Metropolis Monte Carlo sampling

Gibbs Clustering

Morten Nielsen

Department of Health Technology
DTU, Denmark

Metropolis Monte Carlo

- What to do if you cannot do the math or your error function cannot be differentiated?

$$\frac{\partial E}{\partial w_i} = ?$$

Example: Estimating Π by Independent Monte-Carlo Samples

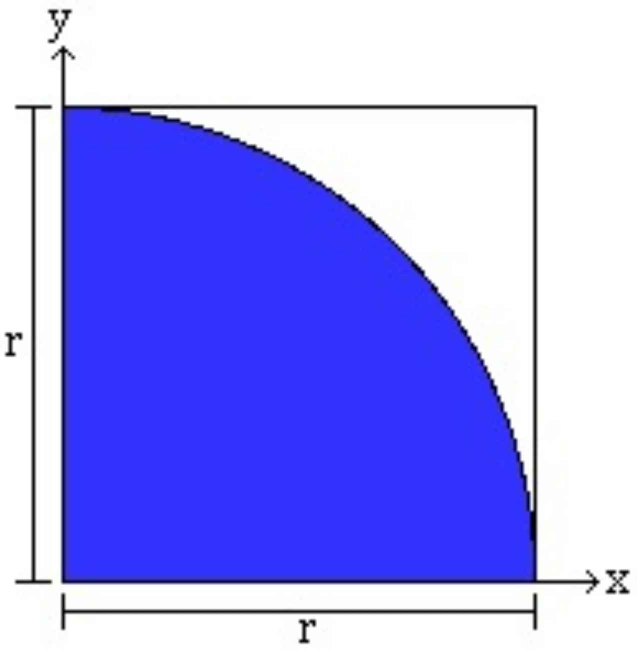
Suppose we throw darts randomly (and uniformly) at the square:

$$\frac{\#\ \text{darts hitting shaded area}}{\#\ \text{darts hitting inside square}} = \frac{\frac{1}{4}\pi r^2}{r^2} = \frac{1}{4}\pi$$

or

$$\pi = 4\,\frac{\#\ \text{darts hitting shaded area}}{\#\ \text{darts hitting inside square}}$$

```
Algorithm:
For i=[1..ntrials]
    x = (random# in [0..r])
    y = (random# in [0..r])
    distance = sqrt (x^2 + y^2)
    if distance ≤ r
        hits++
End
Output: $\frac{4 \times hits}{ntrials}$
```

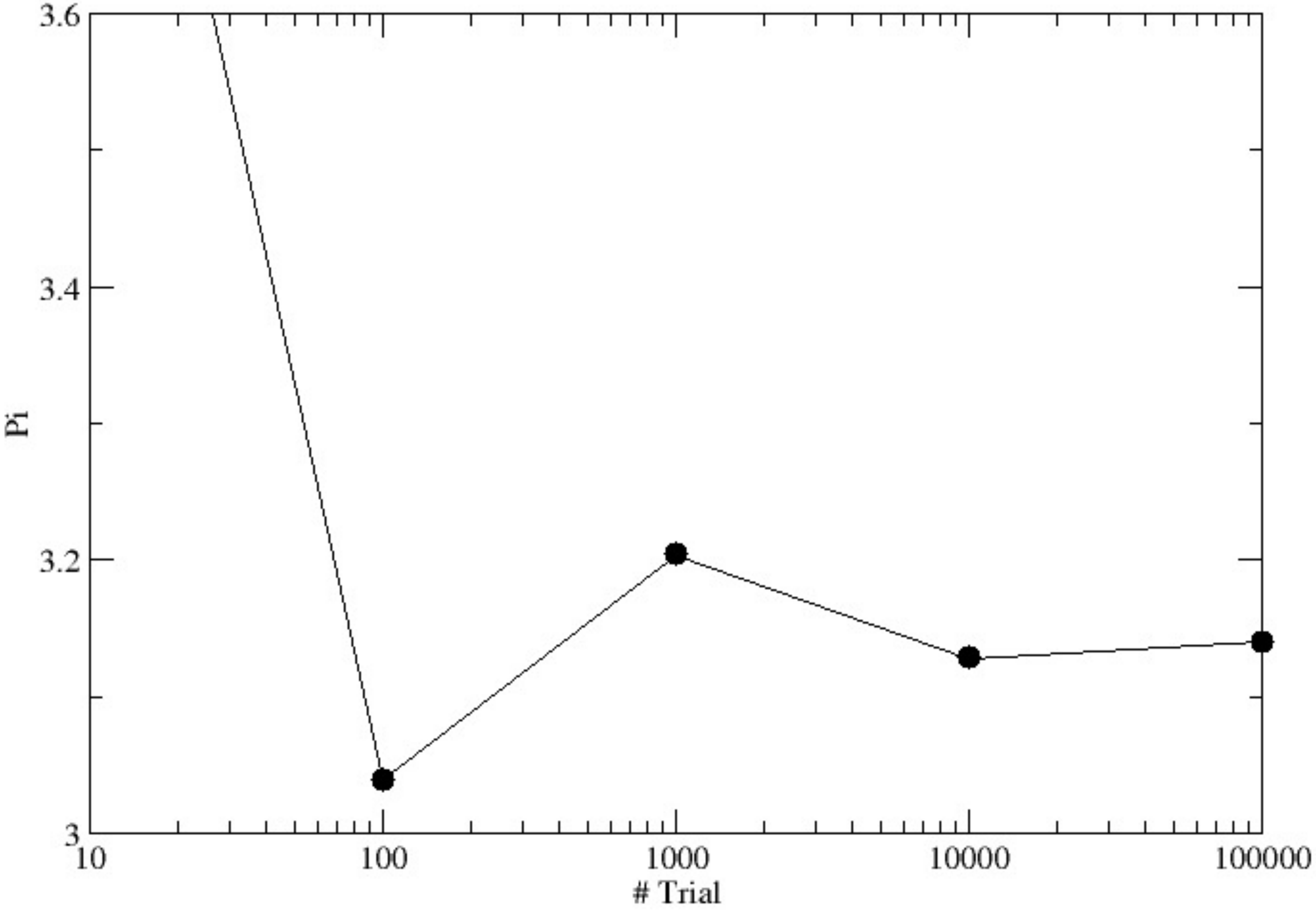


http://www.chem.unl.edu/zeng/joy/mclab/mcintro.html

Adapted from course slides by Craig Douglas

Estimating Π

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS



Monte Carlo

Because of their reliance on repeated computation of random or pseudo-random numbers, Monte Carlo methods are most suited to calculation by a computer. Monte Carlo methods tend to be used when it is unfeasible or impossible to compute an exact result with a deterministic algorithm

Or when you are too stupid to do the math yourself?

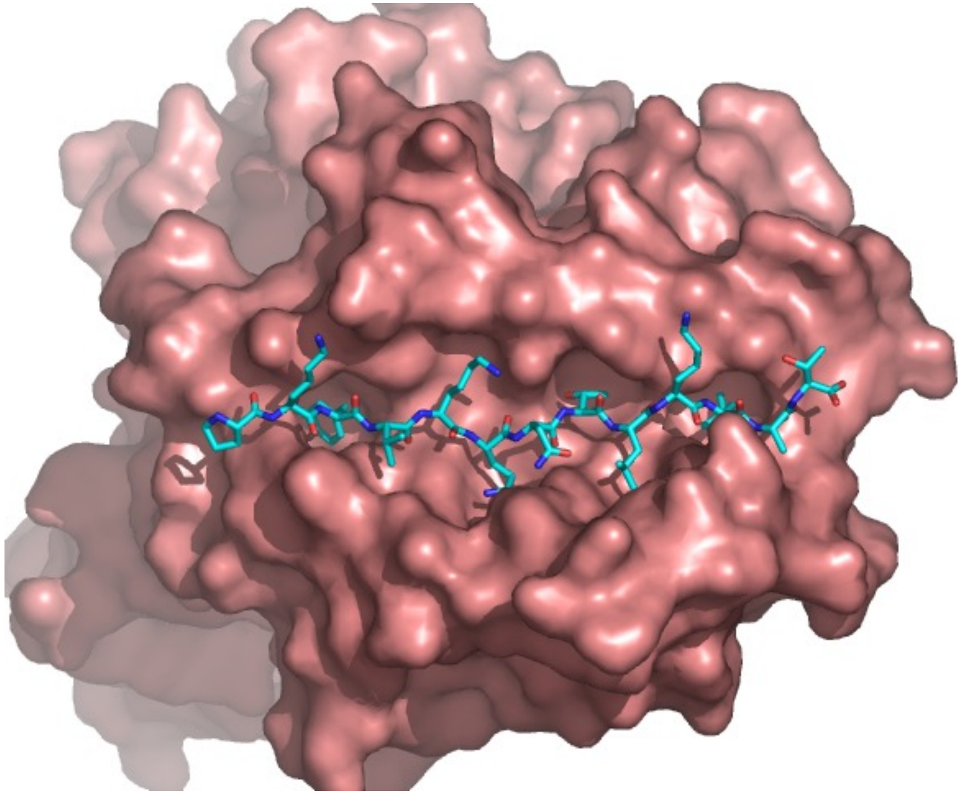
$$E = f(x)$$

$$dE = E_1 - E_0$$

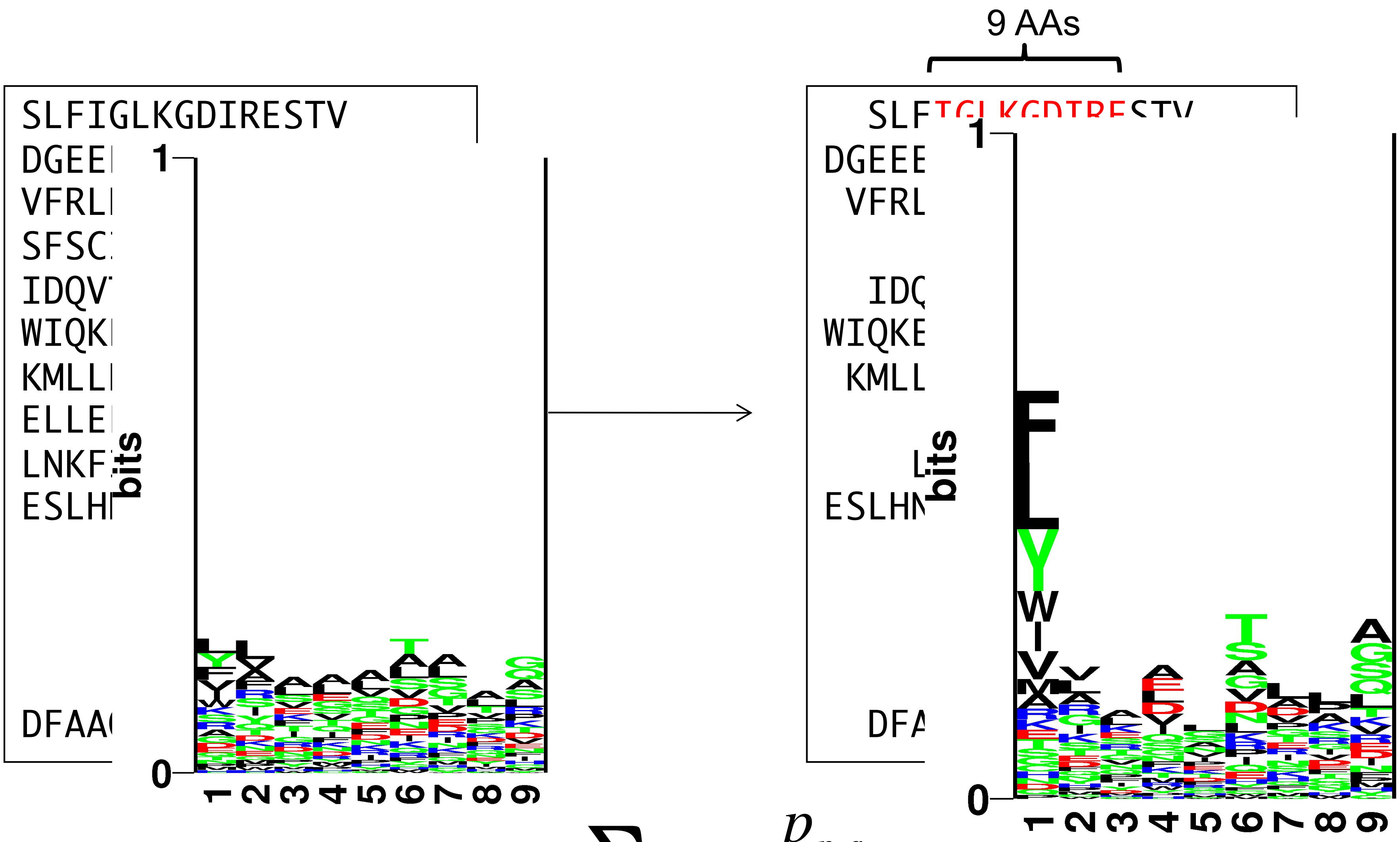
$$P(accept) = \min(1, e^{-dE/T})$$

Class II MHC binding

- MHC class II binds peptides in the class II antigen presentation pathway
- Binds peptides of length 9-18 (even whole proteins can bind!)
- Binding cleft is open
- Binding core is 9 aa



Conventional Gibbs sampling MHC class II binding



$$E = \sum_{peptides} \log \frac{p_{p,a}}{q_a}$$

Gibbs sampling - sequence alignment

Why sampling?

