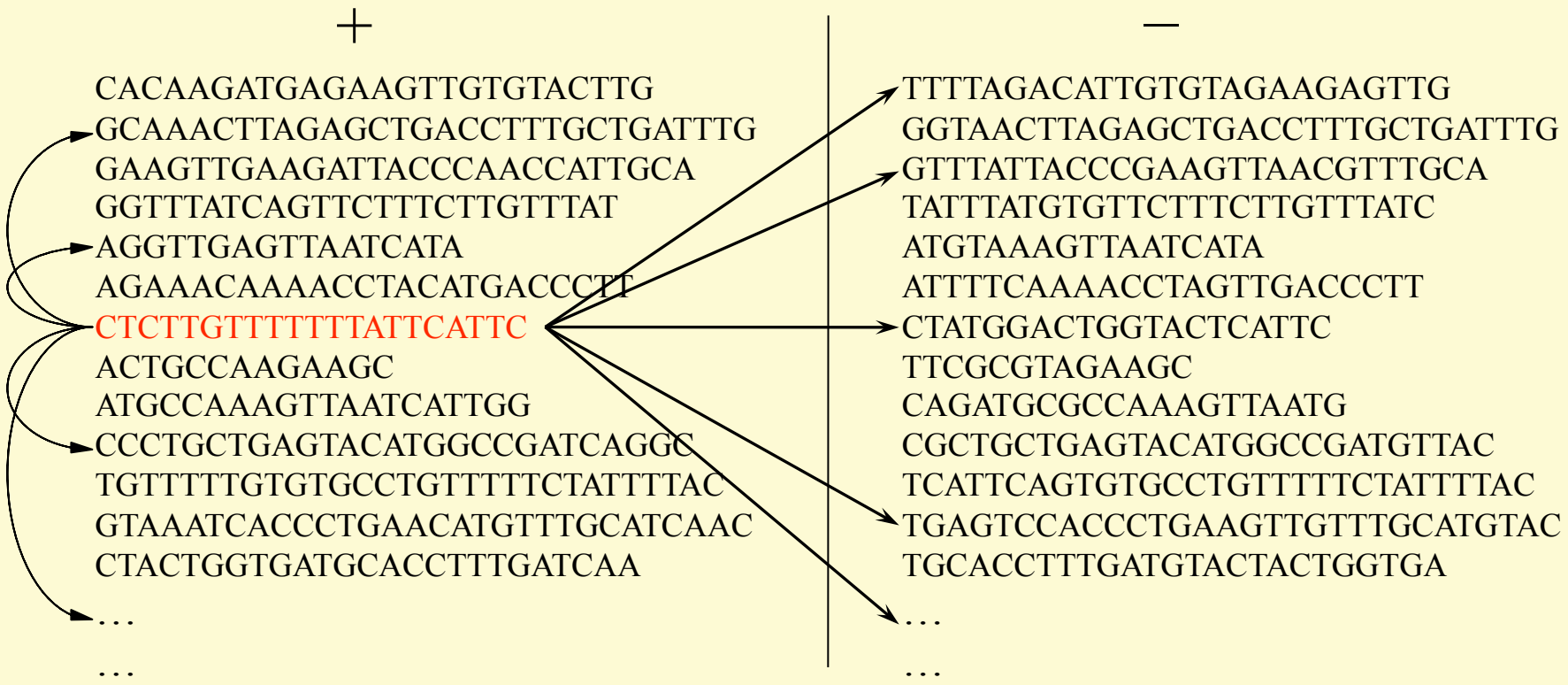
- negative control – a set of sequences of same length chosen randomly from non-coding part of genome

