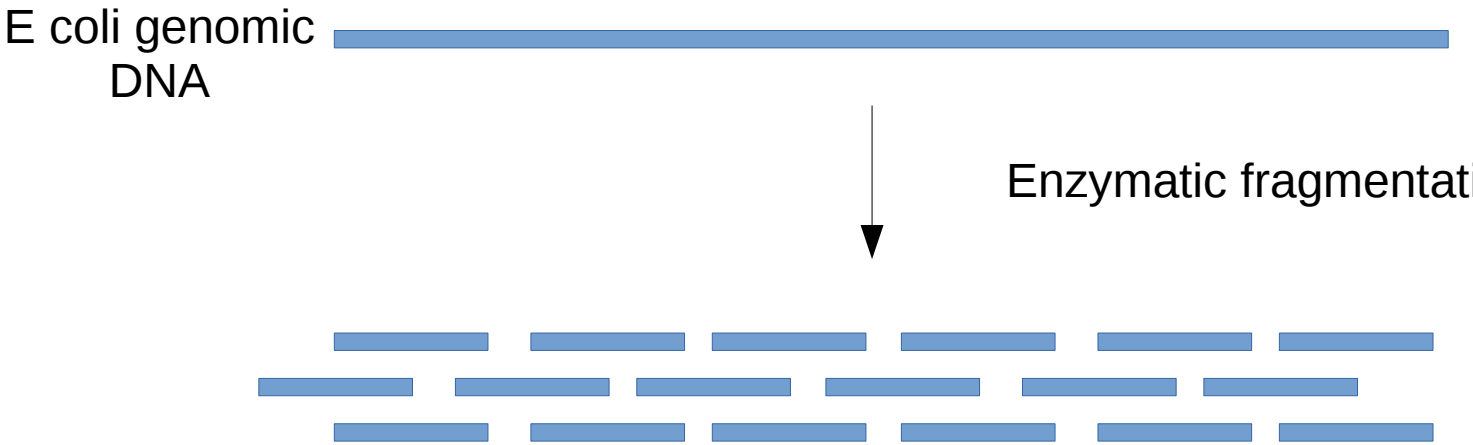
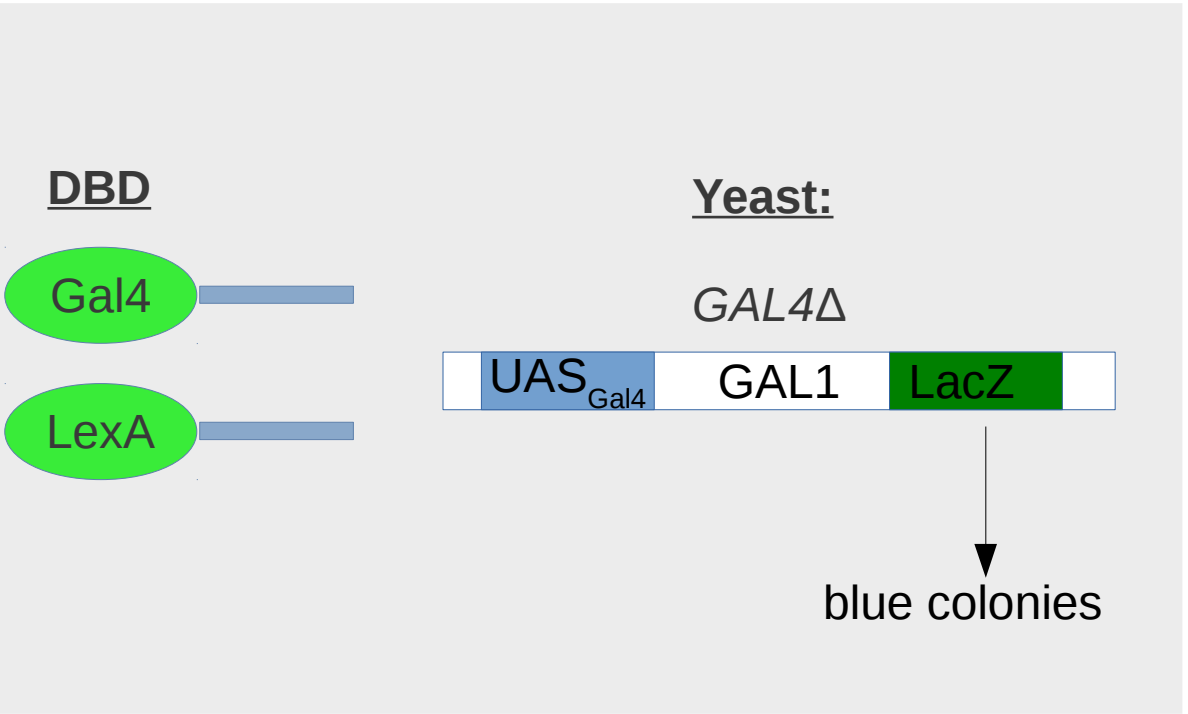
Initial "*motif search*" on TADs





- 15000 transformants
- 0.1-1% function as TA.
- 12 to 81 aa
- Negative charge

Ma & Ptashne, 1987 (Cell)
Ruden et al., 1991 (Nature)

Small molecules mimicking TADs

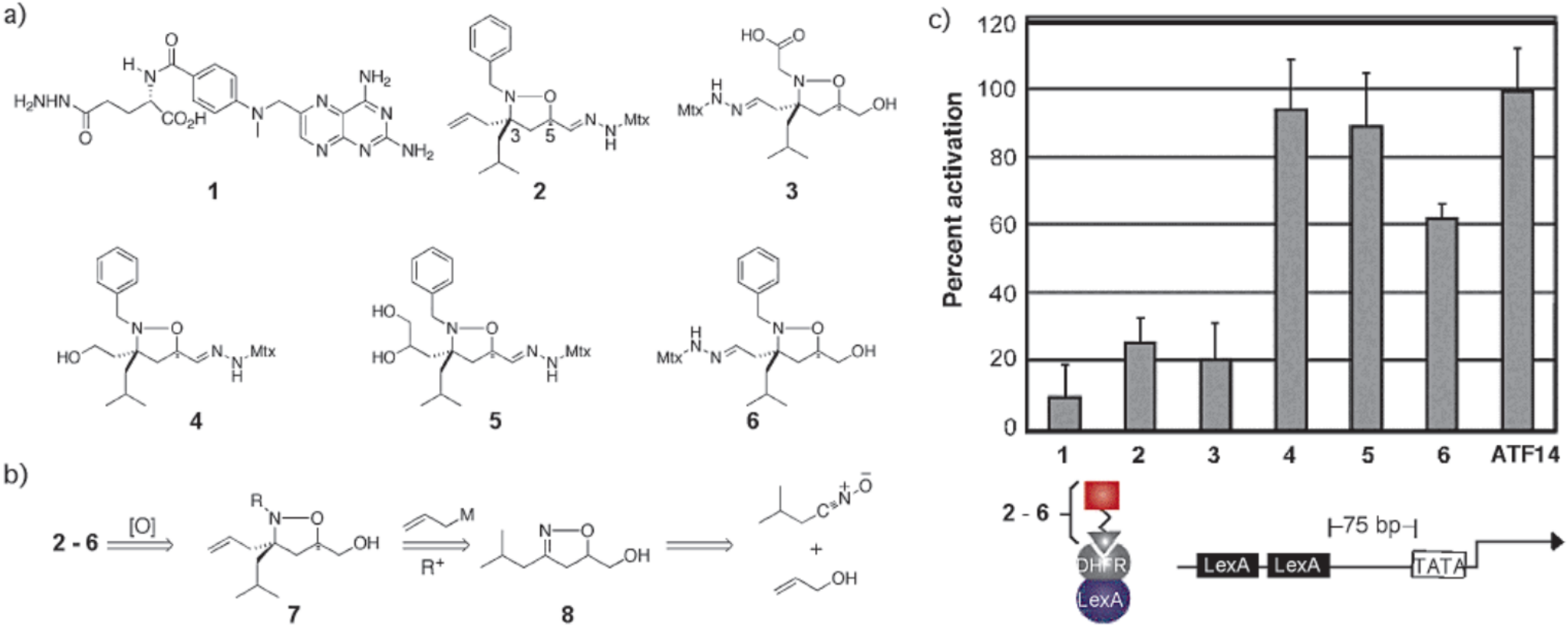


Figure 2. Isoxazolidine-based activation domains. (a) Five isoxazolidines (**2**−**6**) bearing functional groups commonly found in natural activation domains were targeted. (b) Synthetic strategy used to prepare isoxazolidines. (c) Results from in vitro transcription assays. The activity of each compound represents the average of at least three individual experiments with the indicated error (SDOM). For details see the Supporting Information.

Minter et al., 2004 (JACS)

TAD motif?