50 sequences 12 amino acids long

try all possible combinations with a 9-mer overlap

↓

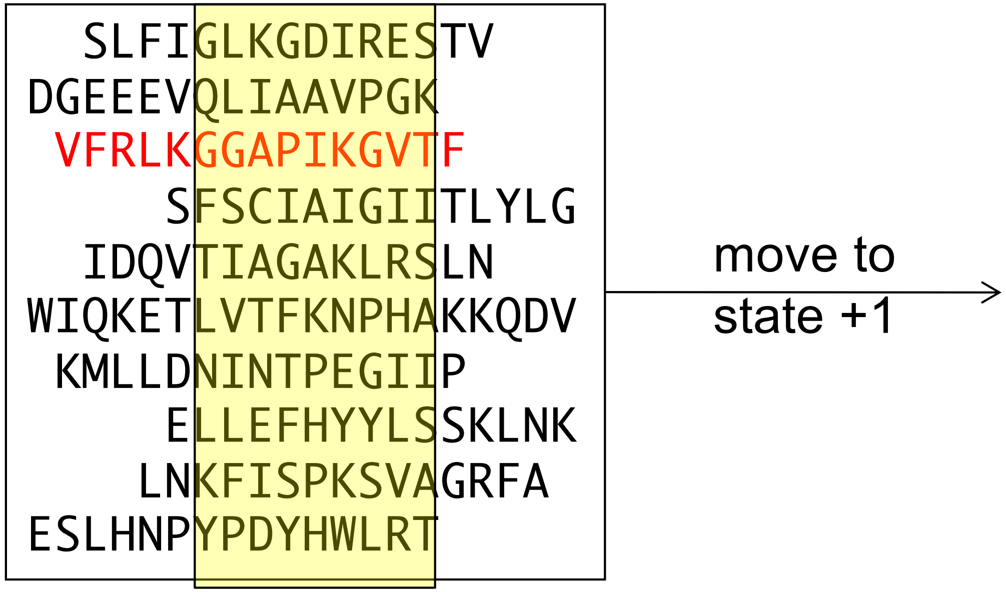
$4^{50} \sim 10^{30}$ possible combinations

...computationally unfeasible

```
  SLFIGLKGDIRESTV
DGEEEVQLIAAVPGK
 VFRLKGGAPIKGVTF
     SFSCIAIGIITLYLG
  IDQVTIAGAKLRSLN
WIQKETLVTFKNPHAKKQDV
 KMLLDNINTPEGIIP
     ELLEFHYYLSSKLNK
    LNKFISPKSVAGRFA
ESLHNPYPDYHWLRT
        ...
        ...
        ...
        ...
  DFAAQVDYPSTGLY
```

Gibbs sampling - sequence alignment

State transition



```
  SLFIGLKGDIRESTV
DGEEEVQLIAAVPGK
 VFRLKGGAPIKGVTF
     SFSCIAIGIITLYLG
  IDQVTIAGAKLRSLN
WIQKETLVTFKNPHAKKQDV
 KMLLDNINTPEGIIP
     ELLEFHYYLSSKLNK
    LNKFISPKSVAGRFA
ESLHNPYPDYHWLRT
```

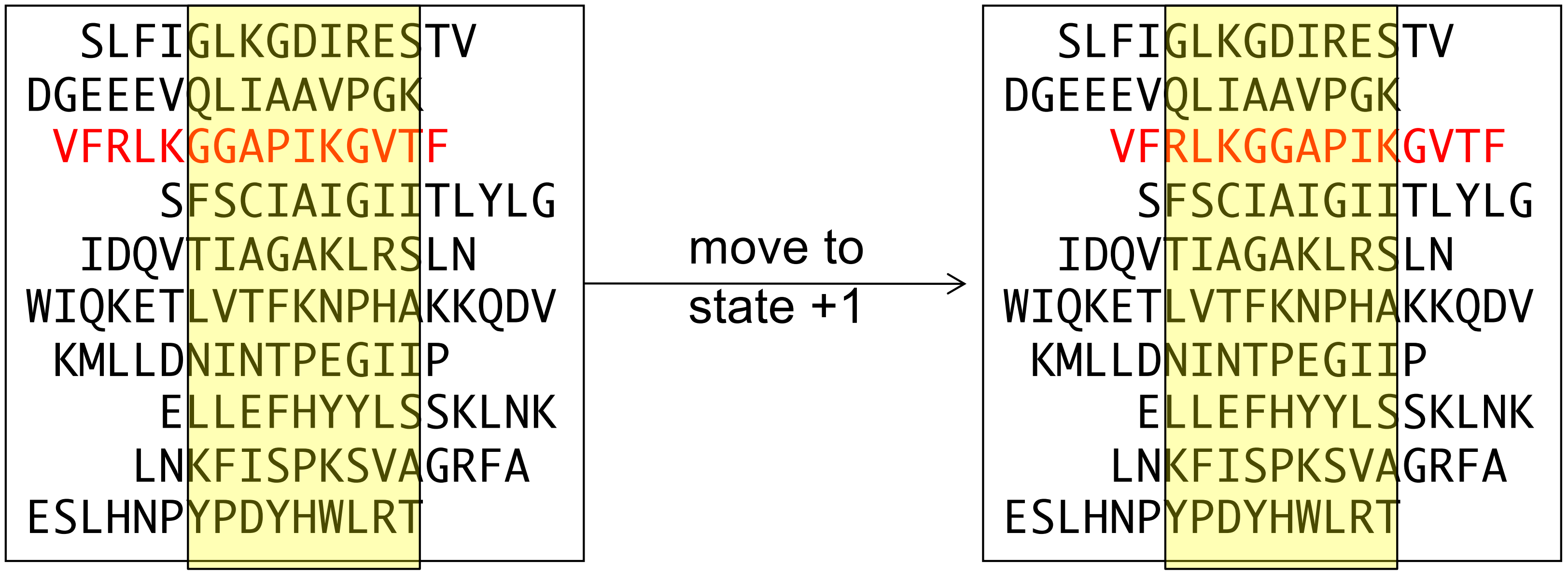
move to state +1

$$E = \sum_{peptides} \log \frac{p_{p,a}}{q_a}$$

$$dE = E_i - E_{i-1}$$

Gibbs sampling - sequence alignment

State transition



```
  SLFIGLKGDIRESTV
DGEEEVQLIAAVPGK
 VFRLKGGAPIKGVTF
     SFSCIAIGIITLYLG
  IDQVTIAGAKLRSLN
WIQKETLVTFKNPHAKKQDV
 KMLLDNINTPEGIIP
     ELLEFHYYLSSKLNK
    LNKFISPKSVAGRFA
ESLHNPYPDYHWLRT
```

move to state +1

```
  SLFIGLKGDIRESTV
DGEEEVQLIAAVPGK
    VFRLKGGAPIKGVTF
     SFSCIAIGIITLYLG
  IDQVTIAGAKLRSLN
WIQKETLVTFKNPHAKKQDV
 KMLLDNINTPEGIIP
     ELLEFHYYLSSKLNK
    LNKFISPKSVAGRFA
ESLHNPYPDYHWLRT
```

$$E = \sum_{peptides} \log \frac{p_{p,a}}{q_a}$$

$$dE = E_i - E_{i-1}$$

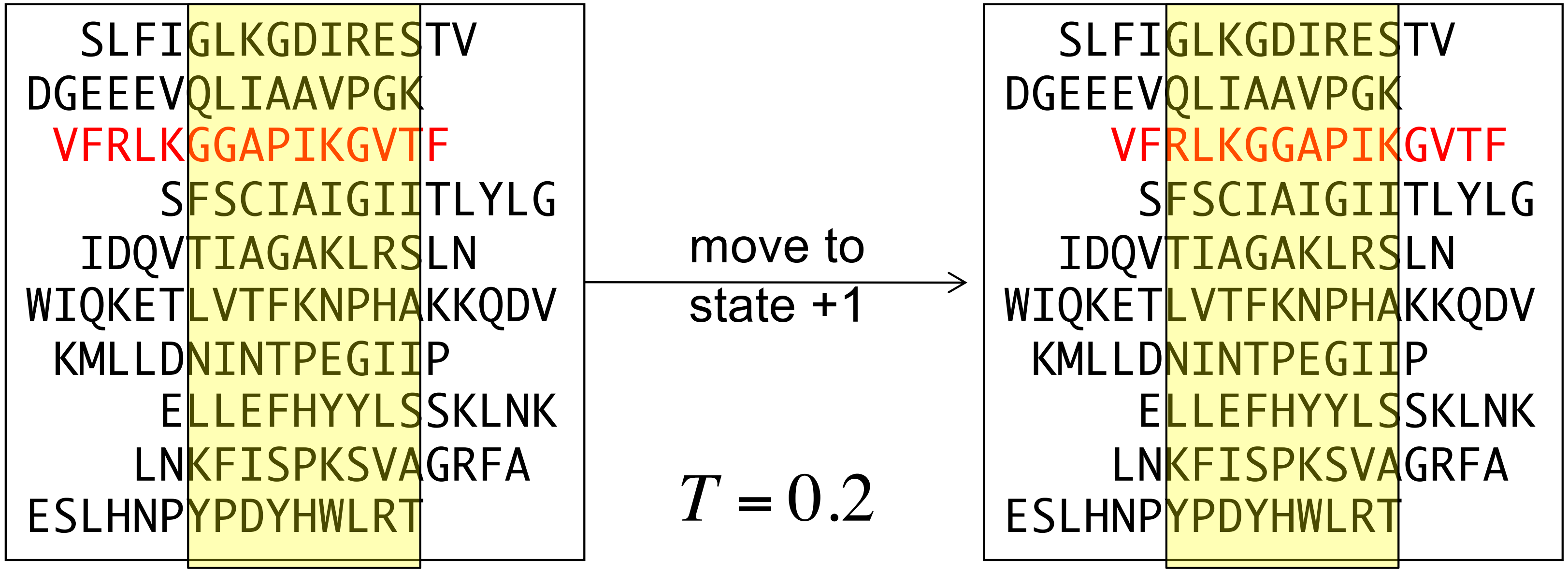
Accept or reject the move?

$$P = \min\left[1, \exp\left(\frac{dE}{T}\right)\right]$$

Note that the probability of going to the new state depends on the previous state only

Gibbs sampling - sequence alignment

Numerical example - 1



```
  SLFIGLKGDIRESTV
DGEEEVQLIAAVPGK
 VFRLKGGAPIKGVTF
     SFSCIAIGIITLYLG
  IDQVTIAGAKLRSLN
WIQKETLVTFKNPHAKKQDV
 KMLLDNINTPEGIIP
     ELLEFHYYLSSKLNK
    LNKFISPKSVAGRFA
ESLHNPYPDYHWLRT
```

move to state +1

```
  SLFIGLKGDIRESTV
DGEEEVQLIAAVPGK
    VFRLKGGAPIKGVTF
     SFSCIAIGIITLYLG
  IDQVTIAGAKLRSLN
WIQKETLVTFKNPHAKKQDV
 KMLLDNINTPEGIIP
     ELLEFHYYLSSKLNK
    LNKFISPKSVAGRFA
ESLHNPYPDYHWLRT
```

$$T = 0.2$$

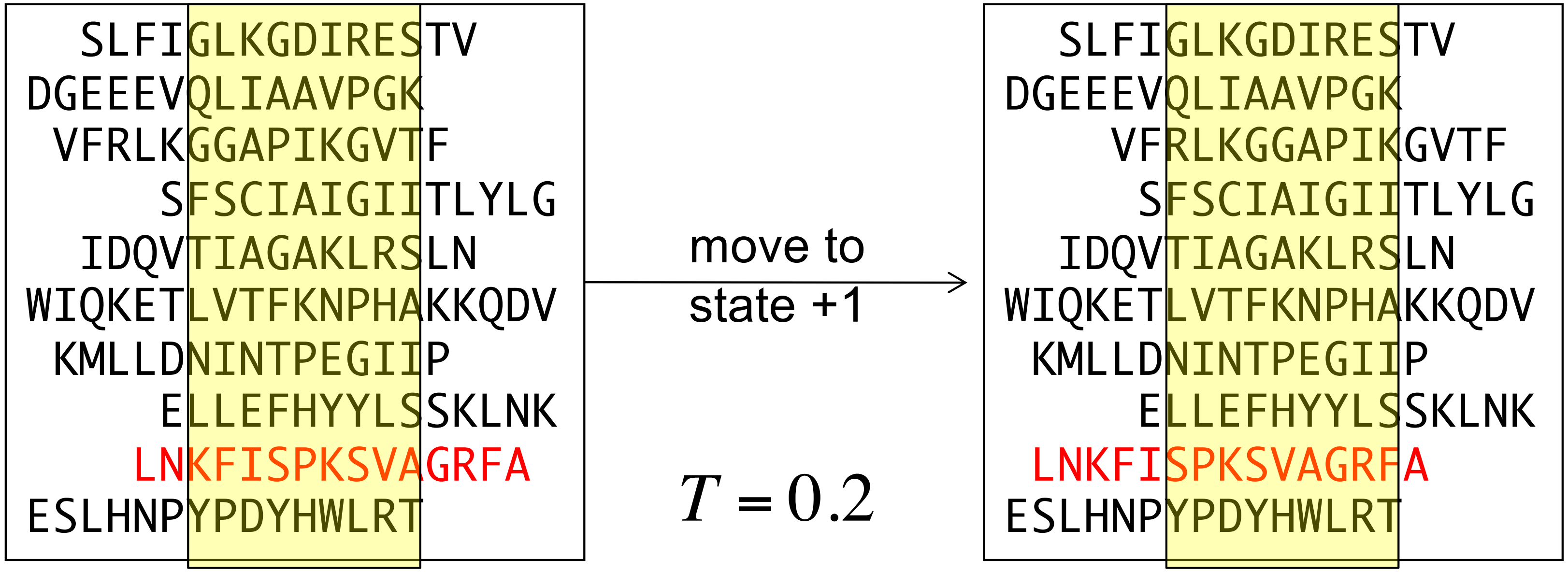
$$E_{i-1} = 2.44 \qquad E_i = 2.52$$

$$P = \min\left[1, \exp\left(\frac{0.08}{0.2}\right)\right] = \min[1\ ,\ 1.49] = 1$$

Accept move with Prob = 100%

Gibbs sampling - sequence alignment

Numerical example - 2



```
  SLFIGLKGDIRESTV
DGEEEVQLIAAVPGK
 VFRLKGGAPIKGVTF
     SFSCIAIGIITLYLG
  IDQVTIAGAKLRSLN
WIQKETLVTFKNPHAKKQDV
 KMLLDNINTPEGIIP
     ELLEFHYYLSSKLNK
    LNKFISPKSVAGRFA
ESLHNPYPDYHWLRT
```

move to state +1

$T = 0.2$

```
  SLFIGLKGDIRESTV
DGEEEVQLIAAVPGK
    VFRLKGGAPIKGVTF
     SFSCIAIGIITLYLG
  IDQVTIAGAKLRSLN
WIQKETLVTFKNPHAKKQDV
 KMLLDNINTPEGIIP
     ELLEFHYYLSSKLNK
 LNKFISPKSVAGRFA
ESLHNPYPDYHWLRT
```

$E_{i-1} = 2.44$

$E_i = 2.35$

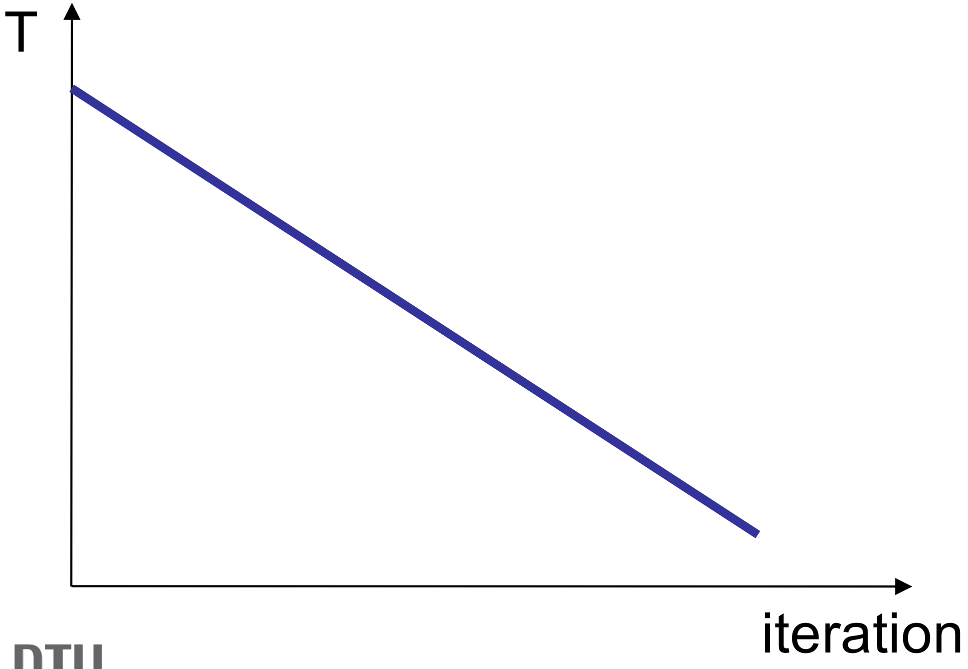
$$P = \min\left[1, \exp\left(\frac{-0.09}{0.2}\right)\right] = \min[1\ ,\ 0.638] = 0.638$$