TTTTAGACATTGTGTAGAAGAGTTG
GGTAACTTAGAGCTGACCTTTGCTGATTTG
GTTTATTACCCGAAGTTAACGTTTGCA
TATTTATGTGTTCTTTCTTGTTTATC
ATGTAAAGTTAATCATA
ATTTTCAAAACCTAGTTGACCCTT
CTATGGACTGGTACTCATTC
TTCGCGTAGAAGC
CAGATGCGCCAAAGTTAATG
CGCTGCTGAGTACATGGCCGATGTTAC
TCATTCAGTGTGCCTGTTTTTCTATTTTAC
TGAGTCCACCCTGAAGTTGTTTGCATGTAC
TGCACCTTTGATGTACTACTGGTGA
…
…

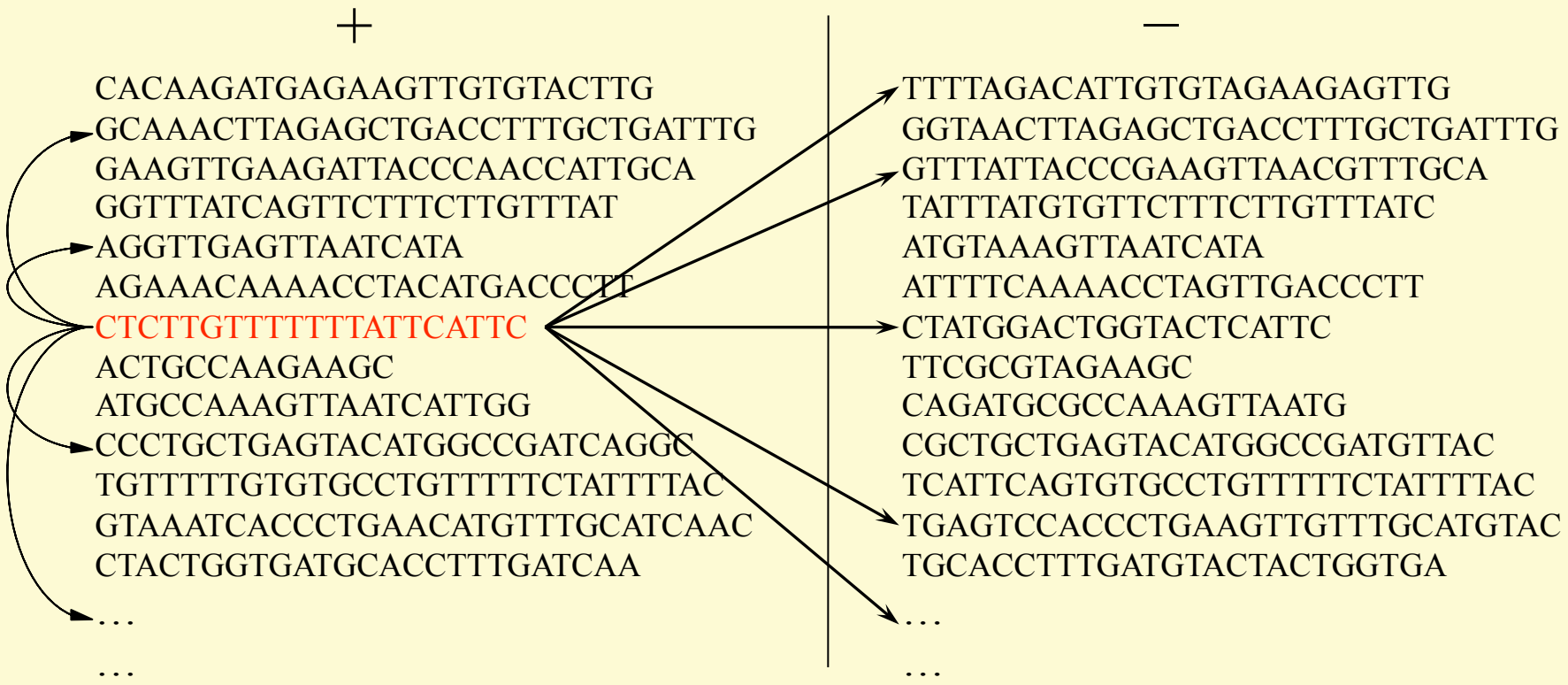
- chose each sequence in turn as the ‘query sequence’. Attempt to classify it as positive or negative control as follows