Supplementation table

Supplemented Oaf1/Pip2/Gal4 pattern:

```
p1:           GSTYWDENQRKH
p2: FLIVAM  +    YWDENQ
p3: FLIVAM  + GSTYW
p4: FLIVAM  +    YWDENQ
p5: FLIVAM  + GSTYWDENQRKH
p6: FLIVAM  +    YWDENQ
p7: FLIVAM  + GSTYW
p8: FLIVAM  + GSTYWDENQRKH
p9: FLIVAM  + GSTYWDENQRKH
```

D

Oaf1/Pip2/Gal4 9aa TAD pattern

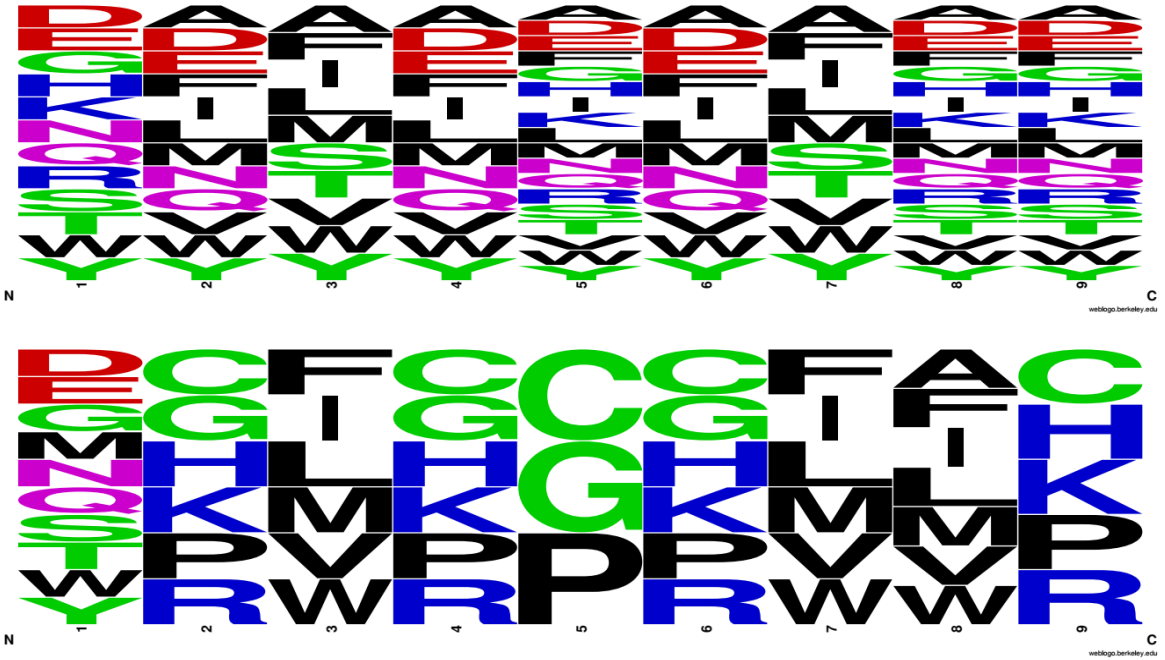
[GSTYWDENQRKH] [FLIVAMYWDENQ] [FLIVAMSTYW] [FLIVAMYWDENQ]
[FLIVAMGSTYWDENQRKH] [FLIVAMYWDENQ] [FLIVAMSTYW]
[FLIVAMGSTYWDENQRKH] [FLIVAMGSTYWDENQRKH]

Yeast 9aa TAD pattern

[GSTDENQWYM] {KRHCGP} [FLIVMW] {KRHCGP} {CGP}
{KRHCGP} [FLIVMW] [FLIVAMW] {KRHCP}

Animal 9aa TAD pattern

[GSTDENQWYM] {KRHCGP} [FLIVMW] {KRHCGP}
{CGP} {CGP} [FLIVMW] {CGP}{CGP}



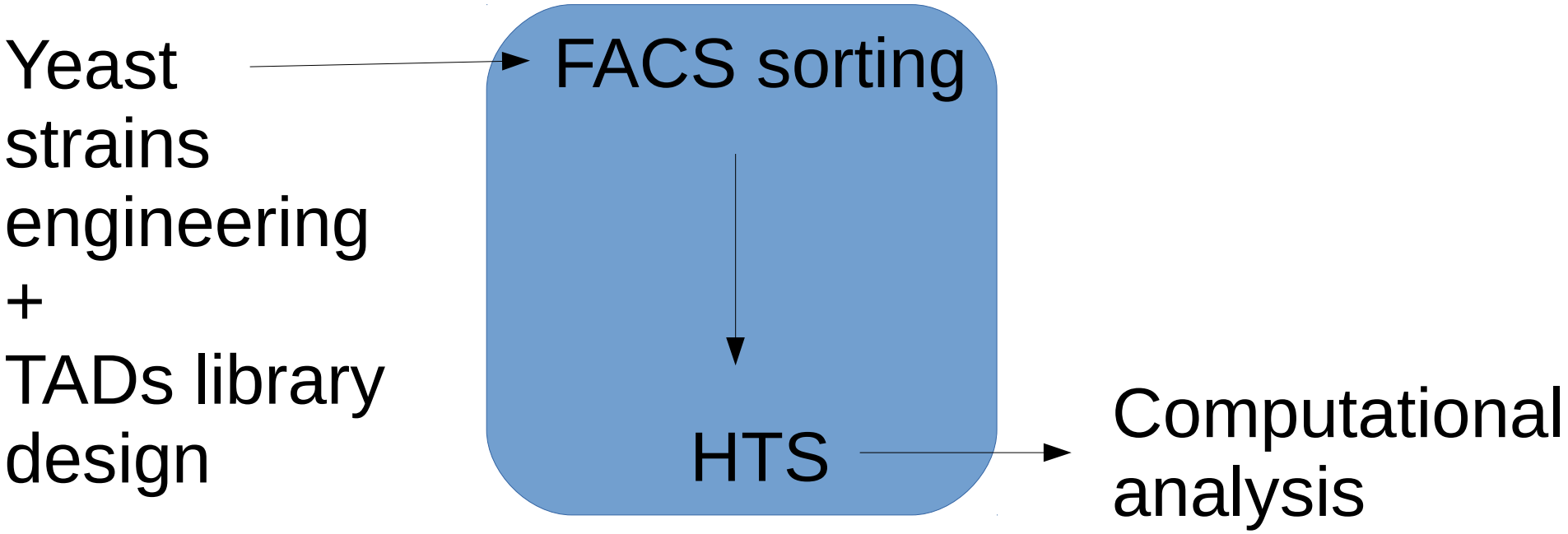
Piskacek et al., 2007 (Genomics)

And yet… it's not clear what TADs have in common.

- **Assumption:**
  There are patterns or motifs common to all TADs.

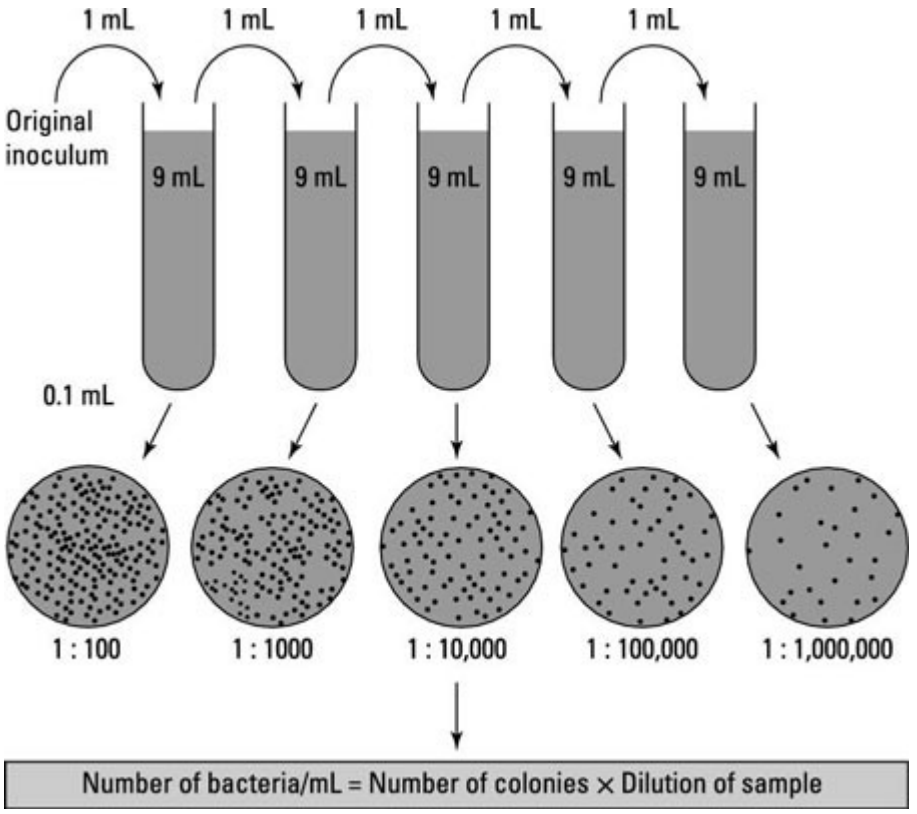
- **Plan:**
  Use *state of the art* experimental and computational methodologies to analyze a big combinatorial space of TAD sequences to find motifs or patterns.

