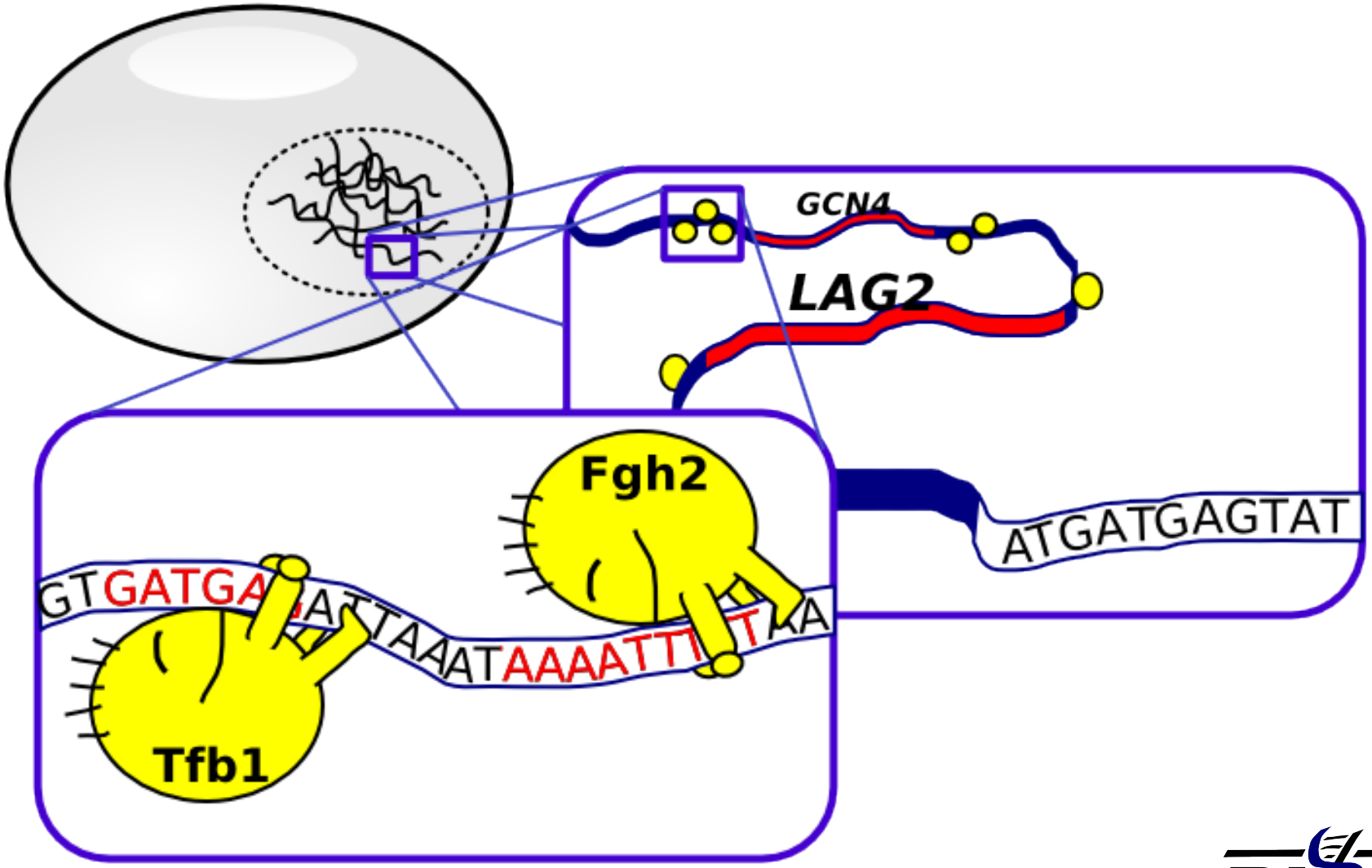
An Evolutionary Model of DNA Substring Distribution

Konstantin Tretyakov (kt@ut.ee)
(joint work with Meelis Kull and Jaak Vilo)