+	−
CACAAGATGAGAAGTTGTGTACTTG	TTTTAGACATTGTGTAGAAGAGTTG
GCAAACTTAGAGCTGACCTTTGCTGATTTG	GGTAACTTAGAGCTGACCTTTGCTGATTTG
GAAGTTGAAGATTACCCAACCATTGCA	GTTTATTACCCGAAGTTAACGTTTGCA
GGTTTATCAGTTCTTTCTTGTTTAT	TATTTATGTGTTCTTTCTTGTTTATC
AGGTTGAGTTAATCATA	ATGTAAAGTTAATCATA
AGAAACAAAACCTACATGACCCTT	ATTTTCAAAACCTAGTTGACCCTT
CTCTTGTTTTTTTATTCATTC	CTATGGACTGGTACTCATTC
ACTGCCAAGAAGC	TTCGCGTAGAAGC
ATGCCAAAGTTAATCATTGG	CAGATGCGCCAAAGTTAATG
CCCTGCTGAGTACATGGCCGATCAGGC	CGCTGCTGAGTACATGGCCGATGTTAC
TGTTTTTGTGTGCCTGTTTTTCTATTTTAC	TCATTCAGTGTGCCTGTTTTTCTATTTTAC
GTAAATCACCCTGAACATGTTTGCATCAAC	TGAGTCCACCCTGAAGTTGTTTGCATGTAC
CTACTGGTGATGCACCTTTGATCAA	TGCACCTTTGATGTACTACTGGTGA
…	…
…	…

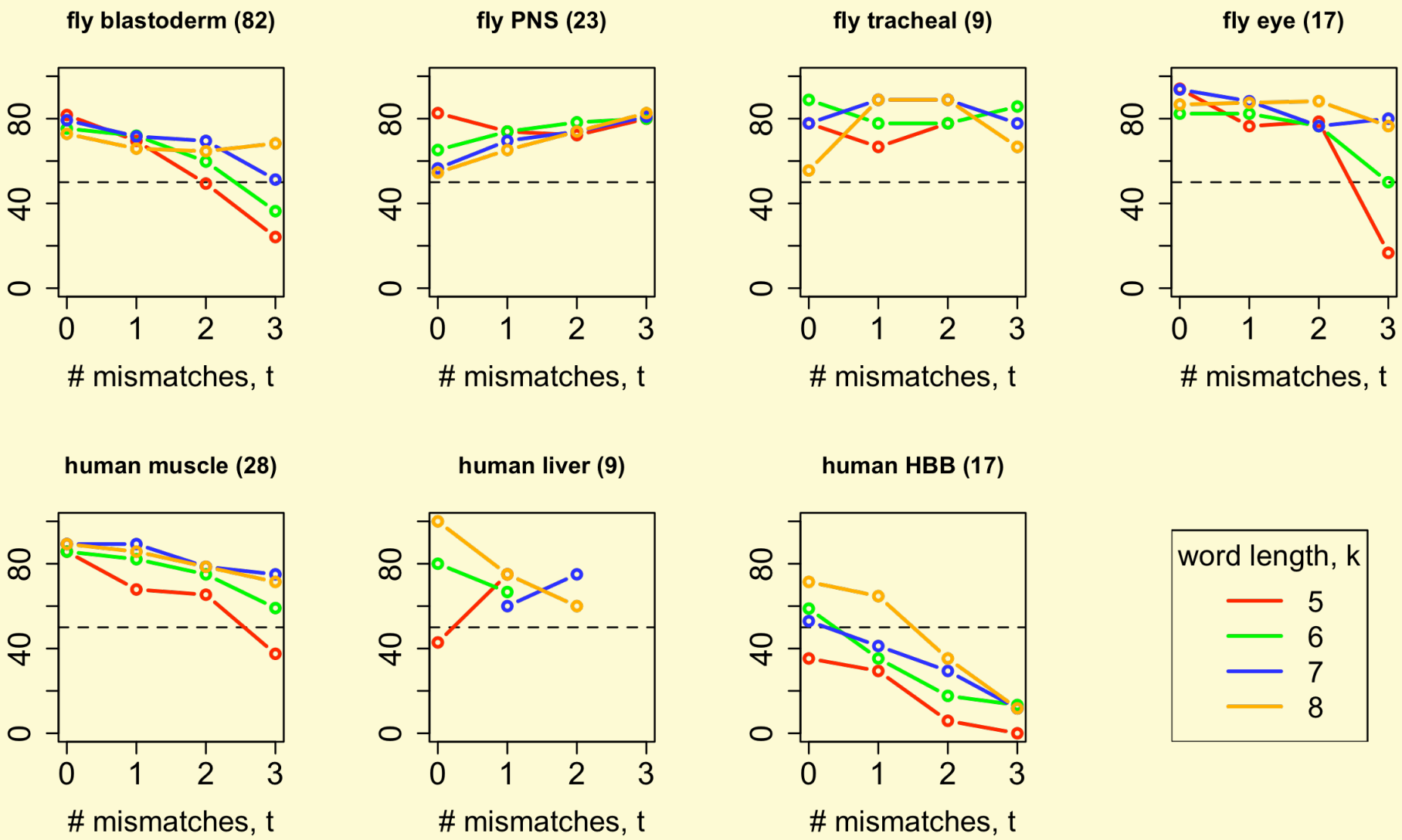
- chose each sequence in turn as the ‘query sequence’. Attempt to classify it as positive or negative control as follows
- measure $D_2^{(t)}$ between query and all other sequences