Workflow

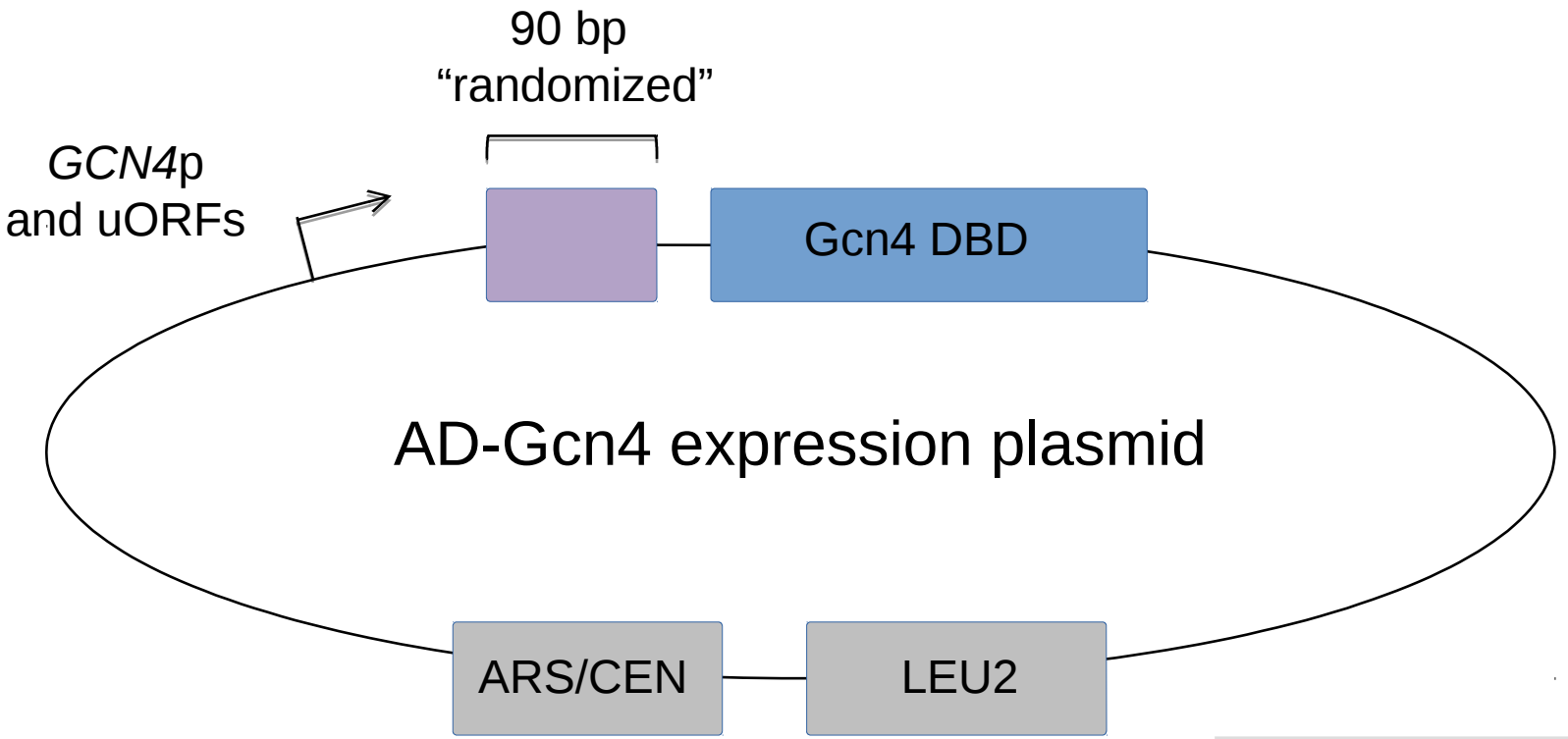

Yeast
strains
engineering
+
TADs library
design
FACS sorting
HTS
Computational
analysis

Bottleneck is yeast transformation

~10^7 **CFU**



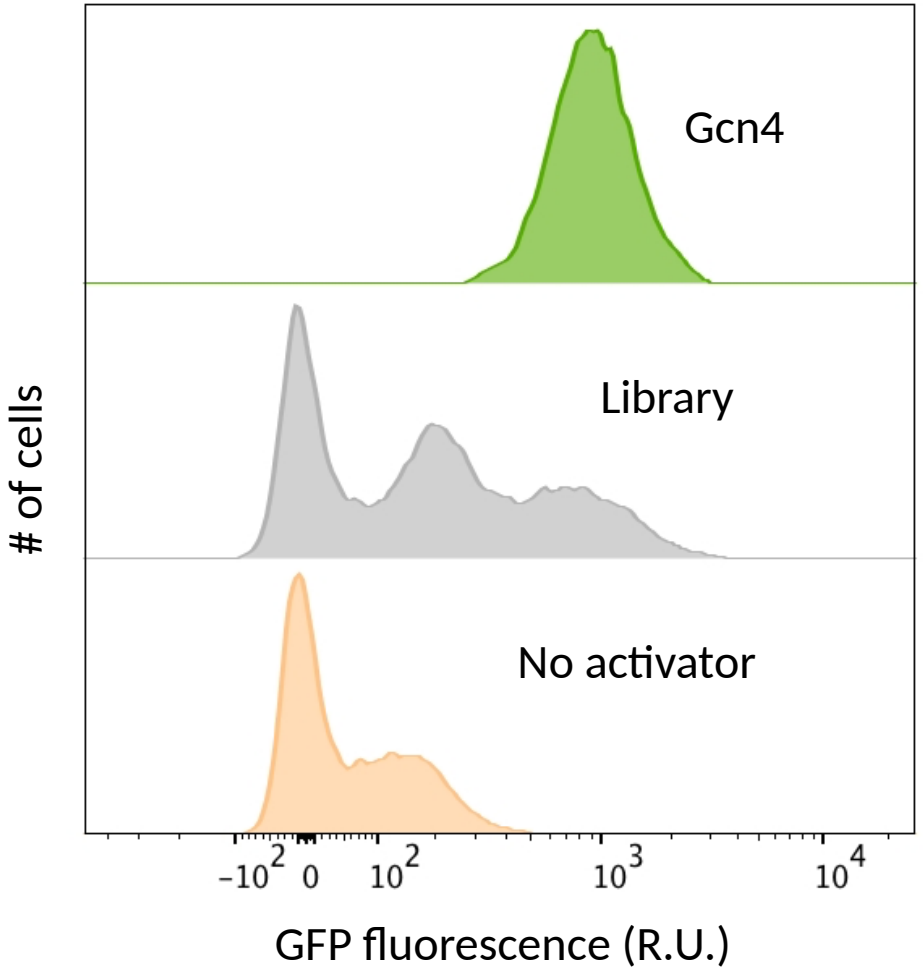
Strategy for high throughput isolation of activation domains



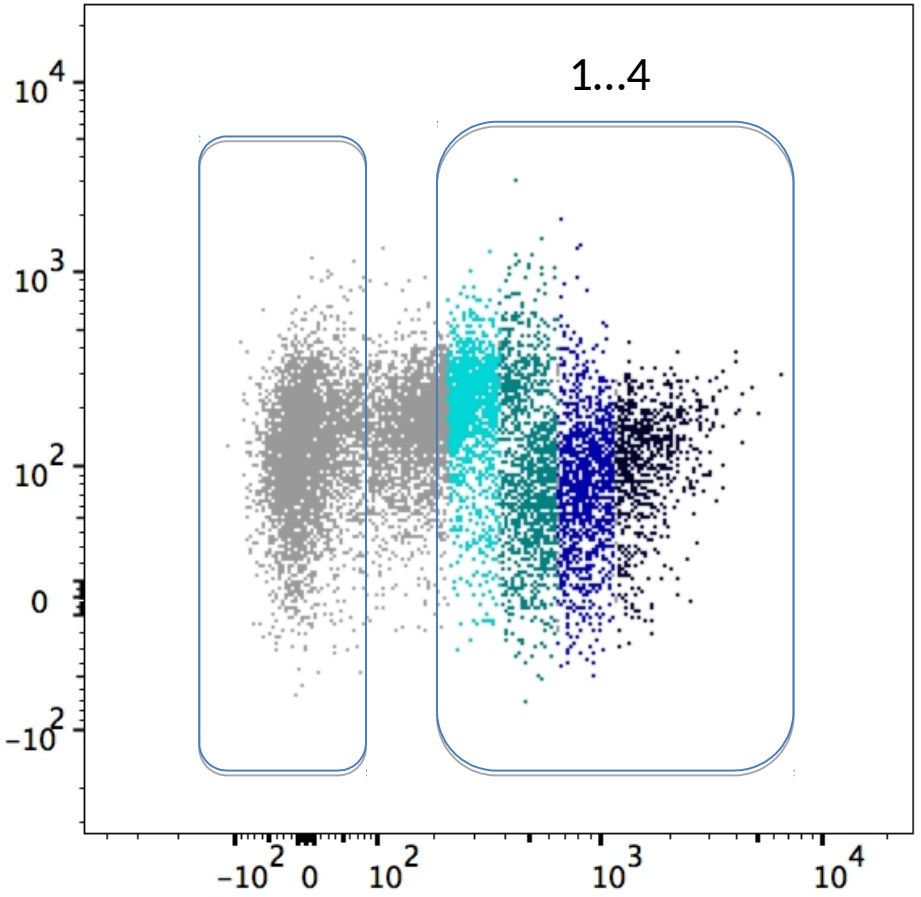
- 3-Amino Triazole O.N
- 3-AT 2hs
- No 3-AT