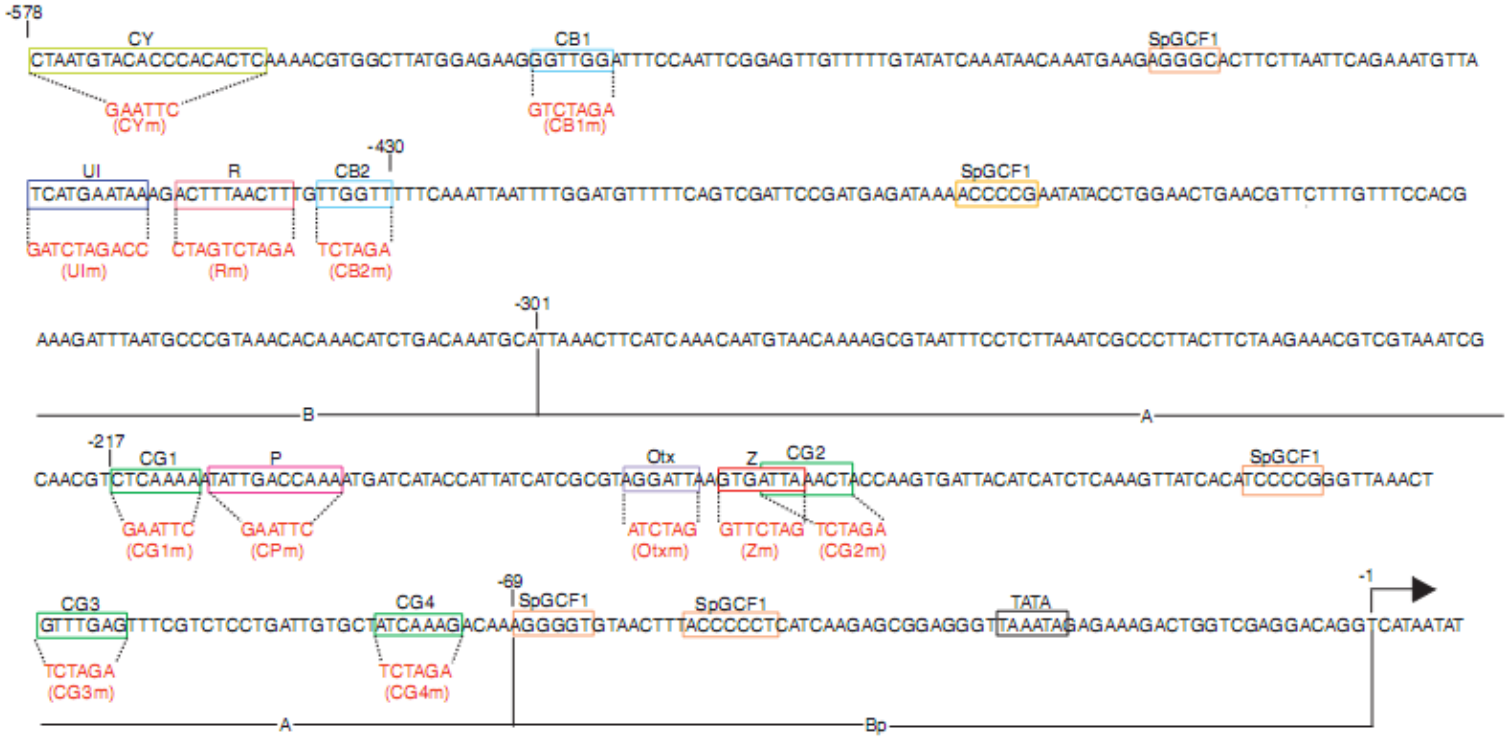
University of Tartu
Bioinformatics & Data Mining Group

DNA Regulatory Regions



DNA Regulatory Regions

Promoter sequence of Endo16 gene



Purple sea urchin

BIOIT

Motif Discovery

FGH: ATCG**GTGATGAGTT**TAGACG
G**GAGATGCGTT**ATCATGTAG
AGATCGAGATACGATAGBAA
GAAAATAGC**GAGATGAGTA**

HGH: TATAATCG**GTGATGAGTT**TAG
AGAGAGABACGGAGATACGA
TAT**CAGATCAGAT**TGGTA

BIOIT

Modeling "Background"