Accept move with Prob = 63.8%

Gibbs sampling - sequence alignment

What is the MC temperature?

it’s a scalar decreased during the simulation

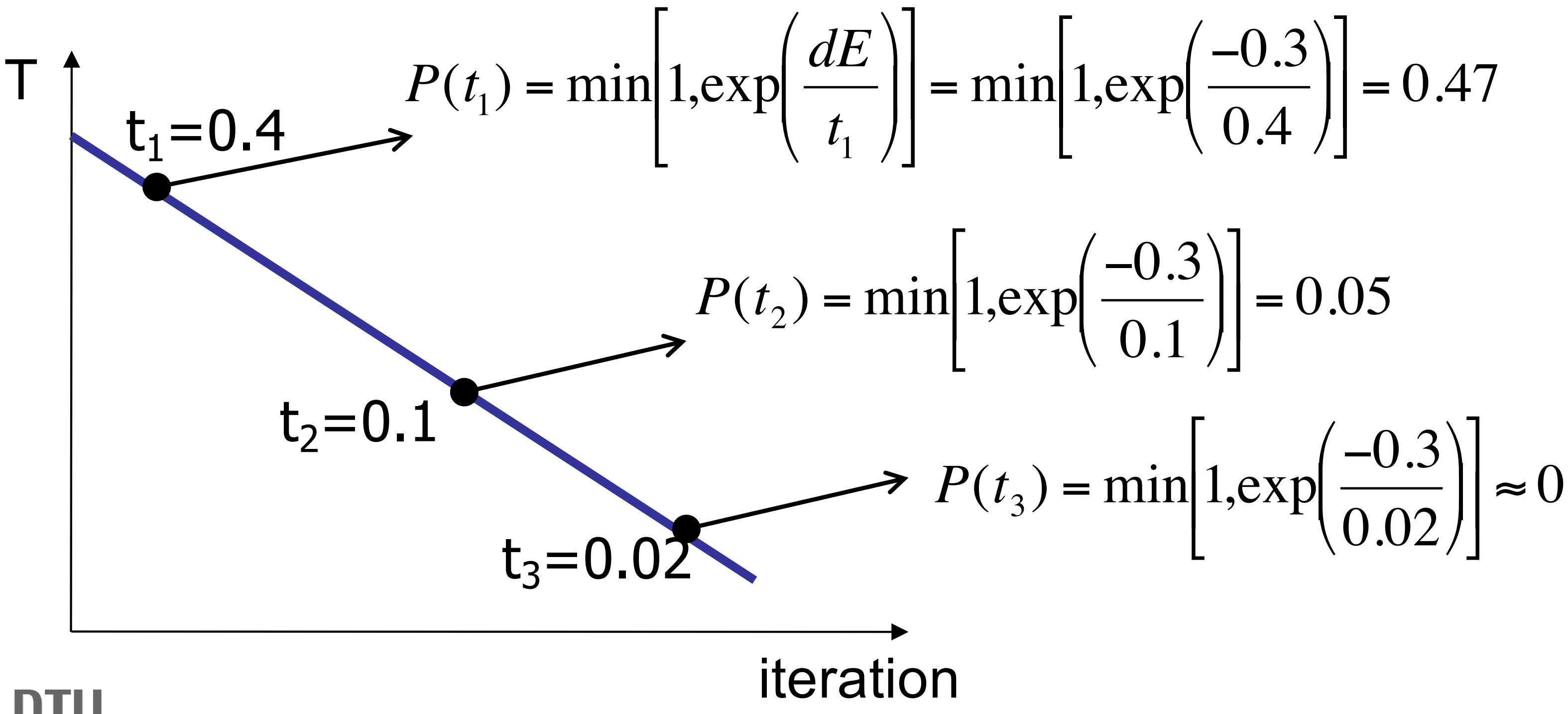


Gibbs sampling - sequence alignment

What is the MC temperature?

it’s a scalar decreased during the simulation

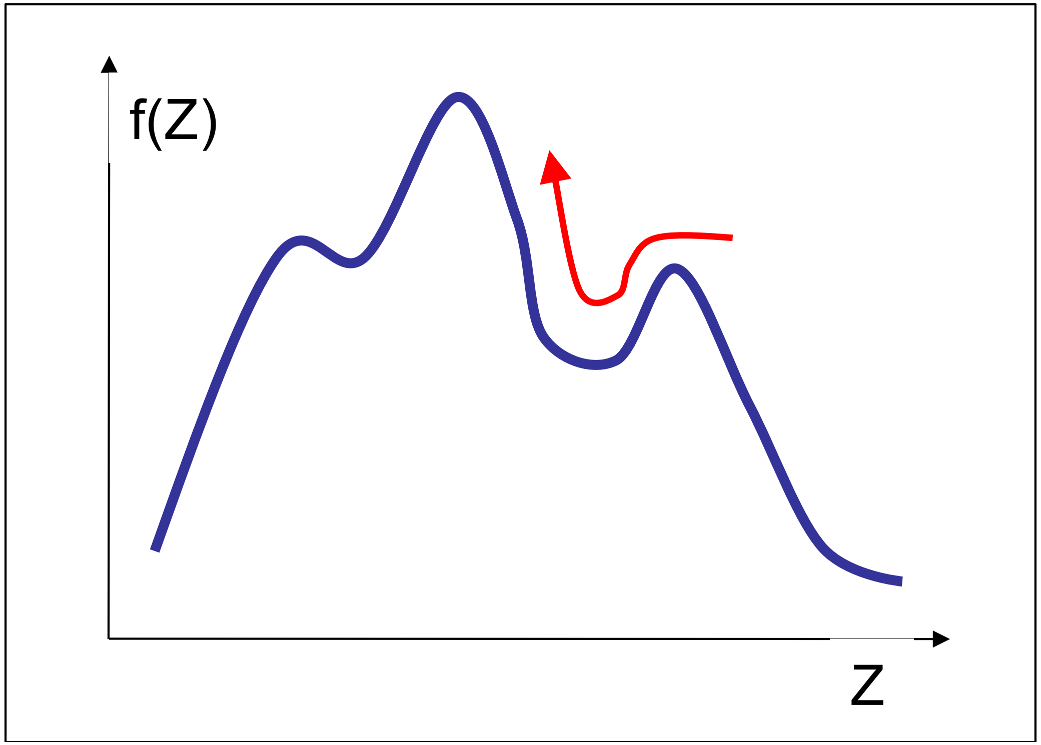
E.g. same **dE=-0.3** but at different temperatures



$$P(t_1) = \min\left[1, \exp\left(\frac{dE}{t_1}\right)\right] = \min\left[1, \exp\left(\frac{-0.3}{0.4}\right)\right] = 0.47$$

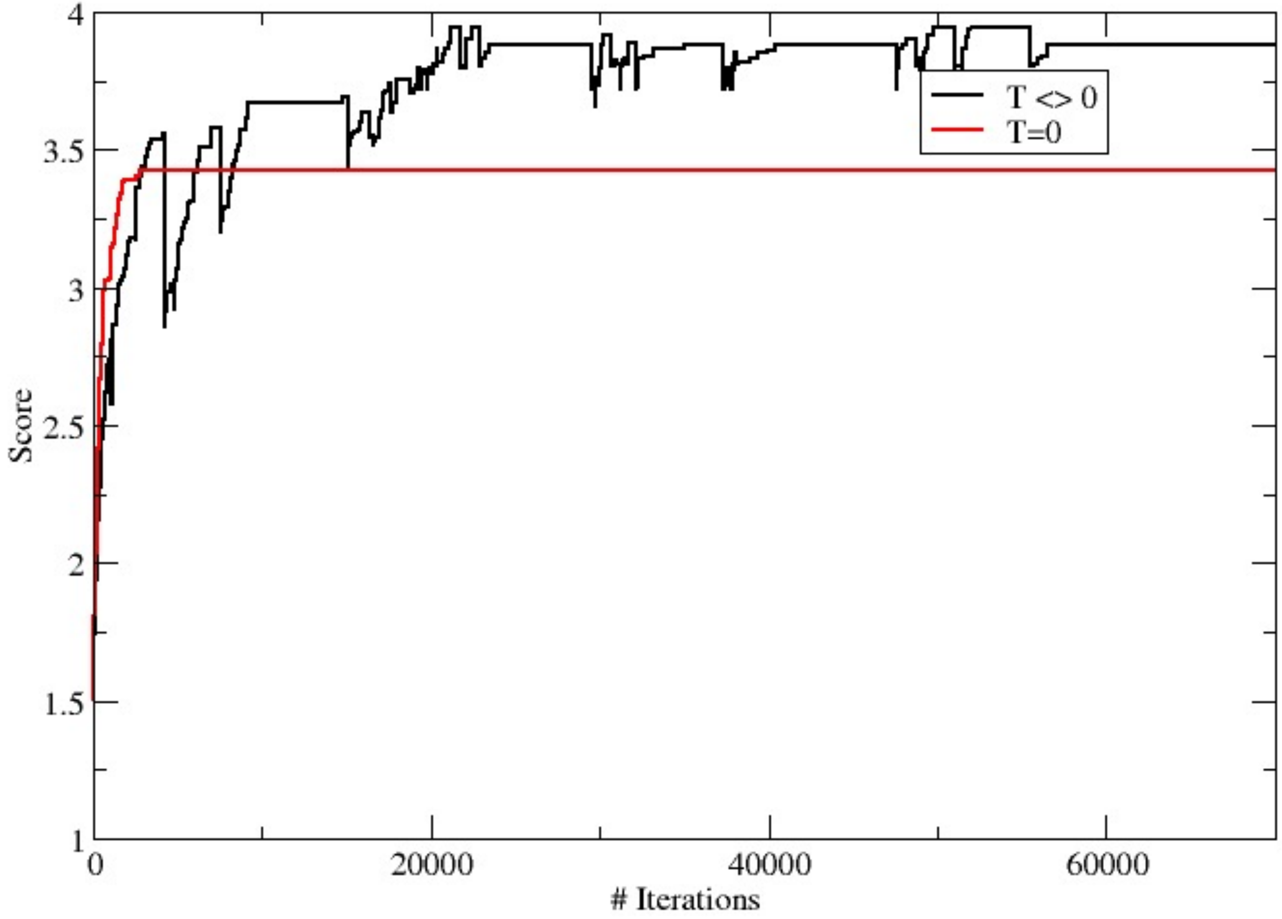
$$P(t_2) = \min\left[1, \exp\left(\frac{-0.3}{0.1}\right)\right] = 0.05$$

$$P(t_3) = \min\left[1, \exp\left(\frac{-0.3}{0.02}\right)\right] \approx 0$$

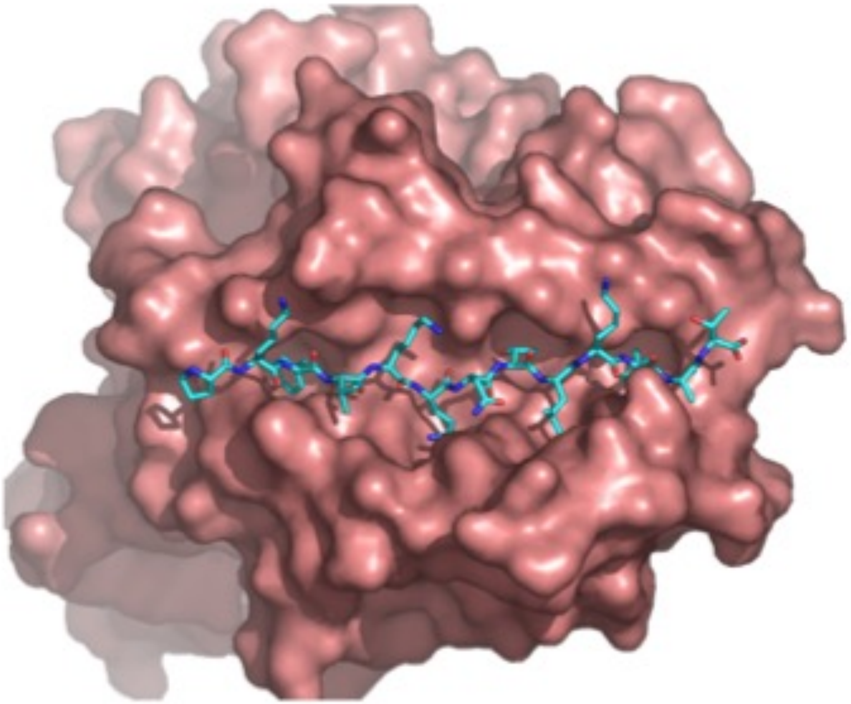


Move freely around states when the system is "warm", then cool it off to force it into a state of high fitness