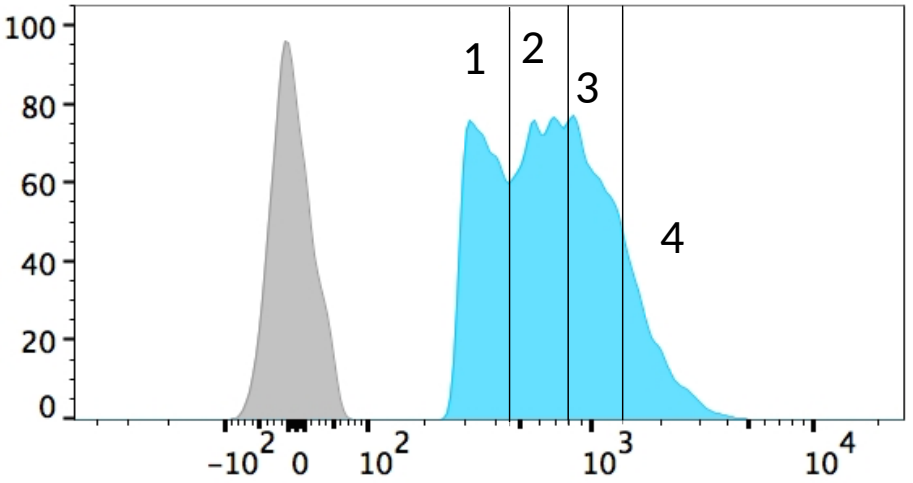
Linda Warfield, Ariel, Erijman, Steven Petesch

FACS selection of TAD libraries



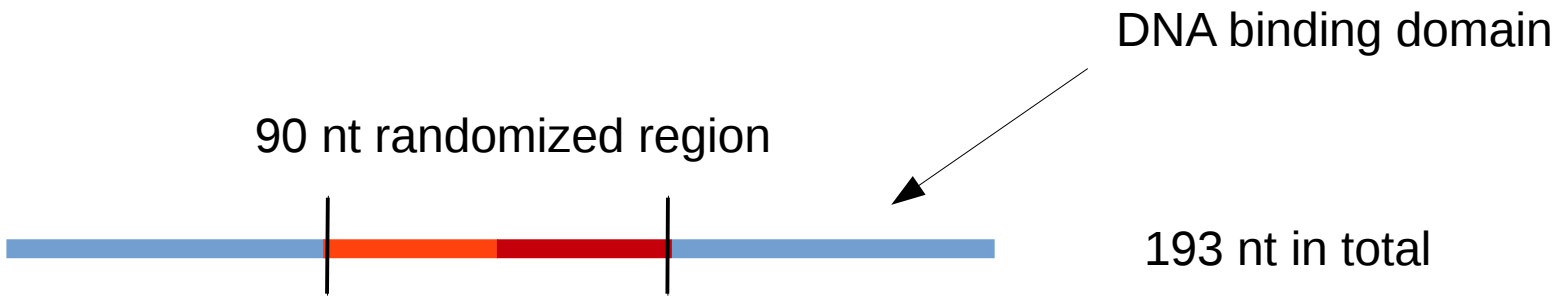
Splitting novel ADs on their strength

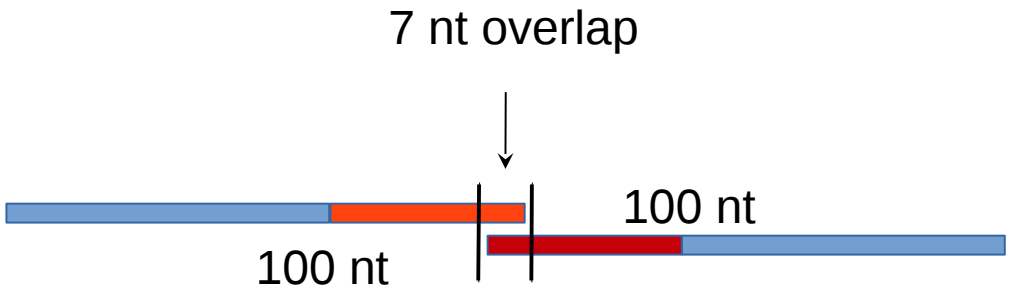




Library sequencing

Platform: HiSeq Illumina (paired end) - 100nt reads with 7nt overlap





HT-seq analysis

- **PAIR READS** (FLASH, PMID: 21903629)
- **TRANSLATION TO AMINO ACID** (custom script)
- **CLUSTER SEQUENCES** (USEARH, pmid: 20709691) - sequences are redundant, probably due to random techinical errors. Clusters allow up to 6 mismatches (20%).
- **SCORE SEQUENCES** based on number or reads/bin
- **PREDICT PHYSICOCHEMICAL PROPERTIES OF 30mers** (*intrinsic disorder*: IUPred, *Secondary Structure*: PSIPRED, *GRAVY scores*: custom scripts)

- **DEEP LEARNING…** you?
- Tried MEME and Gibs sampler without success...

Translation to Amino-acids

(inspect the raw reads in FastaQC program followed by custom scripts – Qual offset=33, HiSeq Illumina v. >1.8)

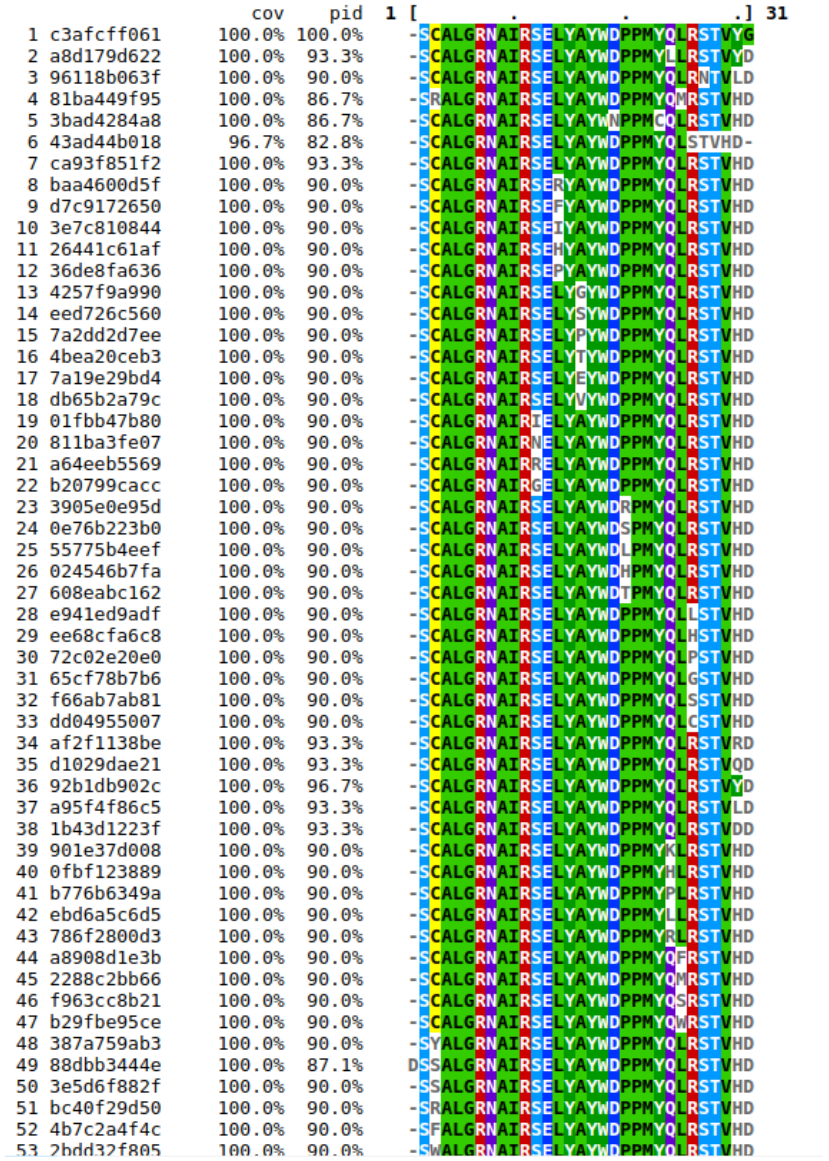
- Include filters for:
  - Early-stop (0.9 – 40%)
  - no5-primer (~3%)
  - no3-primer (5-25%)
  - Frame-shift (~0.3%)
  - Low-quality$_{\text{Phred=30}}$ (15-35%)
  - Short-seq (0.2-20%)