Sequence:

CTAATGTACACCCACACTCAAAACGTGGCTTATGGAGAAGGGTTGGATTTCCAATT
ACTTCTTAATTCAGAAATGTTATCATGAATAAAGACTTTAACTTTGTTGGTTTTTC
AAAACCCCGAATATACCTGGAACTGAACGTTCTTTGTTTCCACGAAAGATTTAATG
CAAACAATGTAACAAAAGCGTAATTTCCTCTTAAATCGCCCTTACTTCTAAGAAAC
TCATACCATTATCATCGCGTAGGATTAAGTGATTAAACTACCAAGTGATTACATCA
TCGTCTCCTGATTGTGCTATCAAAGACAAAGGGGTGTAACTTTACCCCCTCATCAA

Background:

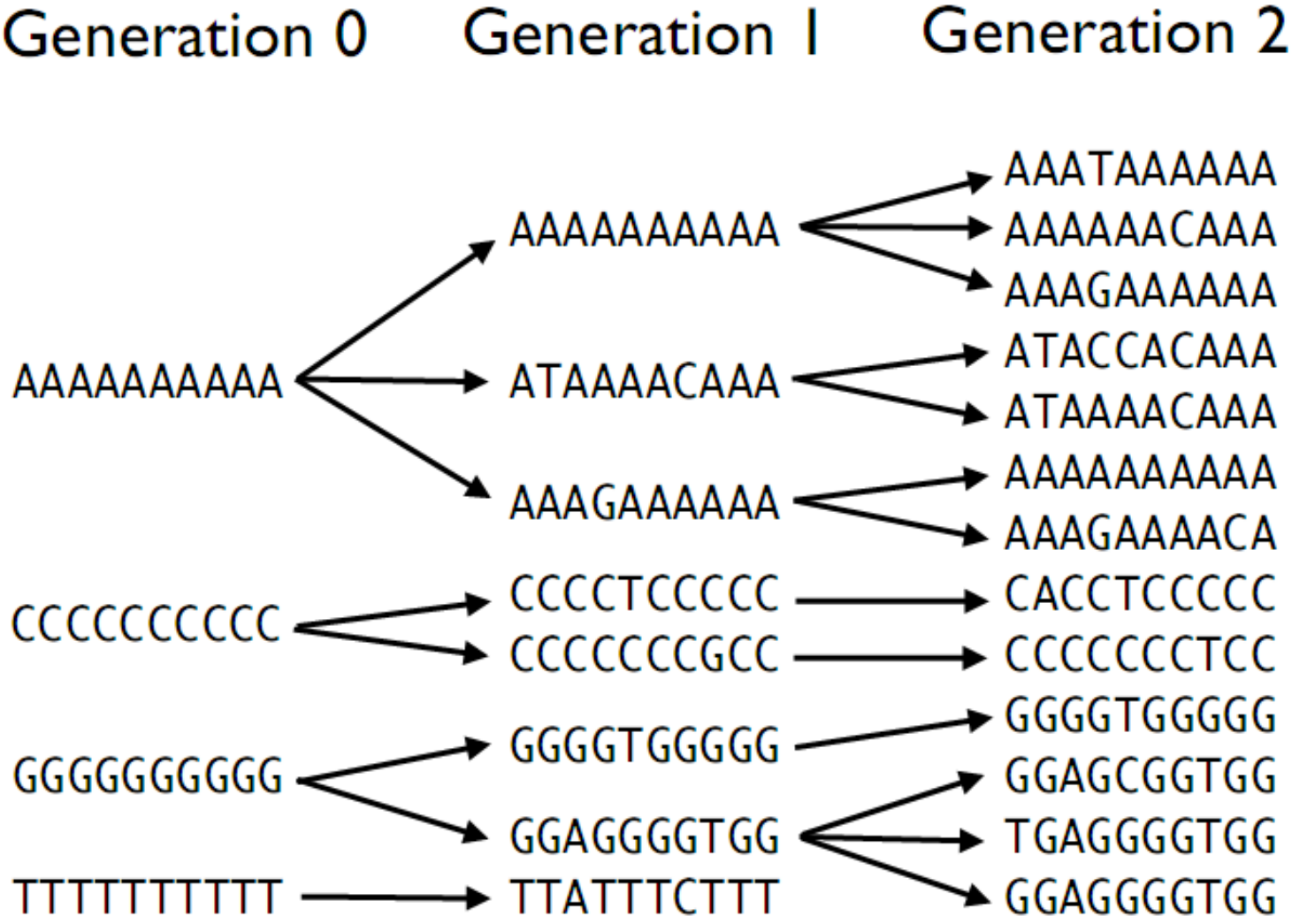
A - 34%	AA - 14%	GA - 6%
C - 19%	AC - 6%	GC - 2%
G - 18%	AG - 5%	GG - 4%
T - 30%	AT - 9%	GT - 6%
	CA - 7%	TA - 7%
	CC - 4%	TC - 7%
	CG - 3%	TG - 6%