- chose each sequence in turn as the ‘query sequence’. Attempt to classify it as positive or negative control as follows
- measure $D_2^{(t)}$ between query and all other sequences
- convert $D_2^{(t)}$ values to (i.i.d.) null hypothesis p-values
- smallest p-value determines whether query belongs to positive or negative control



Percentage of times (+ve) query sequence is correctly classified using minimum p-value



Conclusions

- The k-word count D_2 is a fast and accurate statistic for sequence comparison when alignments are not appropriate
- Approximate word count $D_2^{(t)}$ is slower to calculate, but more accurate, and more appropriate for some applications
- Mean and variance of D_2 and $D_2^{(t)}$ can be computed easily (analytic result for D_2)
- The Beta distribution gives a good empirical estimate of p-values for D_2, and for its extreme value distribution

Papers

- ‘Asymptotic behaviour and optimal word size for exact and approximate word matches between random sequences’, S. Forêt, M.R. Kantorovitz and C.J. Burden, *BMC Bioinformatics*, 7 (2006) S21.

- ‘Asymptotic behaviour of *k*-word matches between two random sequences’, M.R. Kantorovitz, H.S. Booth, C.J. Burden and S.R. Wilson, *J. Appl. Prob*, 44 (2007), 788-805.

- ‘Asymptotic behaviour and optimal word size for exact and approximate word matches’, S.R. Wilson, and C.J. Burden, *Proc. Appl. Math. Mech.*, 7, 11218101.

- ‘Approximate word matches between two random sequences’, C.J. Burden, M.R. Kantorovitz and S.R. Wilson, *Ann. Appl. Prob*., 18 (2008) 1-21.

- ‘Empirical distribution of *k*-word matches in bilogical sequences, S. Forêt, S.R. Wilson and C.J. Burden, *Pattern Recgn*., 42 (2009) 539-548.

- ‘Characterising the D_2 statistic: word matches in biological sequences’, S. Forêt, S.R. Wilson and C.J. Burden, *Stat. Appl. Gen. Mol Biol.* 8 (2009) Art 43.