Remove sequence redundancy that might arise from technical errors

USEARCH

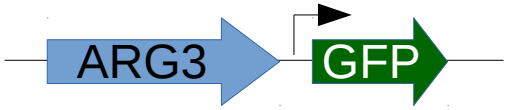
find clusters

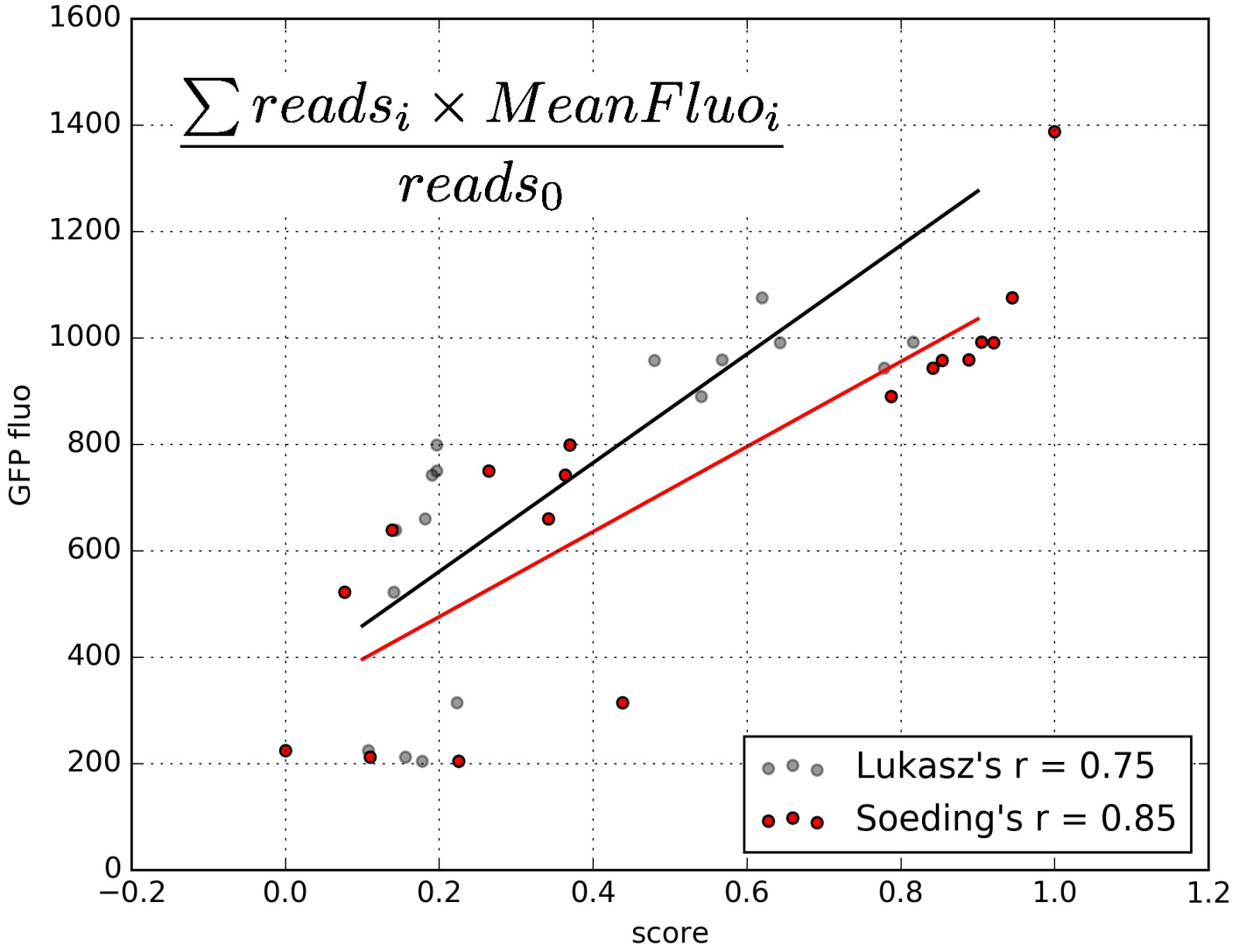
Low reads seqs
Merged into few
Seqs with
high # reads



Scoring the sequences

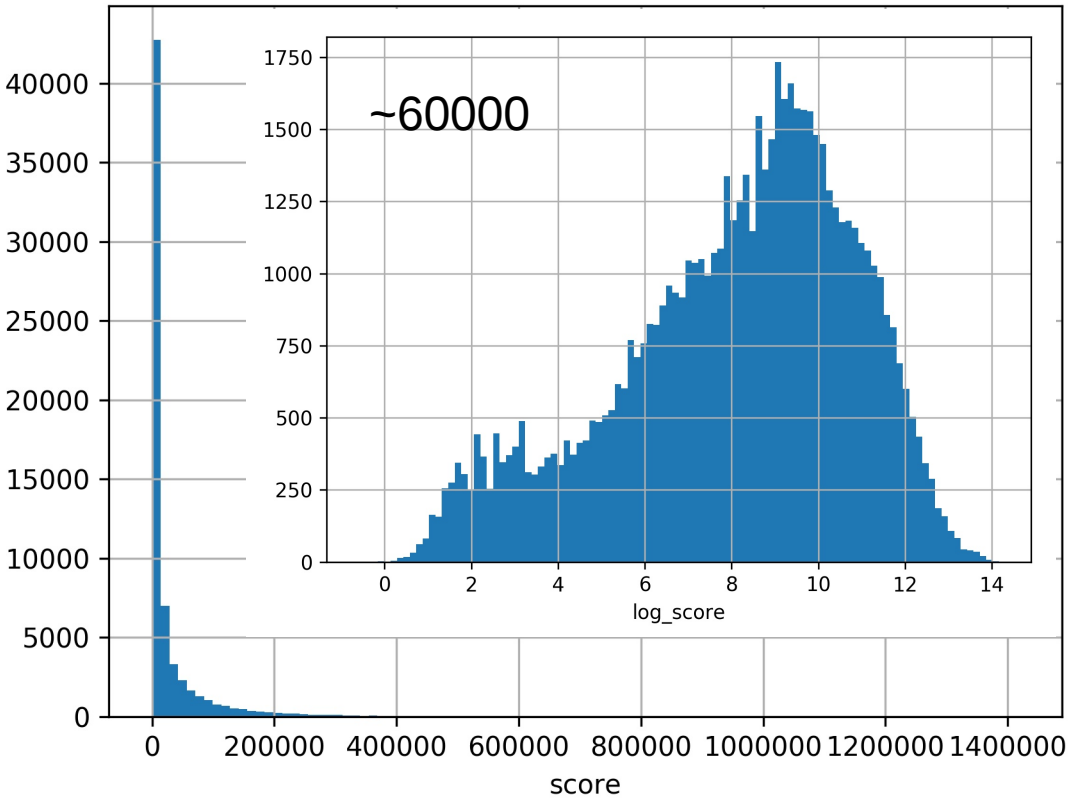
Based on 18 mutants experimentally validated and with known distribution of reads across Bin1-4 and Bkgd



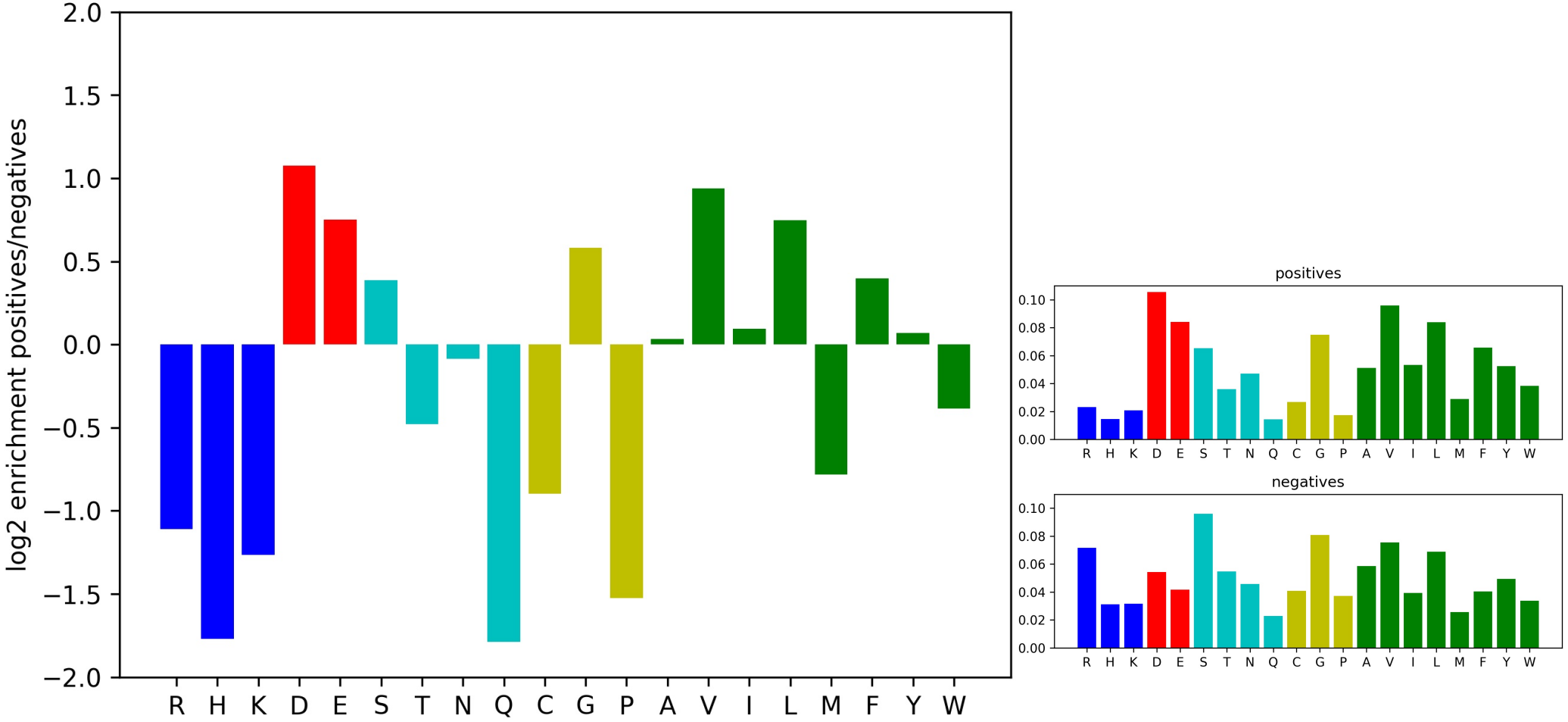


Scores

<5 reads (all bins) → discarded
Positives: bins(3,4) > 2*bins(0,1,2) ~20000
Negatives: bins(0) > 2*bins(2,3,4) ~20000