Modeling "Background"

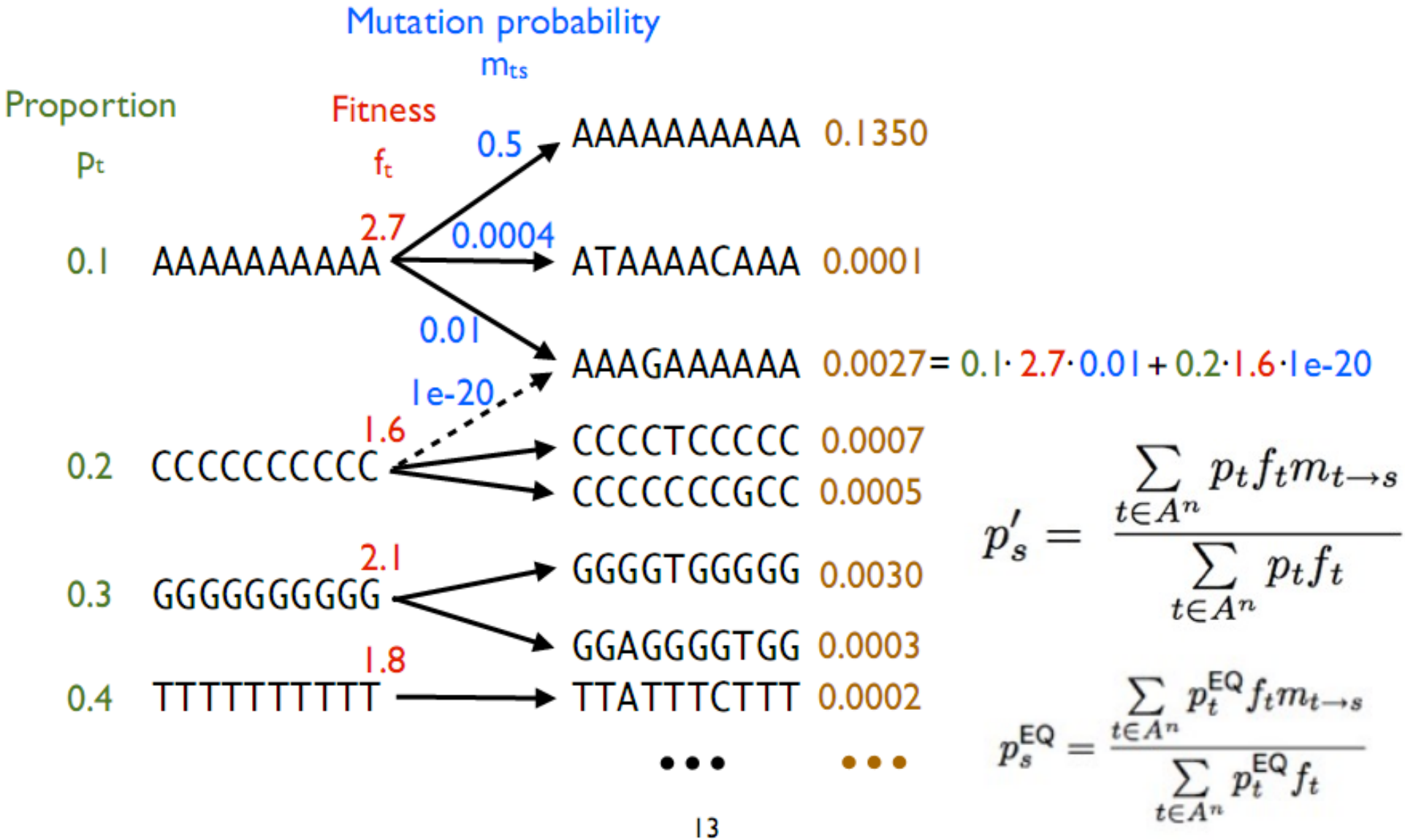
- How does a DNA substring distribution **arise** in the first place?
  - Evolutionary model