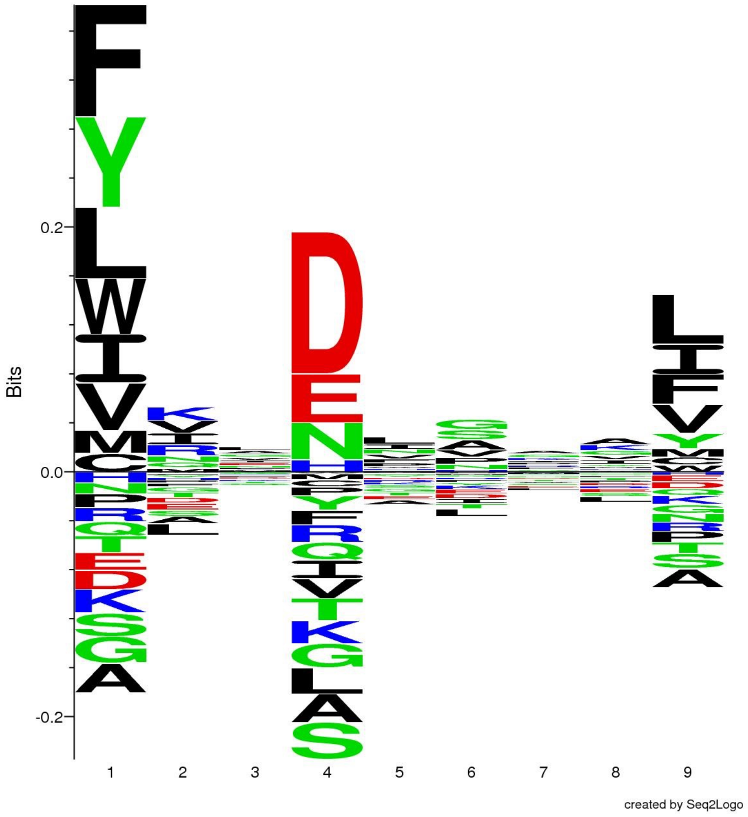
Local minima



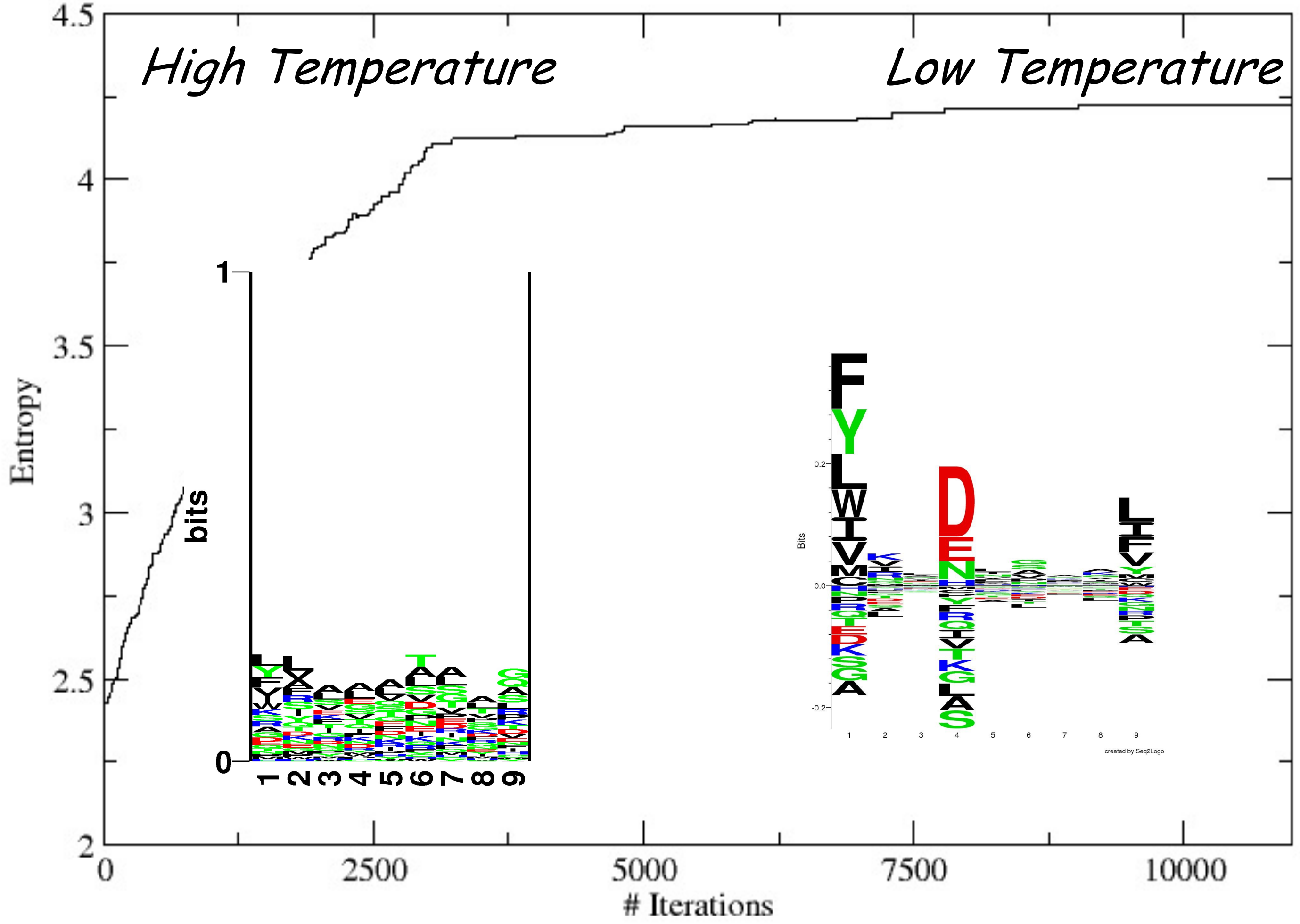
Does it work?



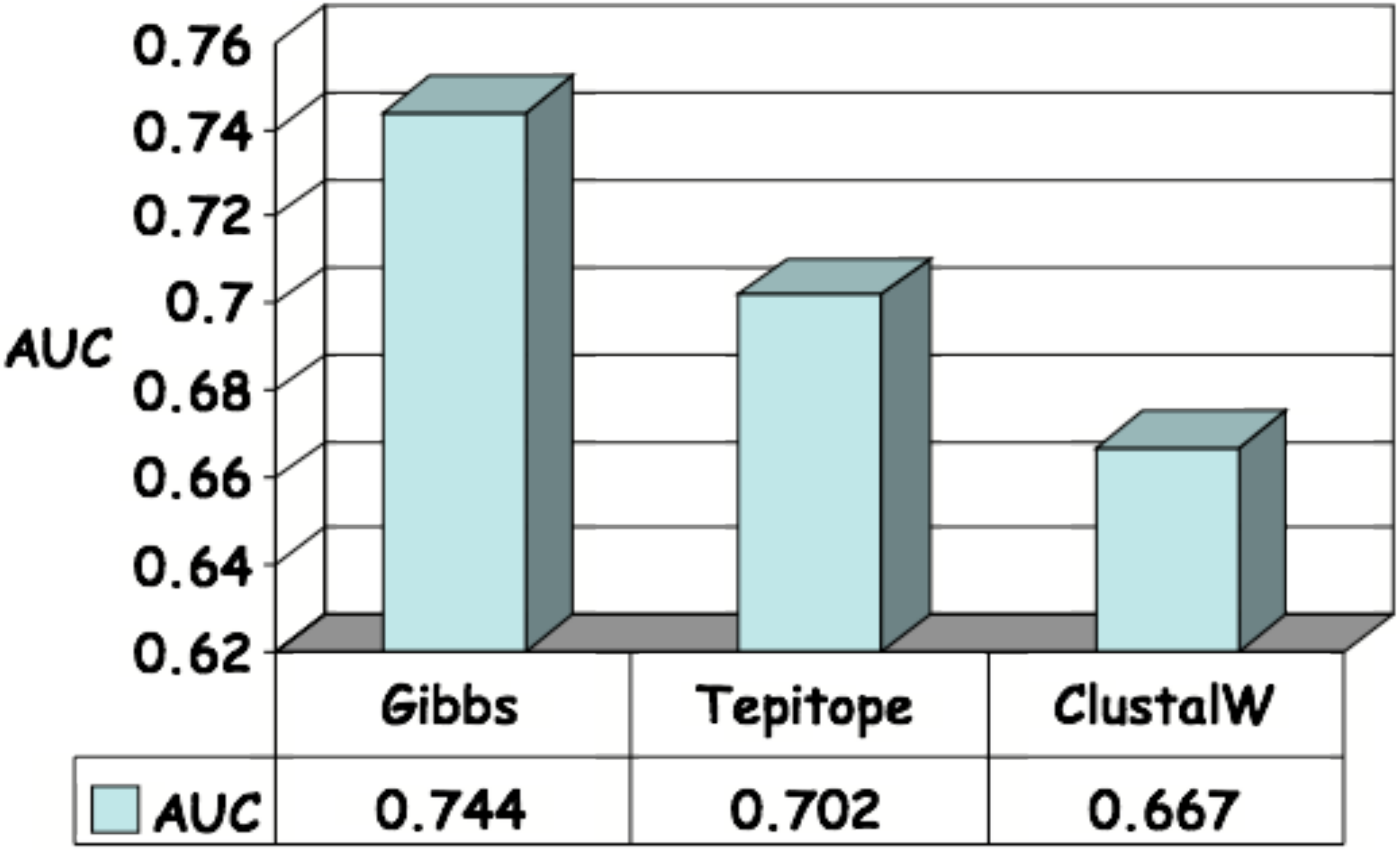
HLA-DRB3*01:01



It works

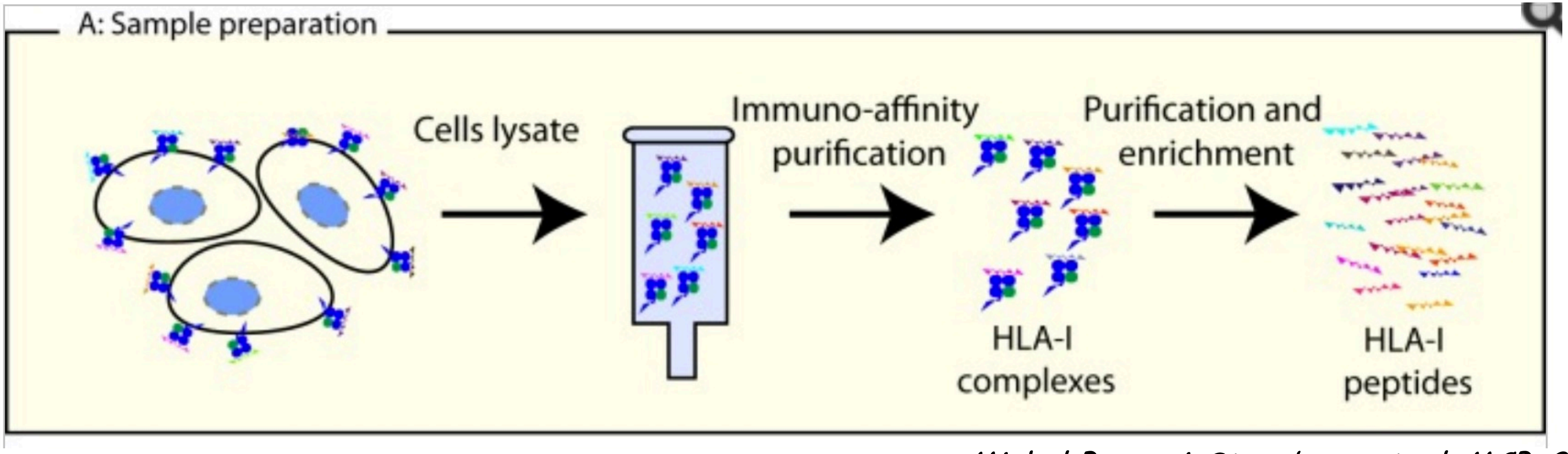

High Temperature
Low Temperature
Entropy
# Iterations
4.5
4
3.5
3
2.5
2
0
2500
5000
7500
10000
bits
1
0
1 2 3 4 5 6 7 8 9
Bits
0.2
0.0
-0.2
created by Seq2Logo

Gibbs sampler. Prediction accuracy



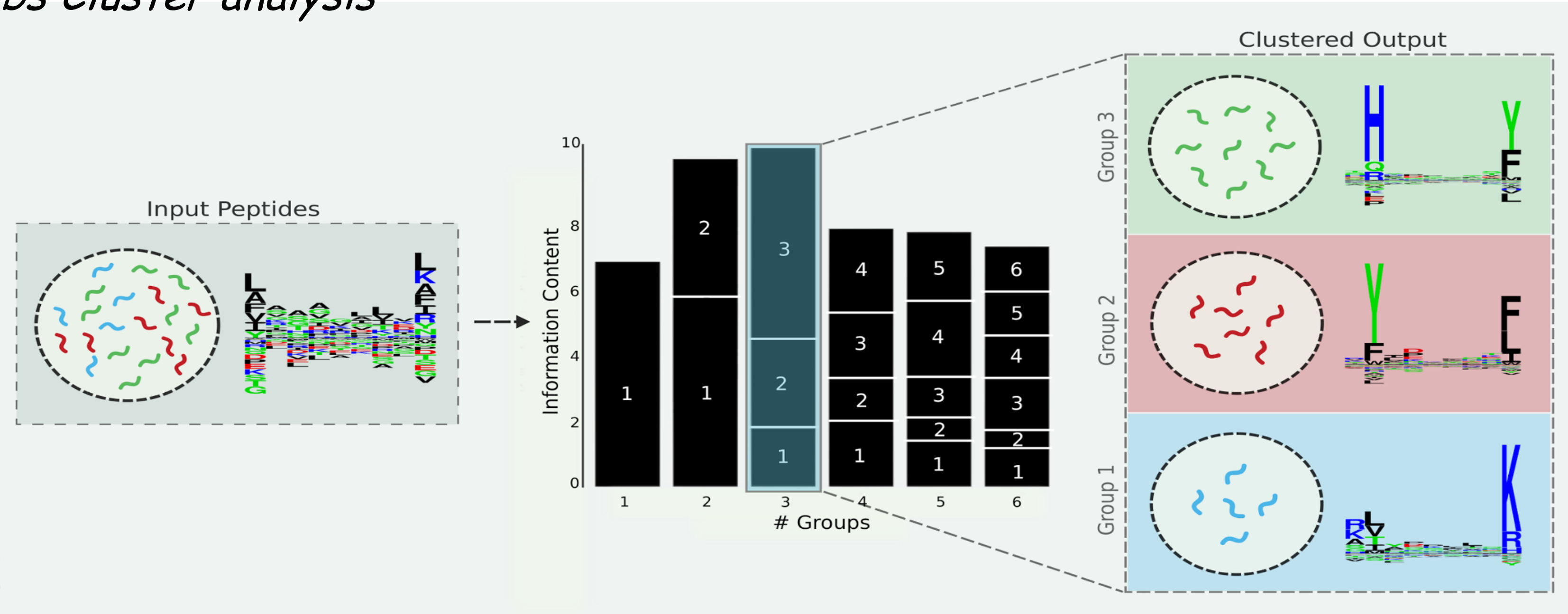
	Gibbs	Tepitope	ClustalW
🟦 AUC	0.744	0.702	0.667

Interpreting and benefitting from MS eluted ligand data sets



Michal Bassani-Sternberg et. al, MCP, 2015

Gibbs Cluster analysis



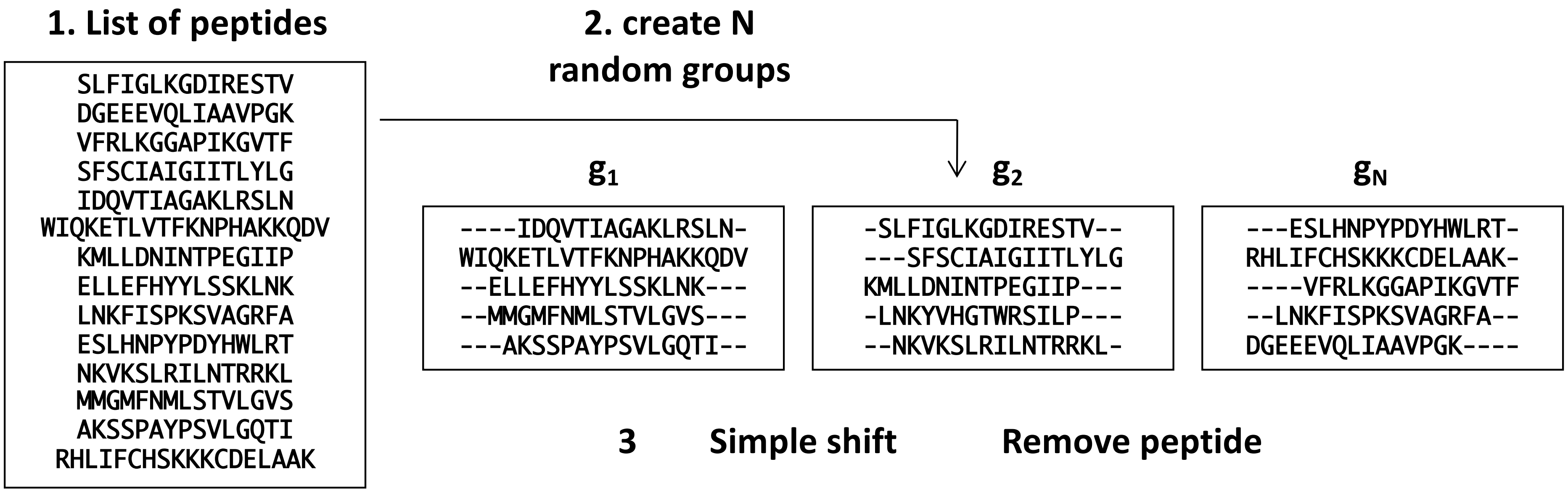
GibbsCluster, Andreatta, Alvarez, Nielsen, NAR, 2017