Enrichment in aa content



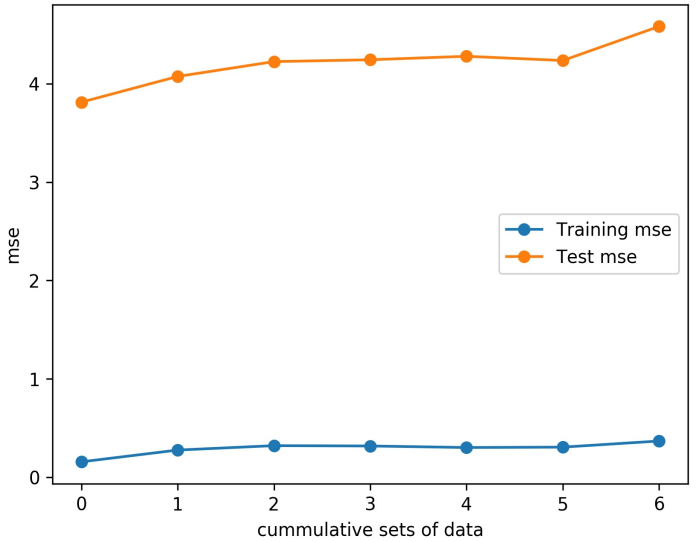
Features for ML

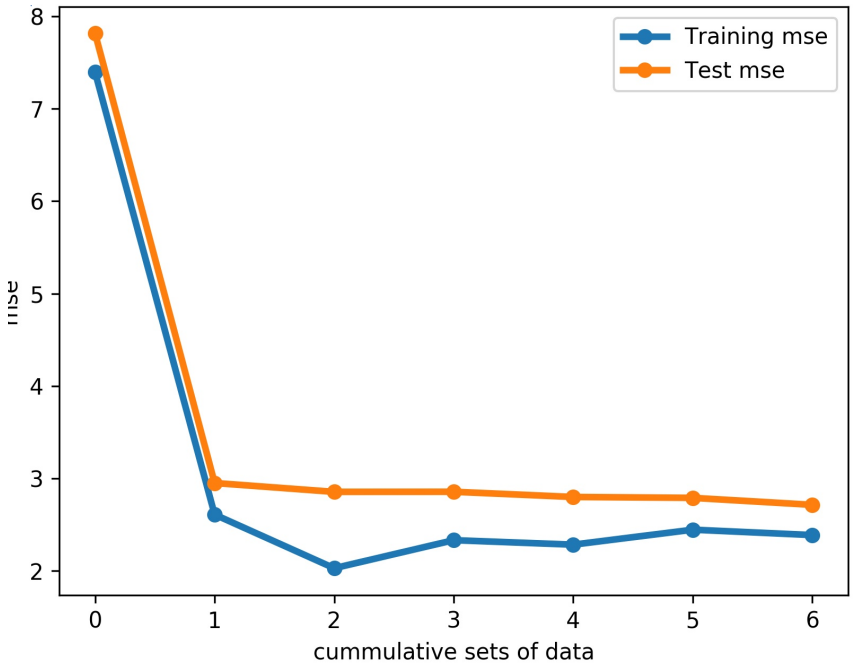
- AA seq
- AA hydrophobicity
- AA Charge
- AA Secondary Structure
- AA Disorder

Developing deep learning models

Regression with a Dense model

- _____
- Layer (type) Output Shape Param #
- ===
- dense_1 (Dense) (None, 500) 30500
- _____
- dropout_1 (Dropout) (None, 500) 0
- _____
- dense_2 (Dense) (None, 250) 125250
- _____
- dropout_2 (Dropout) (None, 250) 0
- _____
- dense_3 (Dense) (None, 60) 15060
- _____
- dropout_3 (Dropout) (None, 60) 0
- _____
- dense_4 (Dense) (None, 1) 61
- ===
- Total params: 170,871
- Trainable params: 170,871
- Non-trainable params: 0
- _____
-
-

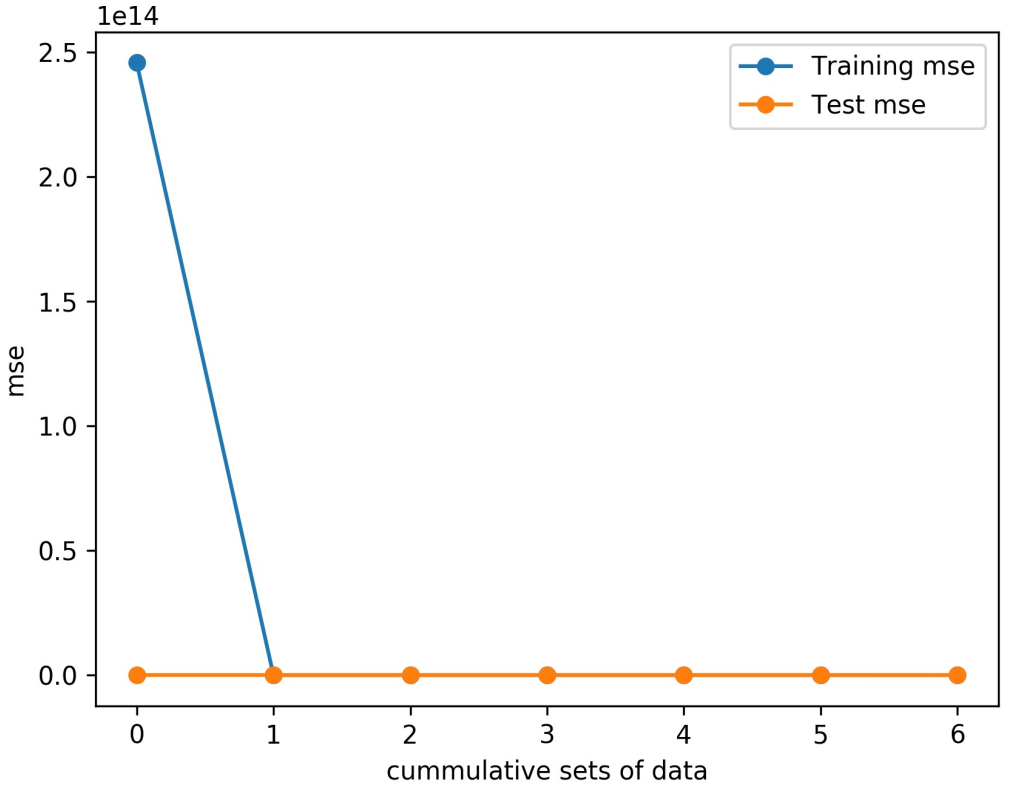


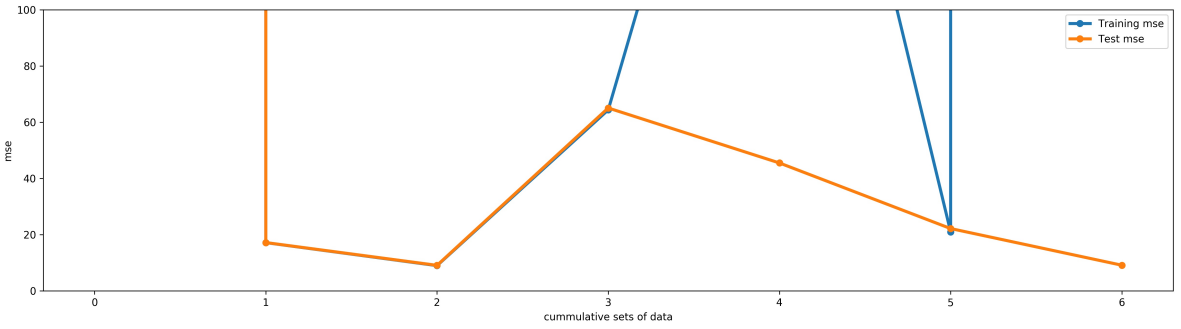


Regression with Convolutional model

Layer (type)	Output Shape	Param #
input_1 (InputLayer)	(None, 60, 1)	0
conv1d_1 (Conv1D)	(None, 60, 100)	1100
conv1d_2 (Conv1D)	(None, 60, 100)	100100
dropout_1 (Dropout)	(None, 60, 100)	0
max_pooling1d_1 (MaxPooling1	(None, 30, 100)	0
dropout_2 (Dropout)	(None, 30, 100)	0
flatten_1 (Flatten)	(None, 3000)	0
dense_1 (Dense)	(None, 500)	1500500
batch_normalization_1 (Batch	(None, 500)	2000
dense_2 (Dense)	(None, 100)	50100
batch_normalization_2 (Batch	(None, 100)	400
dense_3 (Dense)	(None, 50)	5050
batch_normalization_3 (Batch	(None, 50)	200
dense_4 (Dense)	(None, 1)	51

Total params: 1,659,501
Trainable params: 1,658,201
Non-trainable params: 1,300





Classification using convolutional or recurrent models

- Sigmoid activation in output layer
- Loss = binary crossentropy
- Best. accuracy = 0.68 ± 0.06 (benchmark ~50%)

Hyper-parameters tunning

- GridSearchCV(sklearn)
- batch_size = [64, 128, 256]
- epochs = [10]
- kernel_init = ['uniform', ‘normal’]
- pDropout = [0.3-0.5]
- Convolutions2D_shape1= [3,2]
- learning_rate = [0.01, 0.0001] #0.1, 0.01, 0.001]
- Optimizer = [‘RMSprop’, ‘Adam’]
- decay = [1e-4, 1e-6]

Questions

- Stacking ohe-AA and other features into a nD tensor?
- Keeping aa-Ids and other features separately?
- Working with 1D or nD tensors? This for convolutional models and RNN.
- Embedding layers?