BIOIT

Evolutionary Model



Evolutionary Model



k-mer equilibrium distribution

	p_s^{EQ}
AAAAAAAAAA	1.5e-5
AAAAAAAAAC	4.3e-6
AAAAAAAAAG	3.6e-6
AAAAAAAAAT	5.2e-6
AAAAAAAACA	7.1e-6
...	
CAGACAGCAA	3.5e-5
...	
TTTTTTTTTC	9.8e-6
TTTTTTTTTG	1.0e-5
TTTTTTTTTT	2.6e-5

k=3	
AAA	0.052
AAC	0.027
AAG	0.023
AAT	0.015
ACA	0.043 = 1/10 · 7.1e-6 + 2/10 · 3.5e-5 + ...
ACC	0.018
ACG	0.012
ACT	0.010
AGT	0.016
ACA	0.013
...	
TTC	0.021
TTG	0.027
TTT	0.033

Main contribution

- **Given the fitness function and mutation rates, compute the expected frequencies of K-mers.**

- Mustonen and Lässig. Evolutionary population genetics of promoters: predicting binding sites and functional phylogenies. Proc Natl Acad Sci USA (2005) vol. 102 (44) pp. 15936-41
- Moses et al. Large-scale turnover of functional transcription factor binding sites in Drosophila. PLoS Comput. Biol. (2006) vol. 2 (10) pp. e130
- Huang et al. Phylogenetic simulation of promoter evolution: estimation and modeling of binding site turnover events and assessment of their impact on alignment tools. Genome Biol (2007) vol. 8 (10) pp. R225
- Doniger and Fay. Frequent gain and loss of functional transcription factor binding sites. PLoS Comput. Biol. (2007) vol. 3 (5) pp. e99

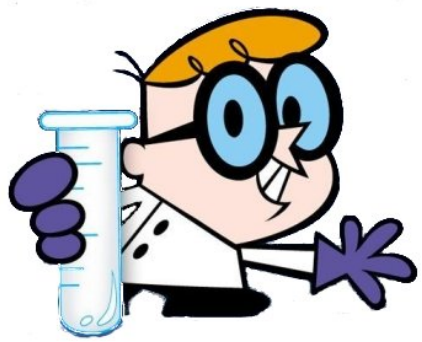
BIOIT

k-mer distribution

$$p_a^{\mathsf{EQ}} = \frac{\sum_{b \in A^k} m_{b \to a} \sum_{u \in A^{n-k}} p_{b \cdot u}^{\mathsf{EQ}} f_{b \cdot u}}{\sum_{b \in A^k} \sum_{u \in A^{n-k}} p_{b \cdot u}^{\mathsf{EQ}} f_{b \cdot u}}$$

$$\sum_{u \in A^{n-k}} p_{b \cdot u}^{\mathsf{EQ}} f_{b \cdot u} \approx \frac{p_b^{\mathsf{EQ}}}{|A^{n-k}|} \sum_{u \in A^{n-k}} f_{b \cdot u}$$

$$p_a^{\mathsf{EQ}_k} = \frac{\sum_{b \in A^k} p_b^{\mathsf{EQ}_k} \cdot f_b m_{b \to a}}{\sum_{b \in A^k} p_b^{\mathsf{EQ}_k} f_b}$$