The algorithm

1. List of peptides

```
    SLFIGLKGDIRESTV
    DGEEEVQLIAAVPGK
    VFRLKGGAPIKGVTF
    SFSCIAIGIITLYLG
    IDQVTIAGAKLRSLN
  WIQKETLVTFKNPHAKKQDV
    KMLLDNINTPEGIIP
    ELLEFHYYLSSKLNK
    LNKFISPKSVAGRFA
    ESLHNPYPDHWLRT
    NKVKSLRILNTRRKL
    MMGMFNMLSTVLGVS
    AKSSPAYPSVLGQTI
   RHLIFCHSKKKCDELAAK
```

The algorithm



1. List of peptides

```
SLFIGLKGDIRESTV
DGEEEVQLIAAVPGK
VFRLKGGAPIKGVTF
SFSCIAIGIITLYLG
IDQVTIAGAKLRSLN
WIQKETLVTFKNPHAKKQDV
KMLLDNINTPEGIIP
ELLEFHYYLSSKLNK
LNKFISPKSVAGRFA
ESLHNPYPDYHWLRT
NKVKSLRILNTRRKL
MMGMFNMLSTVLGVS
AKSSPAYPSVLGQTI
RHLIFCHSKKKCDELAAK
```

2. create N random groups

g_1

```
----IDQVTIAGAKLRSLN-
WIQKETLVTFKNPHAKKQDV
--ELLEFHYYLSSKLNK---
--MMGMFNMLSTVLGVS---
---AKSSPAYPSVLGQTI--
```

g_2

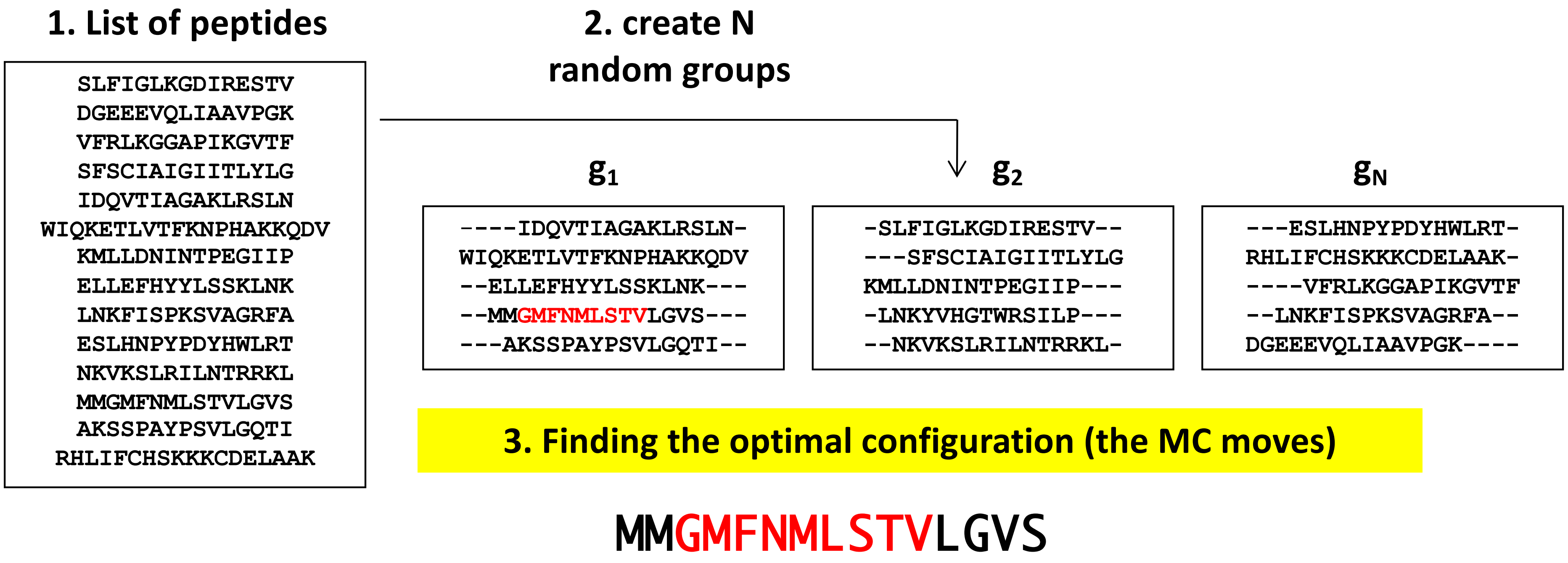
```
-SLFIGLKGDIRESTV--
---SFSCIAIGIITLYLG
KMLLDNINTPEGIIP---
-LNKYVHGTWRSILP---
--NKVKSLRILNTRRKL-
```

g_N

```
---ESLHNPYPDYHWLRT-
RHLIFCHSKKKCDELAAK-
----VFRLKGGAPIKGVTF
--LNKFISPKSVAGRFA--
DGEEEVQLIAAVPGK----
```

3 Simple shift Remove peptide

The algorithm



1. List of peptides

```
SLFIGLKGDIRESTV
DGEEEVQLIAAVPGK
VFRLKGGAPIKGVTF
SFSCIAIGIITLYLG
IDQVTIAGAKLRSLN
WIQKETLVTFKNPHAKKQDV
KMLLDNINTPEGIIP
ELLEFHYYLSSKLNK
LNKFISPKSVAGRFA
ESLHNPYPDYHWLRT
NKVKSLRILNTRRKL
MMGMFNMLSTVLGVS
AKSSPAYPSVLGQTI
RHLIFCHSKKKCDELAAK
```

2. create N random groups

g_1

```
----IDQVTIAGAKLRSLN-
WIQKETLVTFKNPHAKKQDV
--ELLEFHYYLSSKLNK---
--MMGMFNMLSTVLGVS---
---AKSSPAYPSVLGQTI--
```

g_2

```
-SLFIGLKGDIRESTV--
---SFSCIAIGIITLYLG
KMLLDNINTPEGIIP---
-LNKYVHGTWRSILP---
--NKVKSLRILNTRRKL-
```

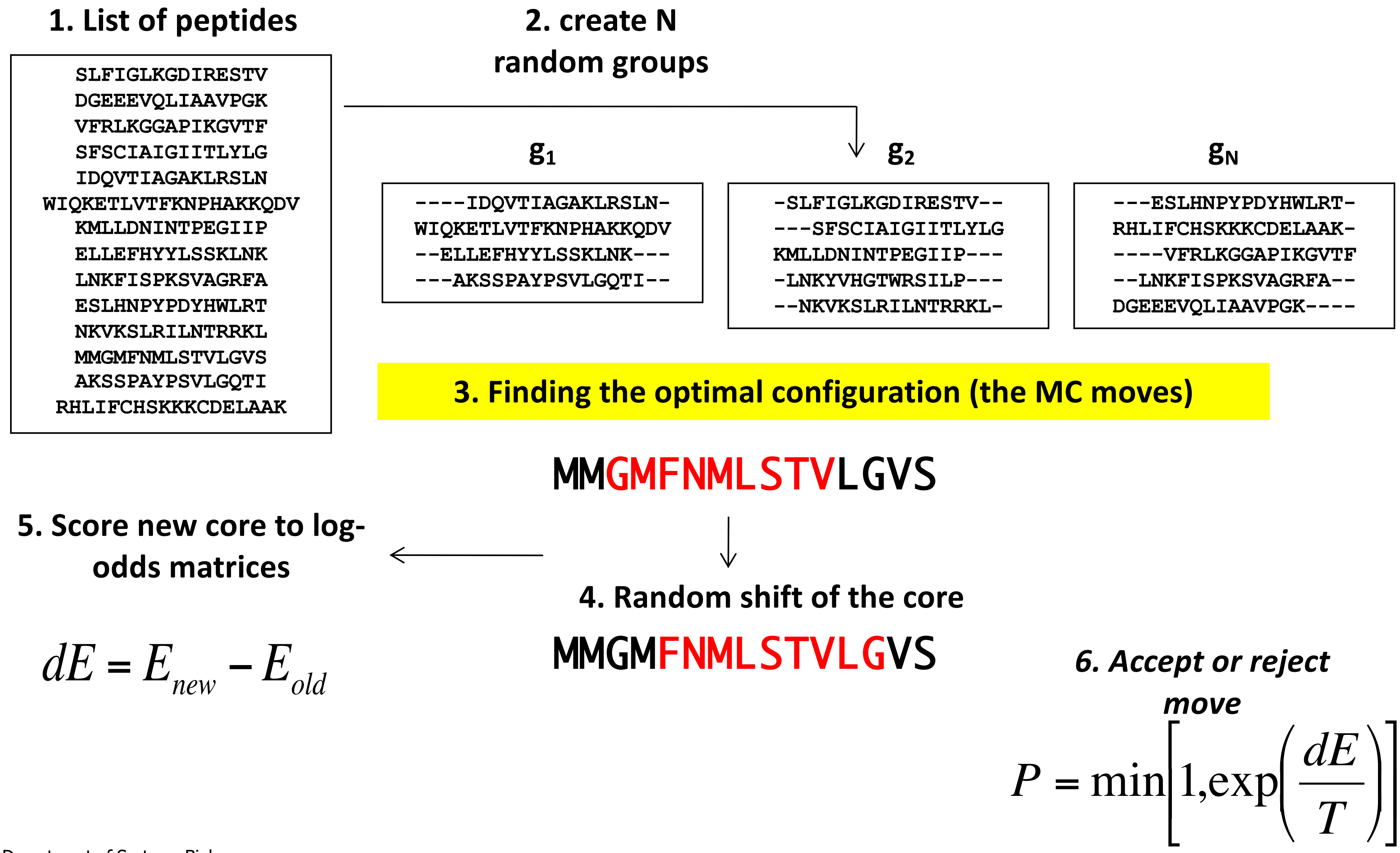
g_N

```
---ESLHNPYPDYHWLRT-
RHLIFCHSKKKCDELAAK-
----VFRLKGGAPIKGVTF
--LNKFISPKSVAGRFA--
DGEEEVQLIAAVPGK----
```

3. Finding the optimal configuration (the MC moves)

MMGMFNMLSTVLGVS

The algorithm



1. List of peptides

```
SLFIGLKGDIRESTV
DGEEEVQLIAAVPGK
VFRLKGGAPIKGVTF
SFSCIAIGIITLYLG
IDQVTIAGAKLRSLN
WIQKETLVTFKNPHAKKQDV
KMLLDNINTPEGIIP
ELLEFHYYLSSKLNK
LNKFISPKSVAGRFA
ESLHNPYPDYHWLRT
NKVKSLRILNTRRKL
MMGMFNMLSTVLGVS
AKSSPAYPSVLGQTI
RHLIFCHSKKKCDELAAK
```

2. create N random groups

g_1

```
----IDQVTIAGAKLRSLN-
WIQKETLVTFKNPHAKKQDV
--ELLEFHYYLSSKLNK---
---AKSSPAYPSVLGQTI--
```

g_2

```
-SLFIGLKGDIRESTV--
---SFSCIAIGIITLYLG
KMLLDNINTPEGIIP---
-LNKYVHGTWRSILP---
--NKVKSLRILNTRRKL-
```

g_N

```
---ESLHNPYPDYHWLRT-
RHLIFCHSKKKCDELAAK-
----VFRLKGGAPIKGVTF
--LNKFISPKSVAGRFA--
DGEEEVQLIAAVPGK----
```

3. Finding the optimal configuration (the MC moves)

MMGMFNMLSTVLGVS

4. Random shift of the core

MMGMFNMLSTVLGVS

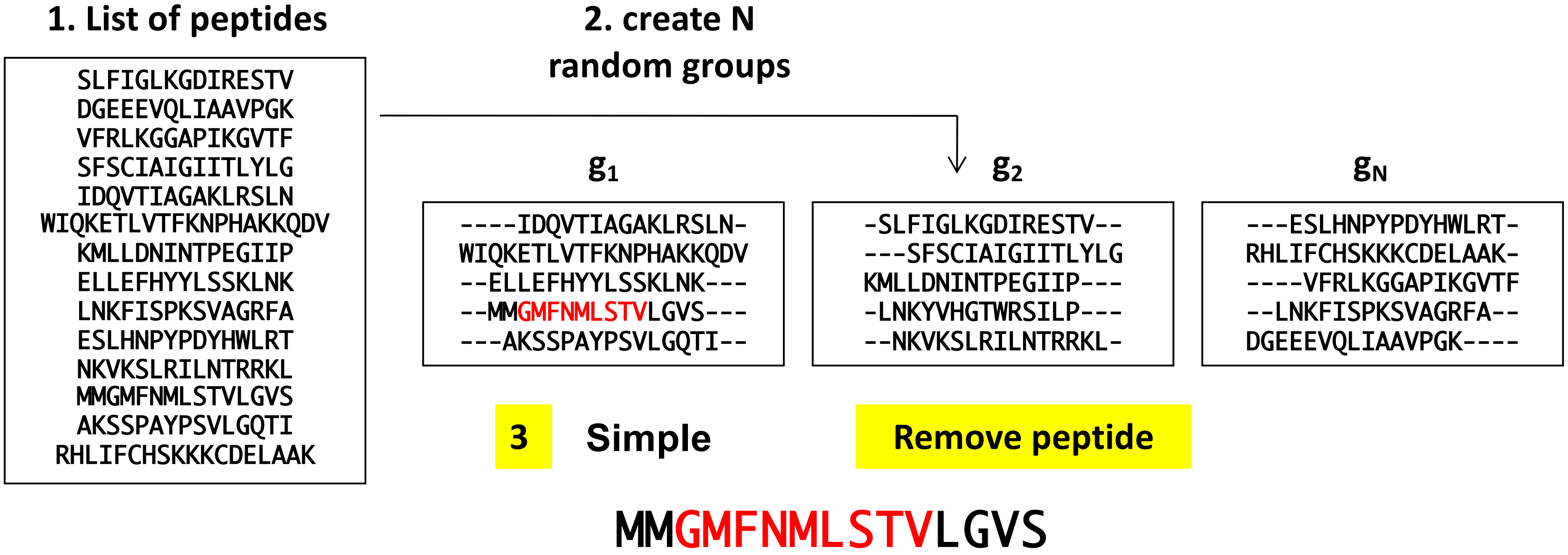
5. Score new core to log-odds matrices

$$dE = E_{new} - E_{old}$$

6. Accept or reject move

$$P = \min\left[1, \exp\left(\frac{dE}{T}\right)\right]$$

The algorithm



1. List of peptides

```
SLFIGLKGDIRESTV
DGEEEVQLIAAVPGK
VFRLKGGAPIKGVTF
SFSCIAIGIITLYLG
IDQVTIAGAKLRSLN
WIQKETLVTFKNPHAKKQDV
KMLLDNINTPEGIIP
ELLEFHYYLSSKLNK
LNKFISPKSVAGRFA
ESLHNPYPDYHWLRT
NKVKSLRILNTRRKL
MMGMFNMLSTVLGVS
AKSSPAYPSVLGQTI
RHLIFCHSKKKCDELAAK
```