Library design and construction

- **NNN** → 3 out of 64 (~5%) are stop codons... → short peptides rather than 30 residues long sequence
- **NNK** or **NNS** → > 3% stop codons
- **NNY** and RNN repeats (Y=primidines, R=purines) avoid Stop codon but do not encode for 2 amino-acids
- **SOLUTION**: Biasing the **ratios of nucleotides** at all three positions in the randomized codons.

- Codon Optimized Libraries

	A	C	G	T
0	0.26	0.26	0.24	0.21
1	0.38	0.19	0.17	0.22
2	0.00	0.46	0.34	0.16

	Ideal	Optimized
F	0,05	0,035
L	0,05	0,08
I	0,05	0,045
M	0,05	0,03
V	0,05	0,065
S	0,05	0,08
P	0,05	0,04
T	0,05	0,05
A	0,05	0,045
Y	0,05	0,04
H	0,05	0,04
Q	0,05	0,03
N	0,05	0,055
K	0,05	0,04
D	0,05	0,045
E	0,05	0,035
C	0,05	0,03
W	0,05	0,025
R	0,05	0,085
G	0,05	0,06
STOP	0	0,03

- Optimized for Equal Ratios
- Optimized for Disordered regions

Searching nucleotide composition space

- Space of all possible sets of 3 nucleotide mixture $X_1X_2X_3$
- Each point in nucleotide space specifies a list of probabilities for the codons and therefore values for aminoacids and stop codons frequencies.
- Difference between ***target*** values and the ***encoded*** amino acid ratios correspond to a cost that we seek to minimize

SPACE

$$C = \sum_{i=1}^{21} \left(t_i - e_i \right)^2$$

Surface, where the deepest valley contains the nucleotide composition that most closely match the design target.

Complete enumeration of the space

Possible values for each dimension of space

Number of possible compositions

$$N = \frac{\sum_{i=1}^{n+1} i(i+1)}{2}$$

1% resolution = 100 possible values

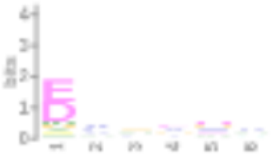
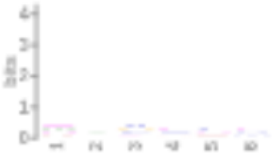
~174000 compositions for one nucleotide
~10^{15} possible 3-based combinations

~30 years to test all possible combinations

Scatter plot – design vs experimental

MEME

Discovered Motifs

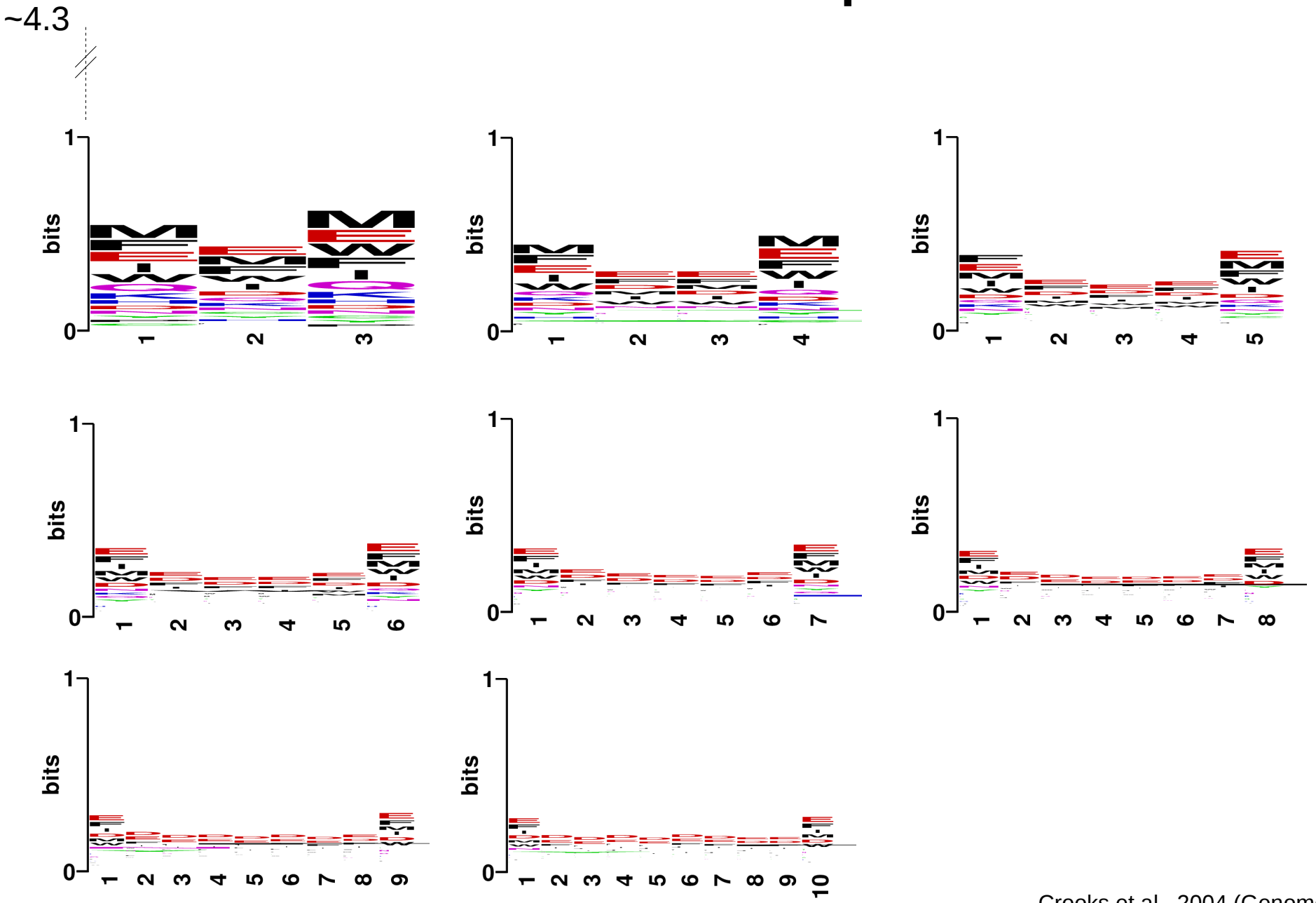
	Logo	E-value	Sites	Width
1.		3.1e+612	19086	7
2.		1.3e+1017	19086	6
3.		1.2e+1299	19086	6

Other Settings

Motif Site Distribution	OOPS: Exactly one site per sequenc
Objective Function	E-value of product of p-values
Starting Point Function	E-value of product of p-values
Site Strand Handling	This alphabet only has one strand
Maximum Number of Motifs	3
Motif E-value Threshold	no limit
Minimum Motif Width	6
Maximum Motif Width	29
Minimum Sites per Motif	19086
Maximum Sites per Motif	19086
Bias on Number of Sites	0.8
Sequence Prior	Dirichlet Mixture
Sequence Prior Source	prior30.plib
Sequence Prior Strength	intrinsic strength
EM Starting Point Source	From substrings in input sequences
EM Starting Point Map Type	Point Accepted Mutation
EM Starting Point Fuzz	120
EM Maximum Iterations	50
EM Improvement Threshold	0.00001
Maximum Search Size	100000
Maximum Number of Sites for E-values	1000
Trim Gap Open Cost	11
Trim Gap Extend Cost	1
End Gap Treatment	Same cost as other gaps

Bailey & Elkan, 1994 (PSICISMB)