Experiments

- To check approximations we have compared p^{EQ} and p^{EQ_k}
- Calculation of p^{EQ} is hard, thus we took sequences of length 8 of letters A and C
- Independent point mutations with probability r at each location
- Fitness defined by the number of occurrences of a motif q

BIOIT

Results

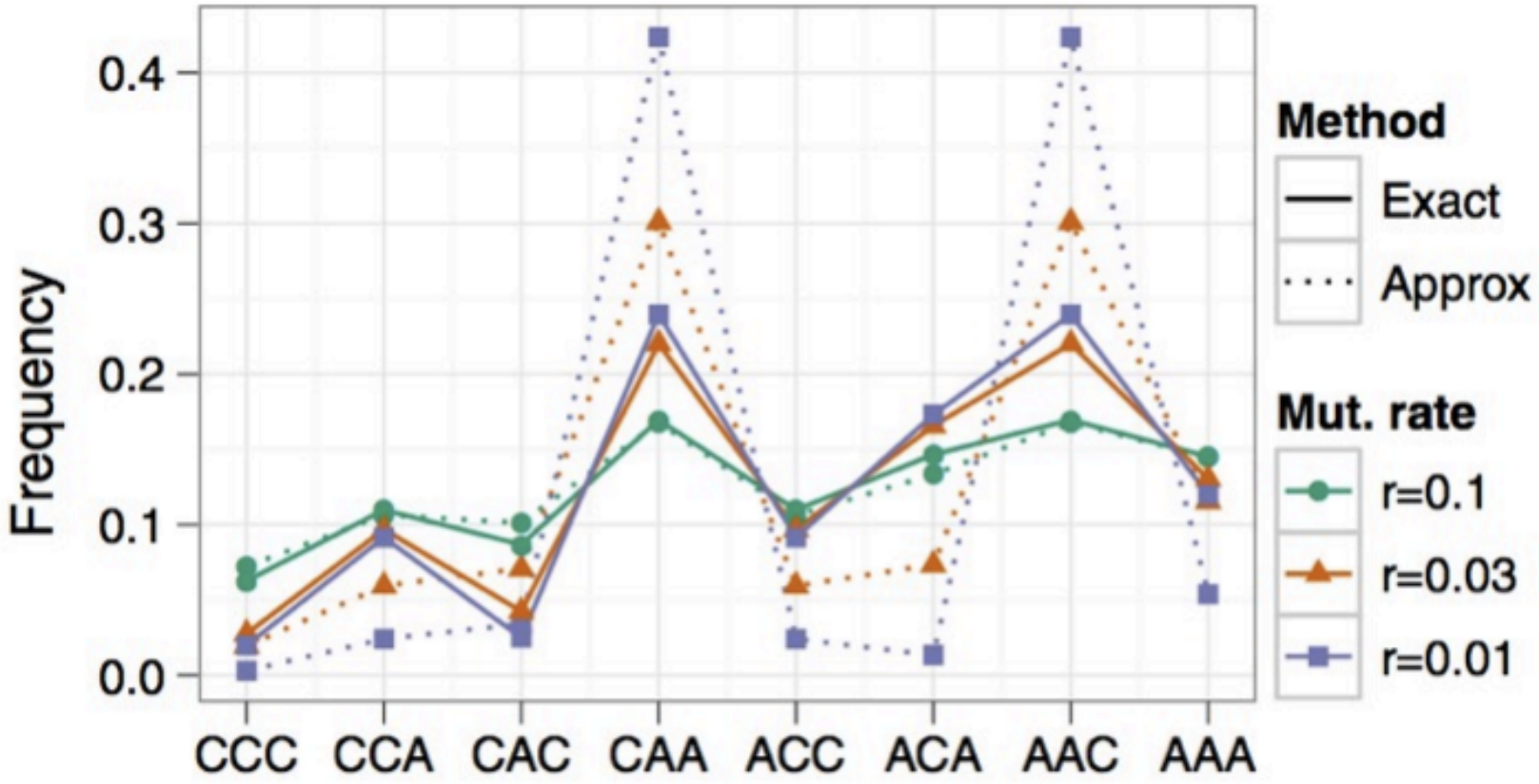


Fig. 2. The exact and approximate 3-mer distributions for point-mutation rates $r = 0.1, 0.03, 0.01$ where fitness is defined with strategy (S2) for substring $q = \text{AAC}$