2. create N random groups

g_1

```
----IDQVTIAGAKLRSLN-
WIQKETLVTFKNPHAKKQDV
--ELLEFHYYLSSKLNK---
--MMGMFNMLSTVLGVS---
---AKSSPAYPSVLGQTI--
```

g_2

```
-SLFIGLKGDIRESTV--
---SFSCIAIGIITLYLG
KMLLDNINTPEGIIP---
-LNKYVHGTWRSILP---
--NKVKSLRILNTRRKL-
```

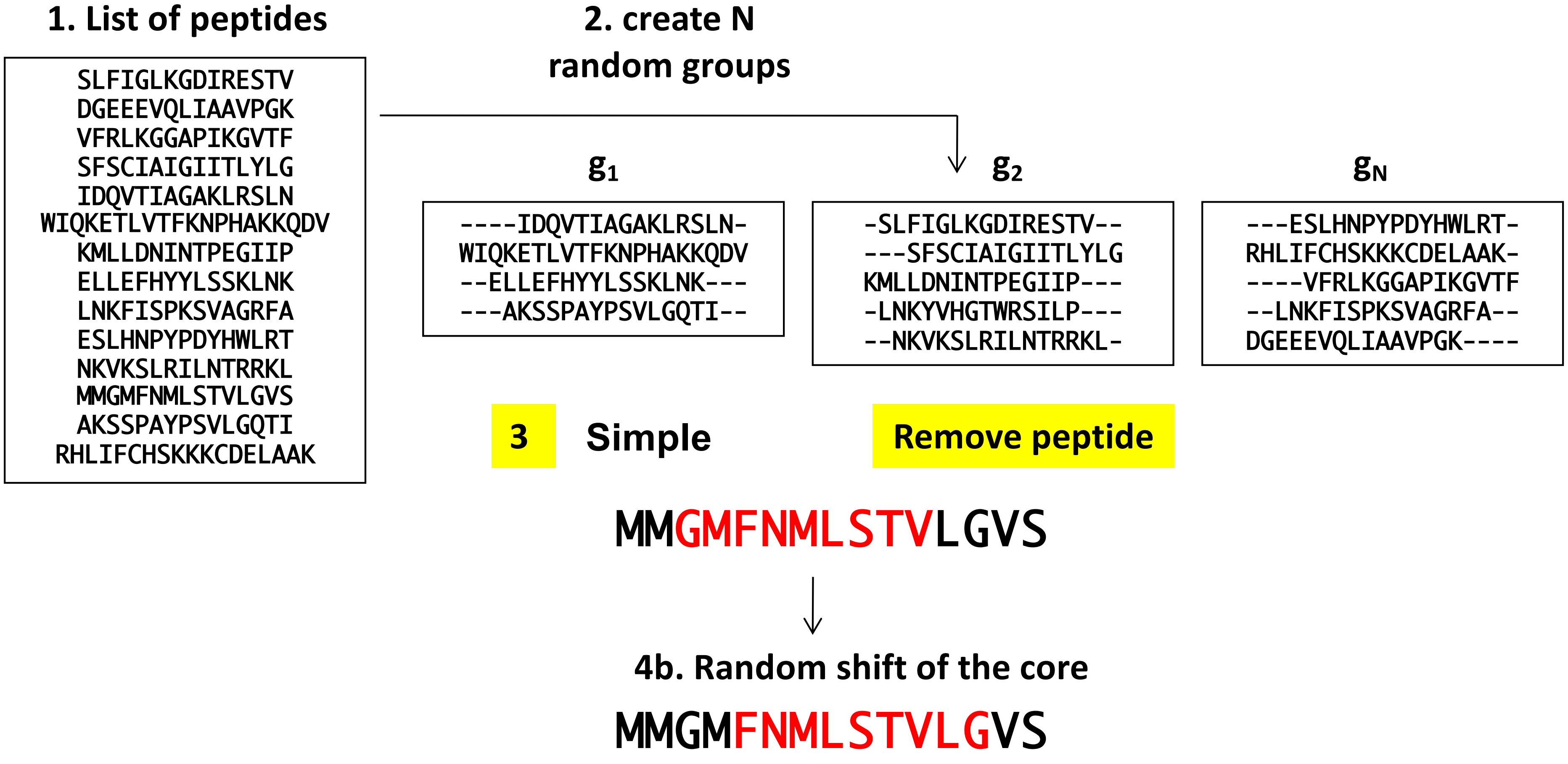
g_N

```
---ESLHNPYPDYHWLRT-
RHLIFCHSKKKCDELAAK-
----VFRLKGGAPIKGVTF
--LNKFISPKSVAGRFA--
DGEEEVQLIAAVPGK----
```

3 **Simple** **Remove peptide**

MMGMFNMLSTVLGVS

The algorithm



1. List of peptides

```
  SLFIGLKGDIRESTV
  DGEEEVQLIAAVPGK
  VFRLKGGAPIKGVTF
  SFSCIAIGIITLYLG
  IDQVTIAGAKLRSLN
WIQKETLVTFKNPHAKKQDV
  KMLLDNINTPEGIIP
  ELLEFHYYLSSKLNK
  LNKFISPKSVAGRFA
  ESLHNPYPDYHWLRT
  NKVKSLRILNTRRKL
  MMGMFNMLSTVLGVS
  AKSSPAYPSVLGQTI
 RHLIFCHSKKKCDELAAK
```

2. create N random groups

g_1

```
----IDQVTIAGAKLRSLN-
WIQKETLVTFKNPHAKKQDV
--ELLEFHYYLSSKLNK---
---AKSSPAYPSVLGQTI--
```

g_2

```
-SLFIGLKGDIRESTV--
---SFSCIAIGIITLYLG
KMLLDNINTPEGIIP---
-LNKYVHGTWRSILP---
--NKVKSLRILNTRRKL-
```

g_N

```
---ESLHNPYPDYHWLRT-
RHLIFCHSKKKCDELAAK-
----VFRLKGGAPIKGVTF
--LNKFISPKSVAGRFA--
DGEEEVQLIAAVPGK----
```

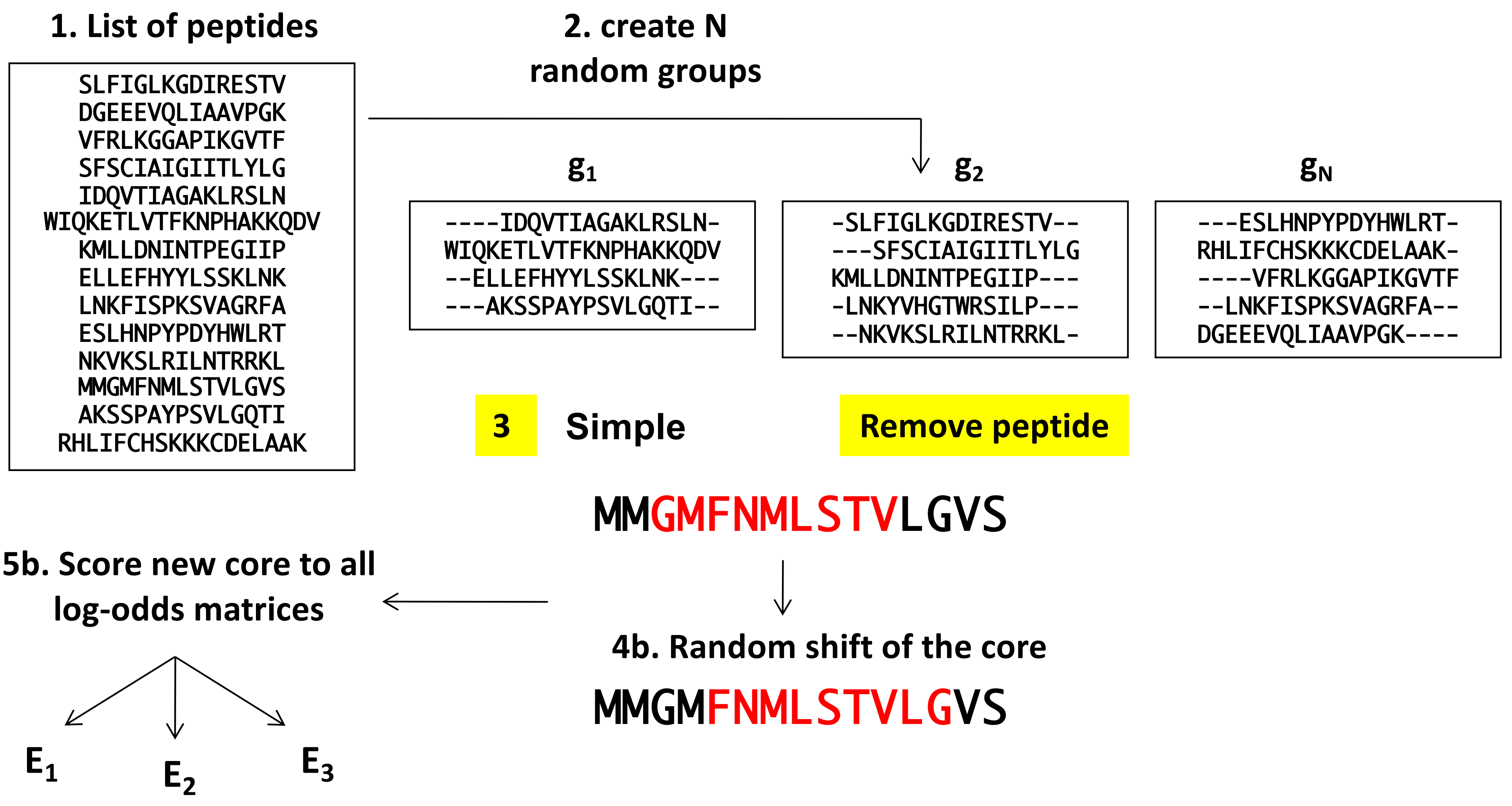
3 **Simple** **Remove peptide**

MM**GMFNMLSTV**LGVS

4b. Random shift of the core

MMGM**FNMLSTVLG**VS

The algorithm



1. List of peptides

```
SLFIGLKGDIRESTV
DGEEEVQLIAAVPGK
VFRLKGGAPIKGVTF
SFSCIAIGIITLYLG
IDQVTIAGAKLRSLN
WIQKETLVTFKNPHAKKQDV
KMLLDNINTPEGIIP
ELLEFHYYLSSKLNK
LNKFISPKSVAGRFA
ESLHNPYPDYHWLRT
NKVKSLRILNTRRKL
MMGMFNMLSTVLGVS
AKSSPAYPSVLGQTI
RHLIFCHSKKKCDELAAK
```

2. create N random groups

g_1

```
----IDQVTIAGAKLRSLN-
WIQKETLVTFKNPHAKKQDV
--ELLEFHYYLSSKLNK---
---AKSSPAYPSVLGQTI--
```

g_2

```
-SLFIGLKGDIRESTV--
---SFSCIAIGIITLYLG
KMLLDNINTPEGIIP---
-LNKYVHGTWRSILP---
--NKVKSLRILNTRRKL-
```

g_N

```
---ESLHNPYPDYHWLRT-
RHLIFCHSKKKCDELAAK-
----VFRLKGGAPIKGVTF
--LNKFISPKSVAGRFA--
DGEEEVQLIAAVPGK----
```

3 **Simple** **Remove peptide**

MMGMFNMLSTVLGVS

4b. Random shift of the core

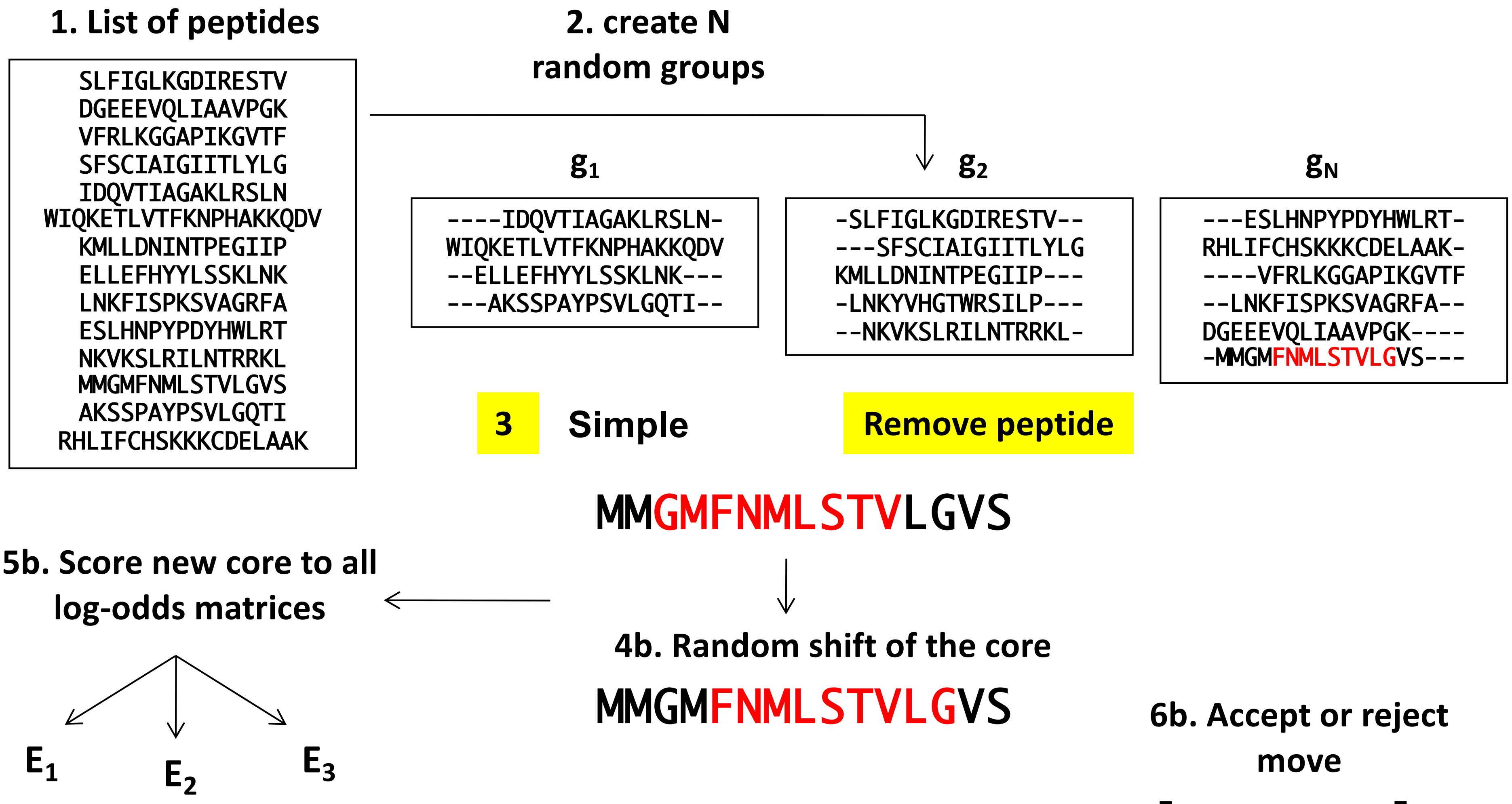
MMGMFNMLSTVLGVS

5b. Score new core to all log-odds matrices

E_1 E_2 E_3

$$dE = E_{before} - \max(E_1, E_2, E_3)$$

The algorithm



1. List of peptides

```
SLFIGLKGDIRESTV
DGEEEVQLIAAVPGK
VFRLKGGAPIKGVTF
SFSCIAIGIITLYLG
IDQVTIAGAKLRSLN
WIQKETLVTFKNPHAKKQDV
KMLLDNINTPEGIIP
ELLEFHYYLSSKLNK
LNKFISPKSVAGRFA
ESLHNPYPDYHWLRT
NKVKSLRILNTRRKL
MMGMFNMLSTVLGVS
AKSSPAYPSVLGQTI
RHLIFCHSKKKCDELAAK
```