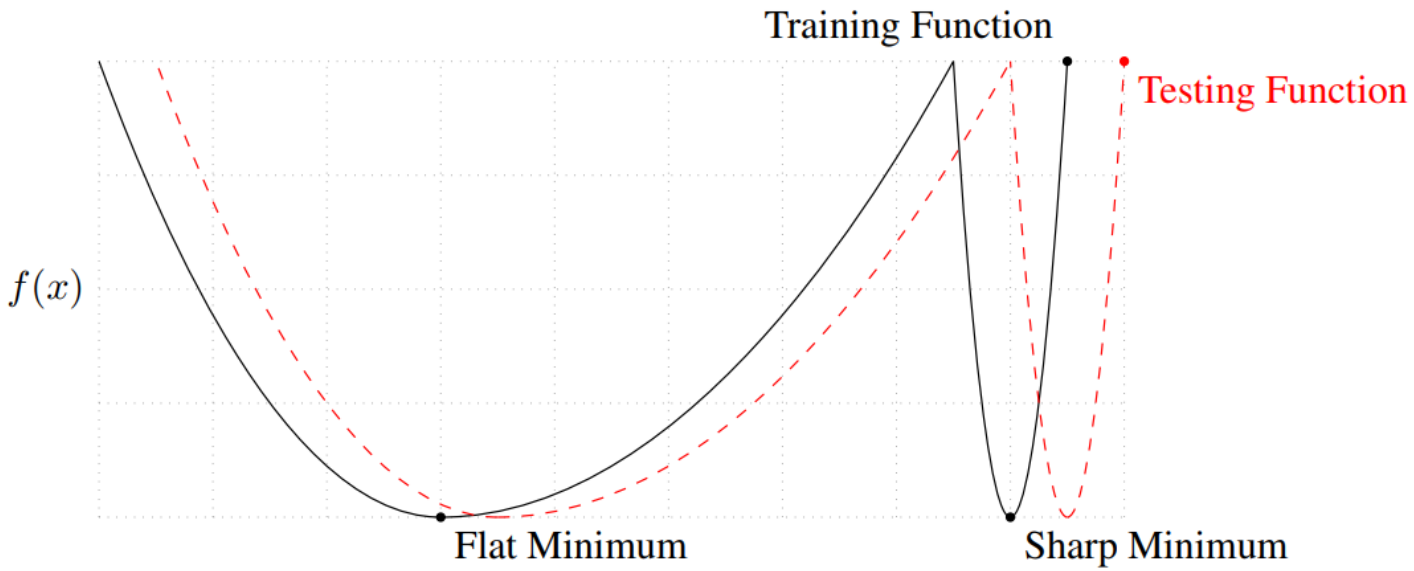
Gibbs sampler



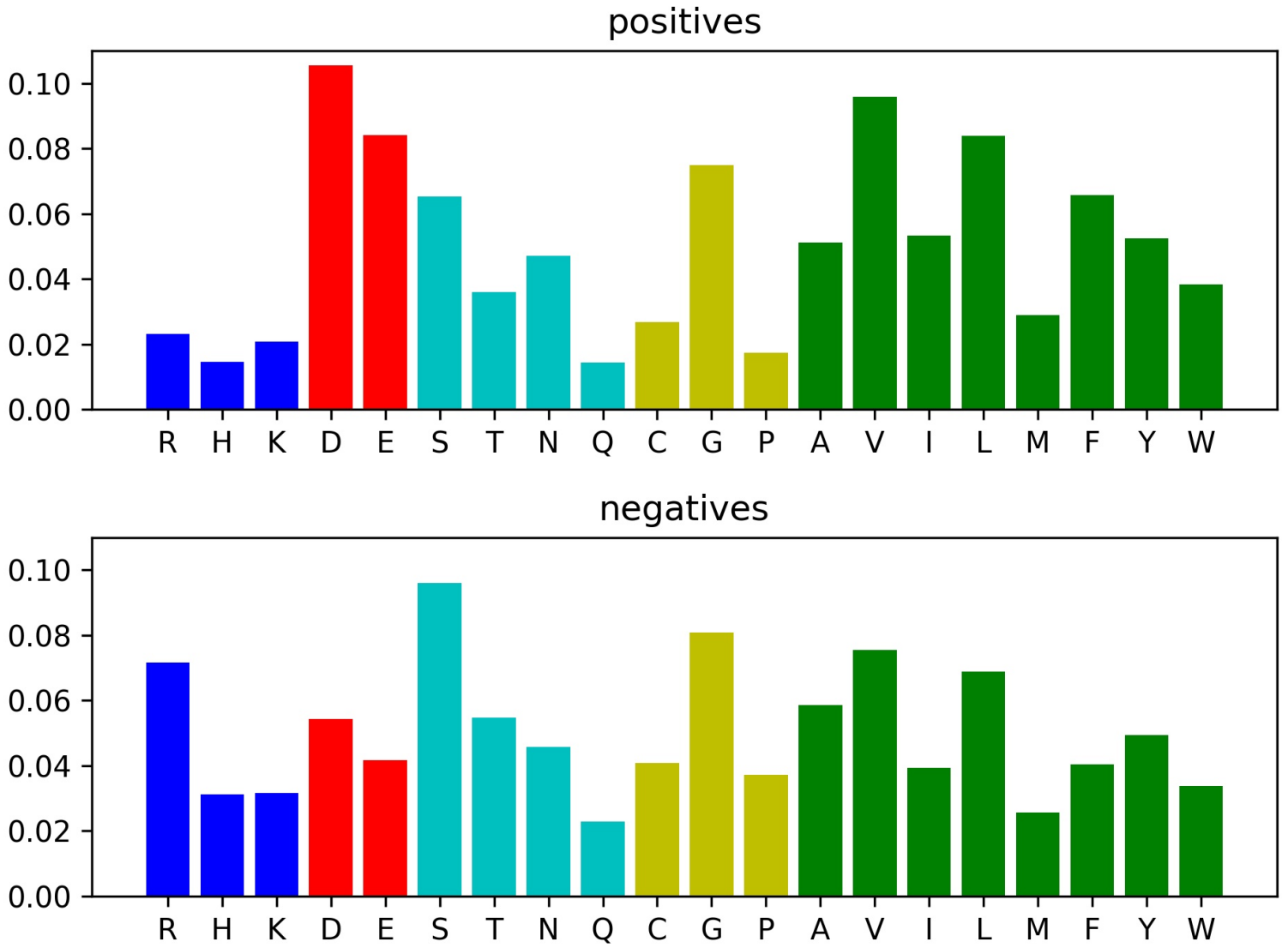
Crooks et al., 2004 (Genome research)

Batch size, not just a matter of learning speed?

0.41 ± 0.29 {'batch_size': **64**, 'decay': 1e-06, 'epochs': 10, 'init': 'uniform', 'k1': 3, 'lr': 0.001, 'pDrop': 0.4}
0.68 ± 0.07 {'batch_size': **128**, 'decay': 1e-06, 'epochs': 10, 'init': 'uniform', 'k1': 3, 'lr': 0.001, 'pDrop': 0.4}
0.55 ± 0.15 {'batch_size': **256**, 'decay': 1e-06, 'epochs': 10, 'init': 'uniform', 'k1': 3, 'lr': 0.001, 'pDrop': 0.4}

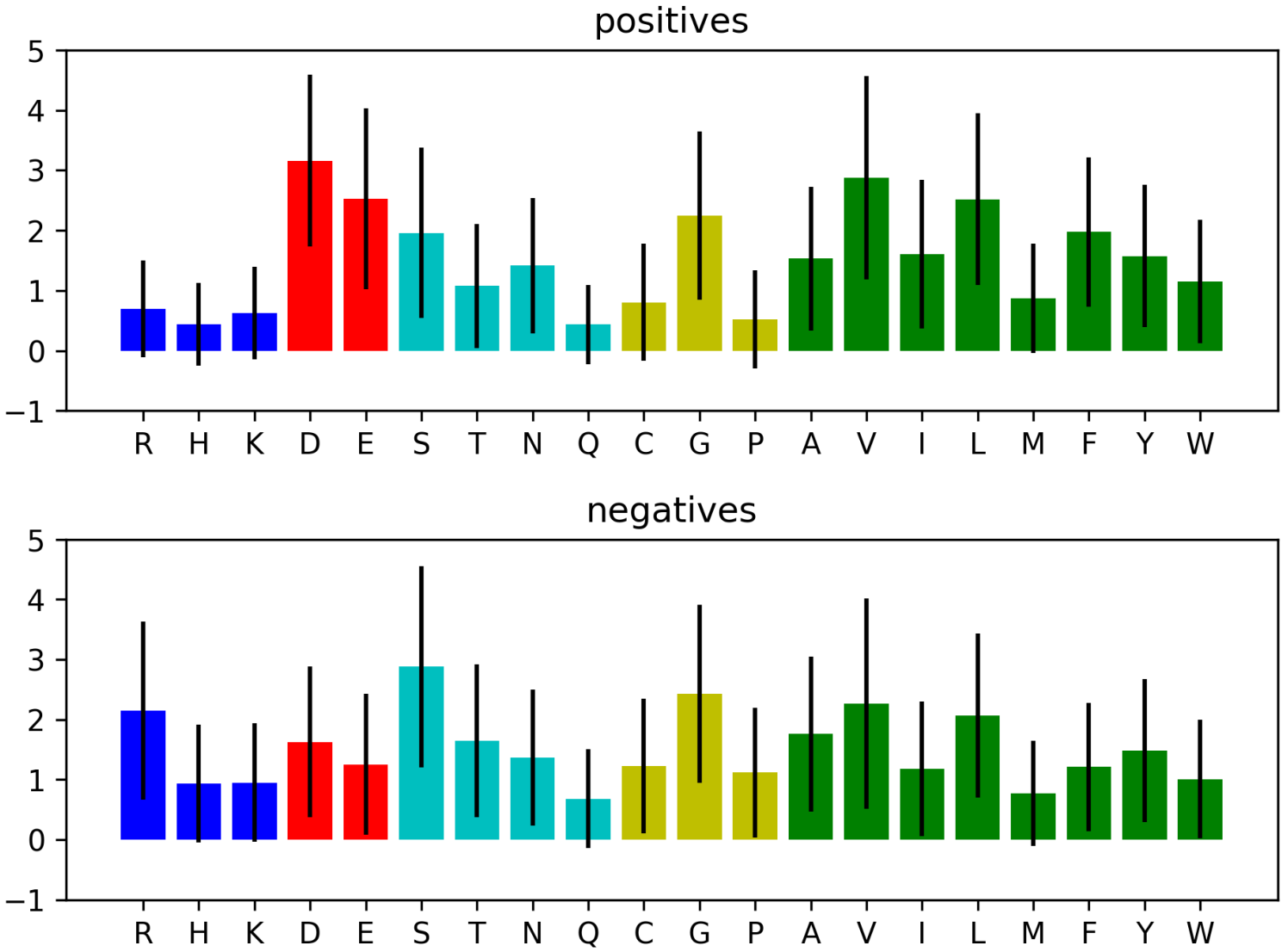


Amino-acid content of positivie and negative sets



Unique sequences

Average aa content per sequence



Unique sequences