Results

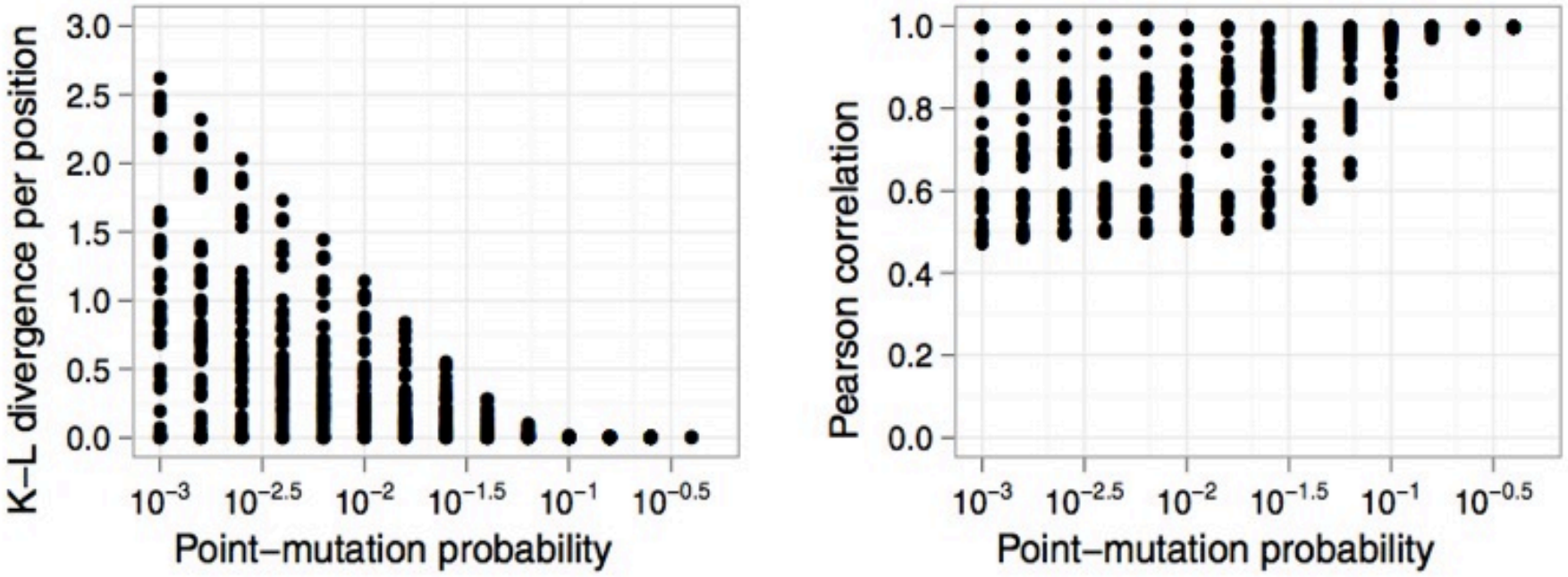


Fig. 1. The effect of point-mutation rate r on the approximation quality measured as Kullback-Leibler divergence per position and correlation between the exact and approximated distributions. Each circle denotes an experiment with a different set of parameters.

Future work

- Matching the model on data
  - Estimating model parameters from data
  - When and to what extent does the genetic equilibrium hold
- Improving approximation
- Applications to motif discovery

BIOIT

Thanks