2. create N random groups

g_1

```
----IDQVTIAGAKLRSLN-
WIQKETLVTFKNPHAKKQDV
--ELLEFHYYLSSKLNK---
---AKSSPAYPSVLGQTI--
```

g_2

```
-SLFIGLKGDIRESTV--
---SFSCIAIGIITLYLG
KMLLDNINTPEGIIP---
-LNKYVHGTWRSILP---
--NKVKSLRILNTRRKL-
```

g_N

```
---ESLHNPYPDYHWLRT-
RHLIFCHSKKKCDELAAK-
----VFRLKGGAPIKGVTF
--LNKFISPKSVAGRFA--
DGEEEVQLIAAVPGK----
-MMGMFNMLSTVLGVS---
```

3 **Simple** **Remove peptide**

MMGMFNMLSTVLGVS

4b. Random shift of the core

MMGMFNMLSTVLGVS

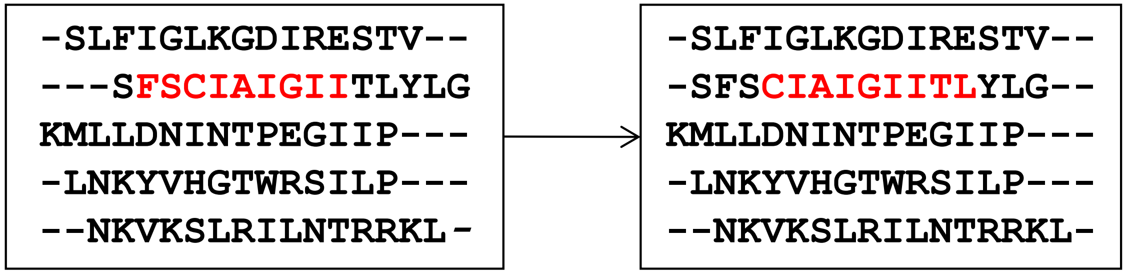
5b. Score new core to all log-odds matrices

E_1 E_2 E_3

6b. Accept or reject move

$$dE = E_{before} - \max(E_1, E_2, E_3) \qquad P = \min\left[1, \exp\left(\frac{dE}{T}\right)\right]$$

The scoring function



```
-SLFIGLKGDIRESTV--
---SFSCIAIGIITLYLG
KMLLDNINTPEGIIP---
-LNKYVHGTWRSILP---
--NKVKSLRILNTRRKL-
```

```
-SLFIGLKGDIRESTV--
-SFSCIAIGIITLYLG--
KMLLDNINTPEGIIP---
-LNKYVHGTWRSILP---
--NKVKSLRILNTRRKL-
```

$$LO_{A,j} = \frac{n}{n+\sigma}\log\frac{p_{A,j}'}{q_A}$$

Avoid small specialized clusters (σ = 10)

$$E = \sum_j LO_{A,j} \qquad E_i^* = E_i - \lambda \max_{\substack{1\le n\le g\\ n\neq i}}\left(E_n,0\right)$$

Maximize intra cluster similarity whilst minimize inter cluster similarity

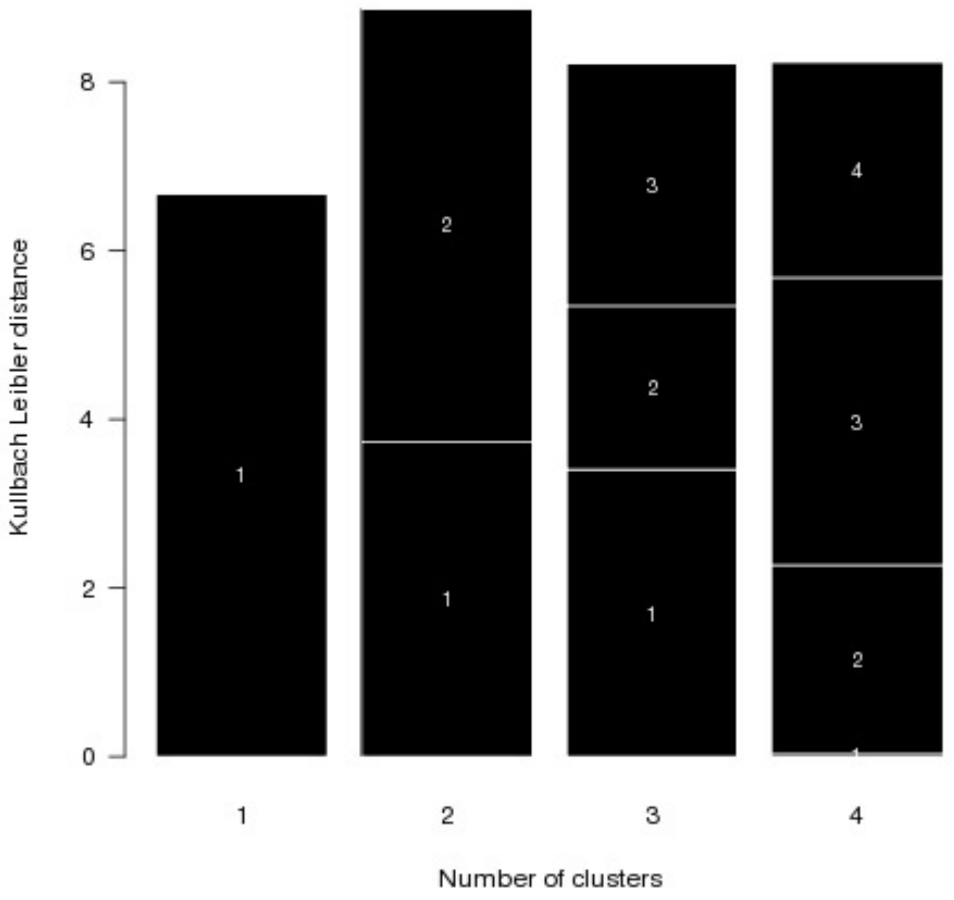
$$dE = E^*_{before} - E^*_{after}$$

$$P = \min\left[1,\exp\left(\frac{dE}{T}\right)\right]$$

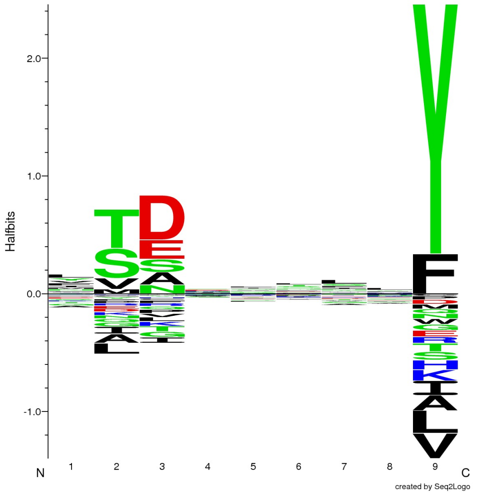
P = Probability of accepting the move

A mixture of 500 9mer peptides. How many motifs?

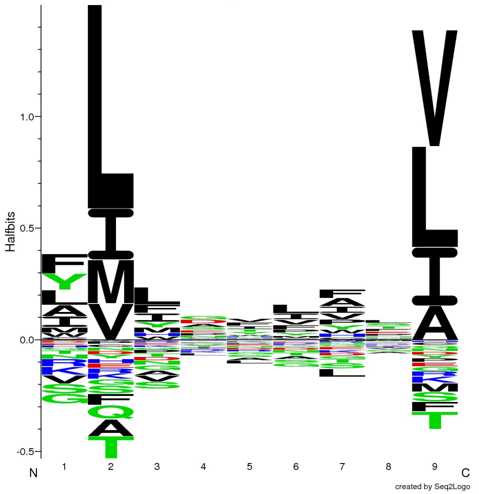
A mixture of 500 9mer peptides. How many motifs?



Group 1 of 2



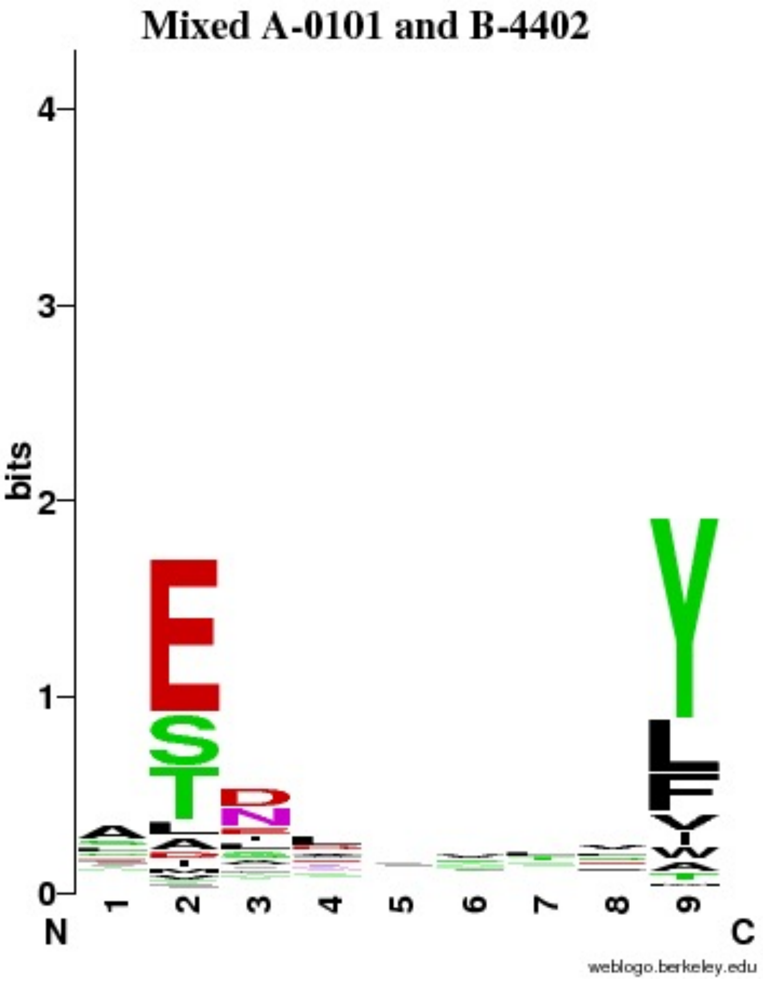
Group 2 of 2



Two MHC class I alleles: HLA-A*0101 and HLA-B*4402

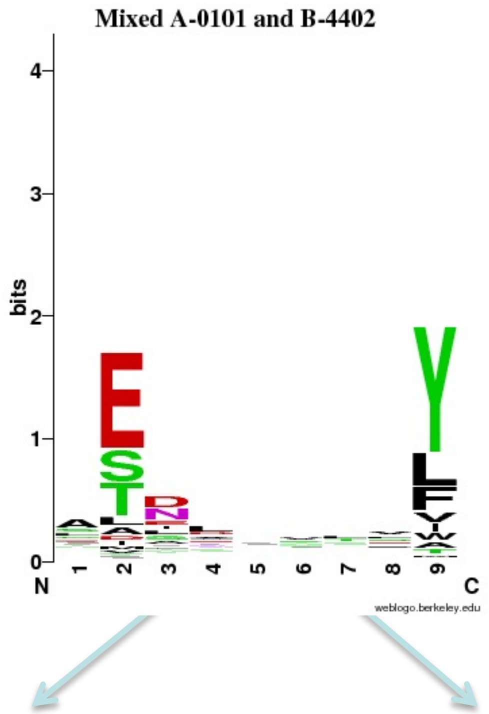
Mixture of 100 binders
for the two alleles

ATDKAAAAY	A*0101
EVDQTKIQY	A*0101
AETGSQGVY	B*4402
ITDITKYLY	A*0101
AEMKTDAAT	B*4402
FEIKSAKKF	B*4402
LSEMLNKEY	A*0101
GELDRWEKI	B*4402
LTDSSTLLV	A*0101
FTIDFKLKY	A*0101
TTTIKPVSY	A*0101
EEKAFSPEV	B*4402
AENLWVPVY	B*4402



Two MHC class I alleles: HLA-A*0101 and HLA-B*4402

Mixed

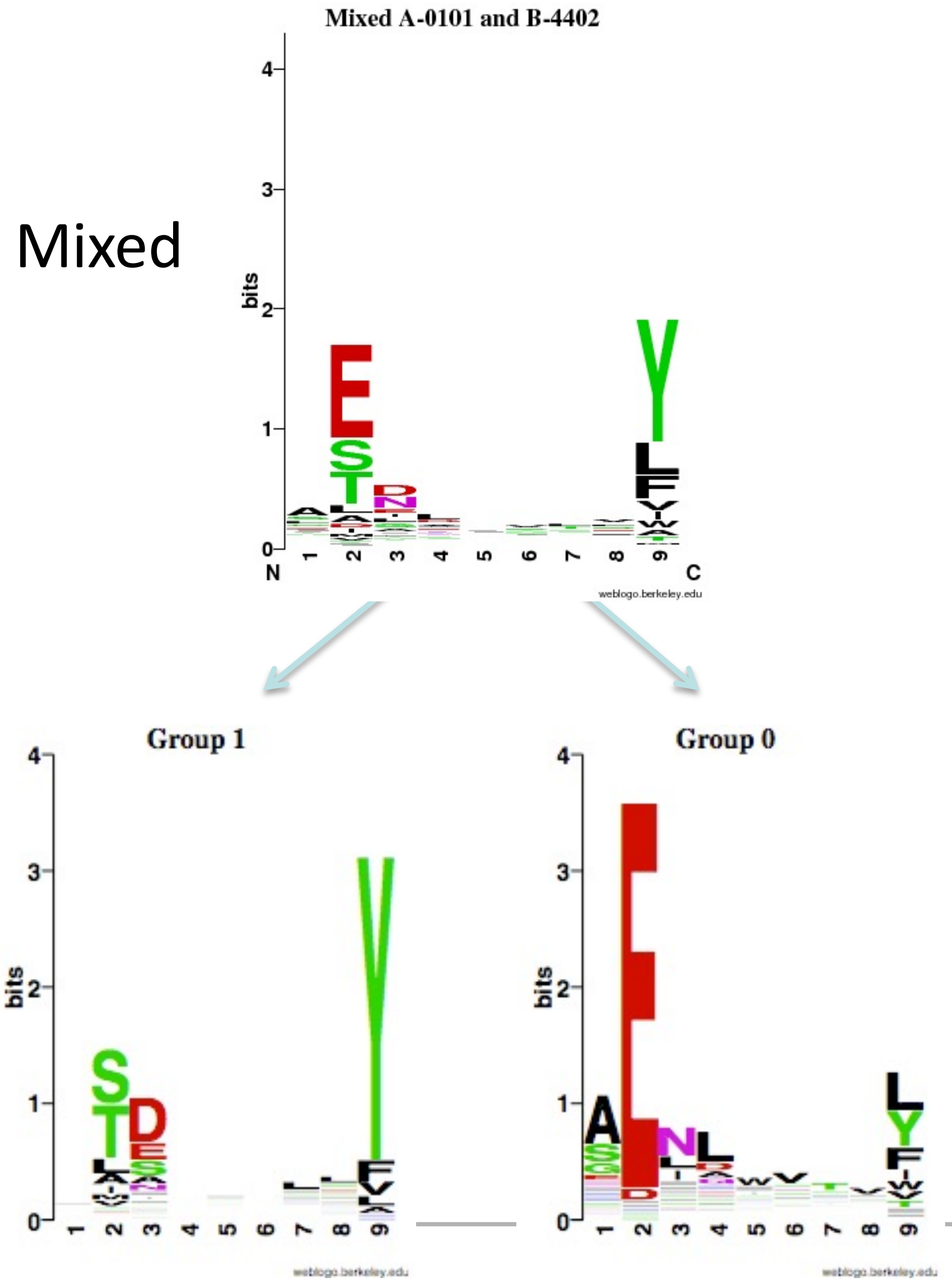


	A0101	B4402
G 1		
G 2		

Two MHC class I alleles: HLA-A*0101 and HLA-B*4402

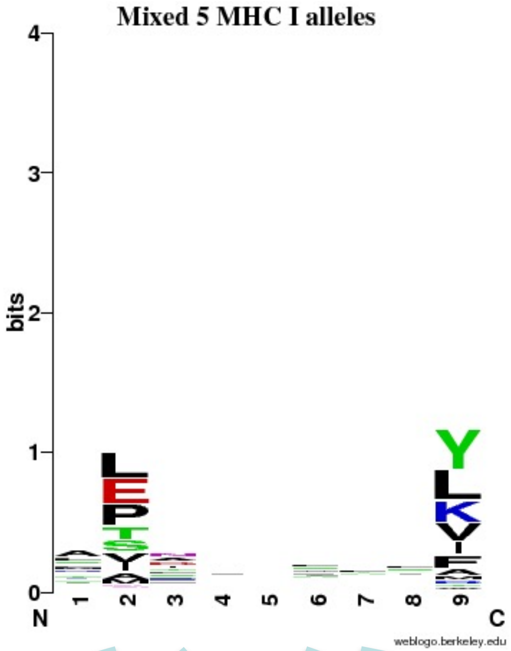
Mixed

Resolved



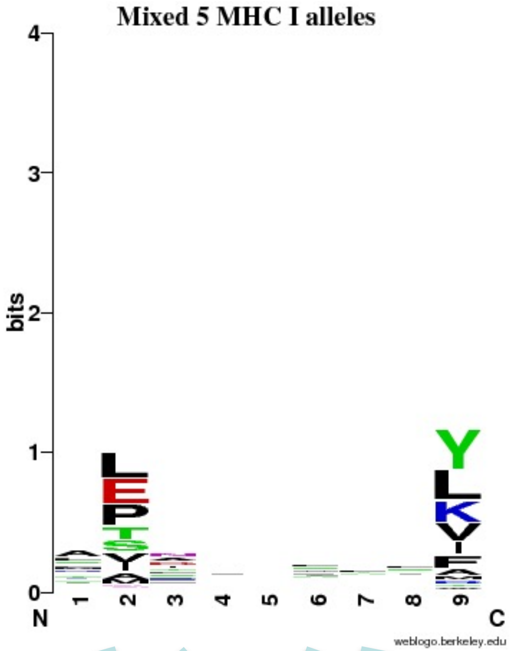
	A0101	B4402
G 1	97	3
G 2	3	97

Five MHC class I alleles

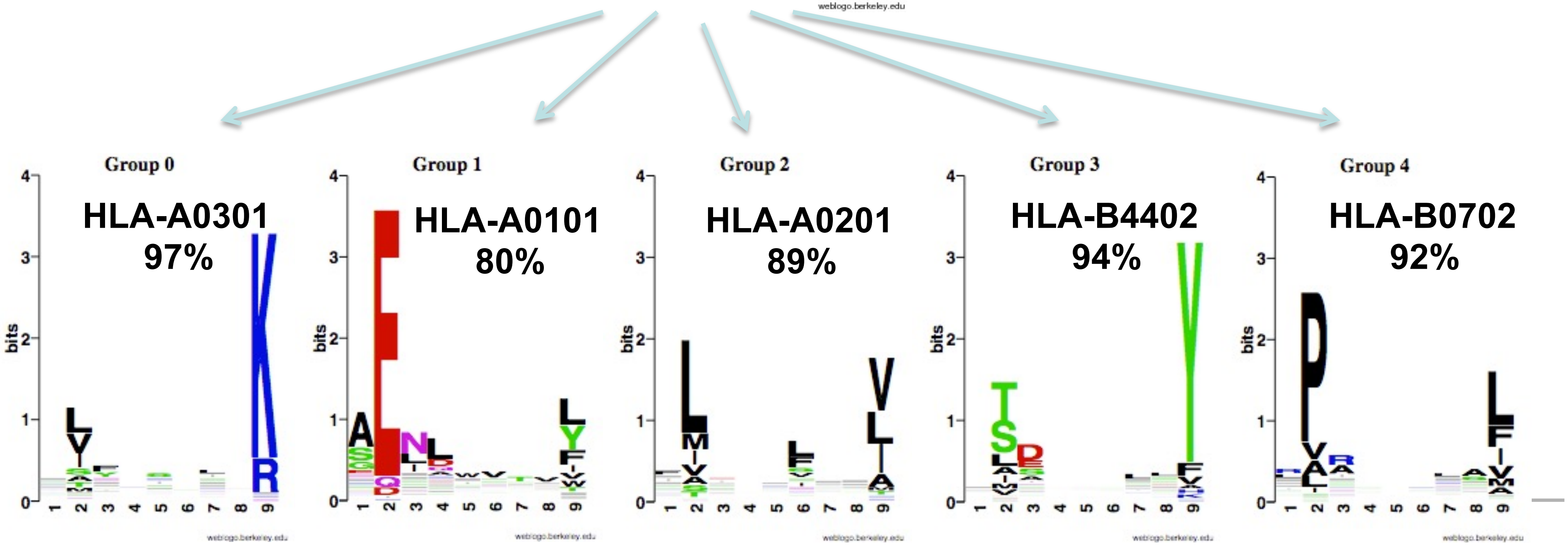


	A010	A0201	A0301	B0702	B4402
G 0					
G 1					
G 2					
G 3					
G 4					

Five MHC class I alleles



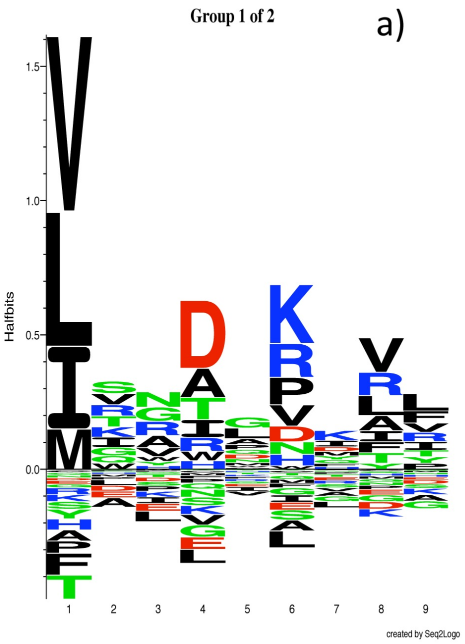
	A0101	A0201	A0301	B0702	B4402
G 0	0	1	76	1	0
G 1	2	4	0	0	95
G 2	5	87	5	1	0
G 3	93	2	19	0	2
G 4	0	6	0	98	3



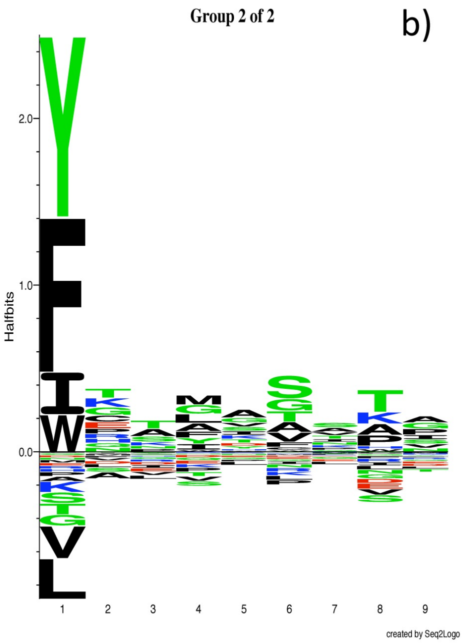
Adding in alignment (MHC class II)

SLFIGLKGDIRESTV
DGEEEVQLIAAVPGK
VFRLKGGAPIKGVTF
SFSCIAIGIITLYLG
IDQVTIAGAKLRSLN
WIQKETLVTFKNPHAKKQDV
KMLLDNINTPEGIIP
ELLEFHYYLSSKLNK
LNKFISPKSVAGRFA
ESLHNPYPDHWLRT
NKVKSLRILNTRRKL
MMGMFNMLSTVLGVS
AKSSPAYPSVLGQTI
RHLIFCHSKKKCDELAAK

HLA-DRB1*03:01



HLA-DRB1*04:01



c)

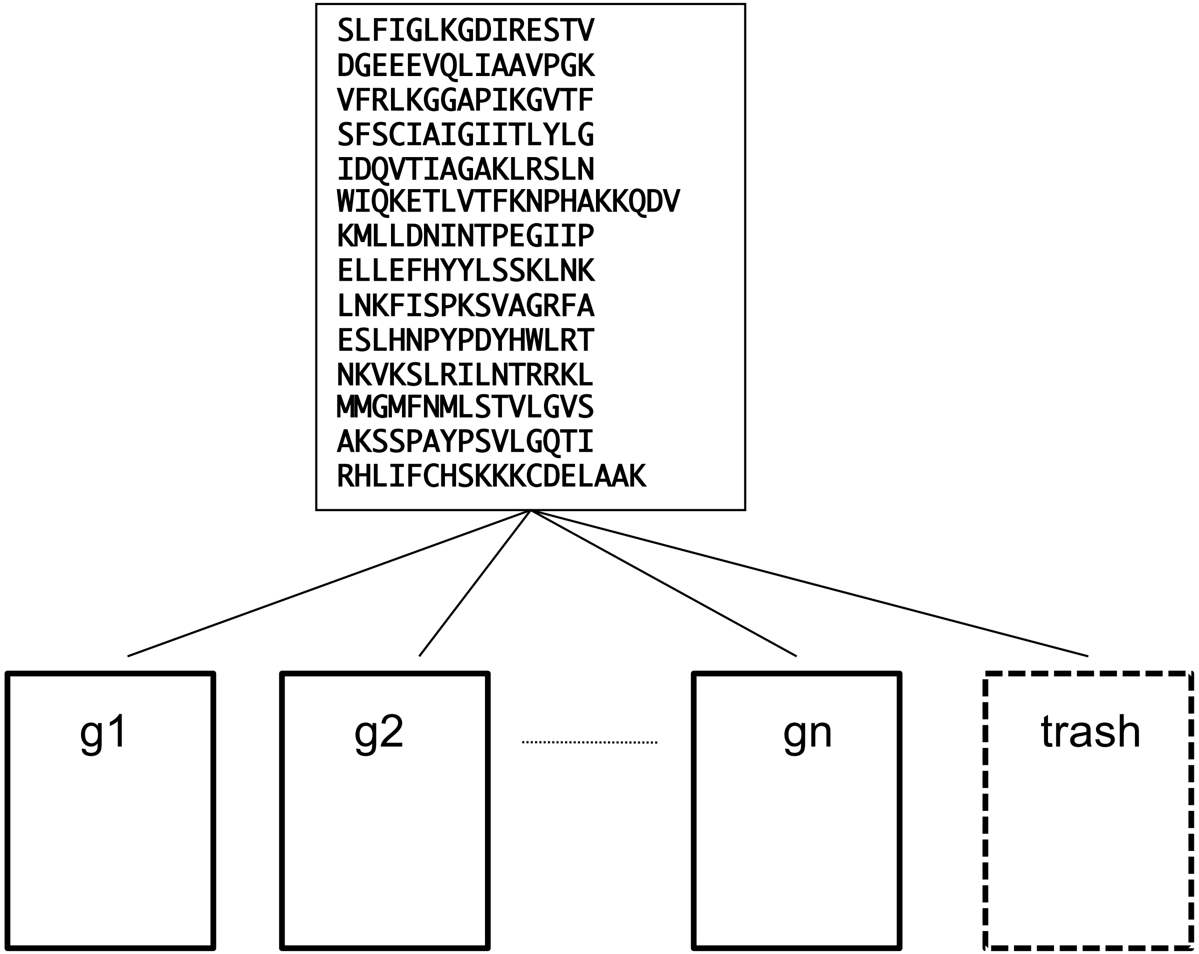
	DRB1*03:01	DRB1*04:01
C1	170	53
C2	31	149

Dealing with noisy data

- Experimental data often contain false positives
- Outliers do not match any recurrent motif
- Introduce a garbage bin to collect outliers

Dealing with noisy data

- Introduce a garbage bin to collect outliers



```
SLFIGLKGDIRESTV
DGEEEVQLIAAVPGK
VFRLKGGAPIKGVTF
SFSCIAIGIITLYLG
IDQVTIAGAKLRSLN
WIQKETLVTFKNPHAKKQDV
KMLLDNINTPEGIIP
ELLEFHYYLSSKLNK
LNKFISPKSVAGRFA
ESLHNPYPDYHWLRT
NKVKSLRILNTRRKL
MMGMFNMLSTVLGVS
AKSSPAYPSVLGQTI
RHLIFCHSKKKCDELAAK
```

Move to the trash cluster peptides that do not match any motif

Dealing with noisy data

200 binders to 3 MHC class I alleles

50 random sequences are added to the data set

3 alleles	HLA-A0101	HLA-B0702	HLA-B4001	Random	
g0	0	197	2	6	205
g1	199	1	0	2	202
g2	0	0	196	5	201
trash	1	2	2	37	42
	200	200	200	50	

Dealing with noisy data

200 binders to 3 MHC class I alleles

50 random sequences are added to the data set

3 alleles	HLA-A0101	HLA-B0702	HLA-B4001	Random	
g0	0	197	2	6	205
g1	199	1	0	2	202
g2	0	0	196	5	201
trash	1	2	2	37	42
	200	200	200	50	

DHHFTPQII

NAFGWENAY
SQTSYQYLI

ELPIVTPAL
ADKNLIKCS

Dealing with noisy data

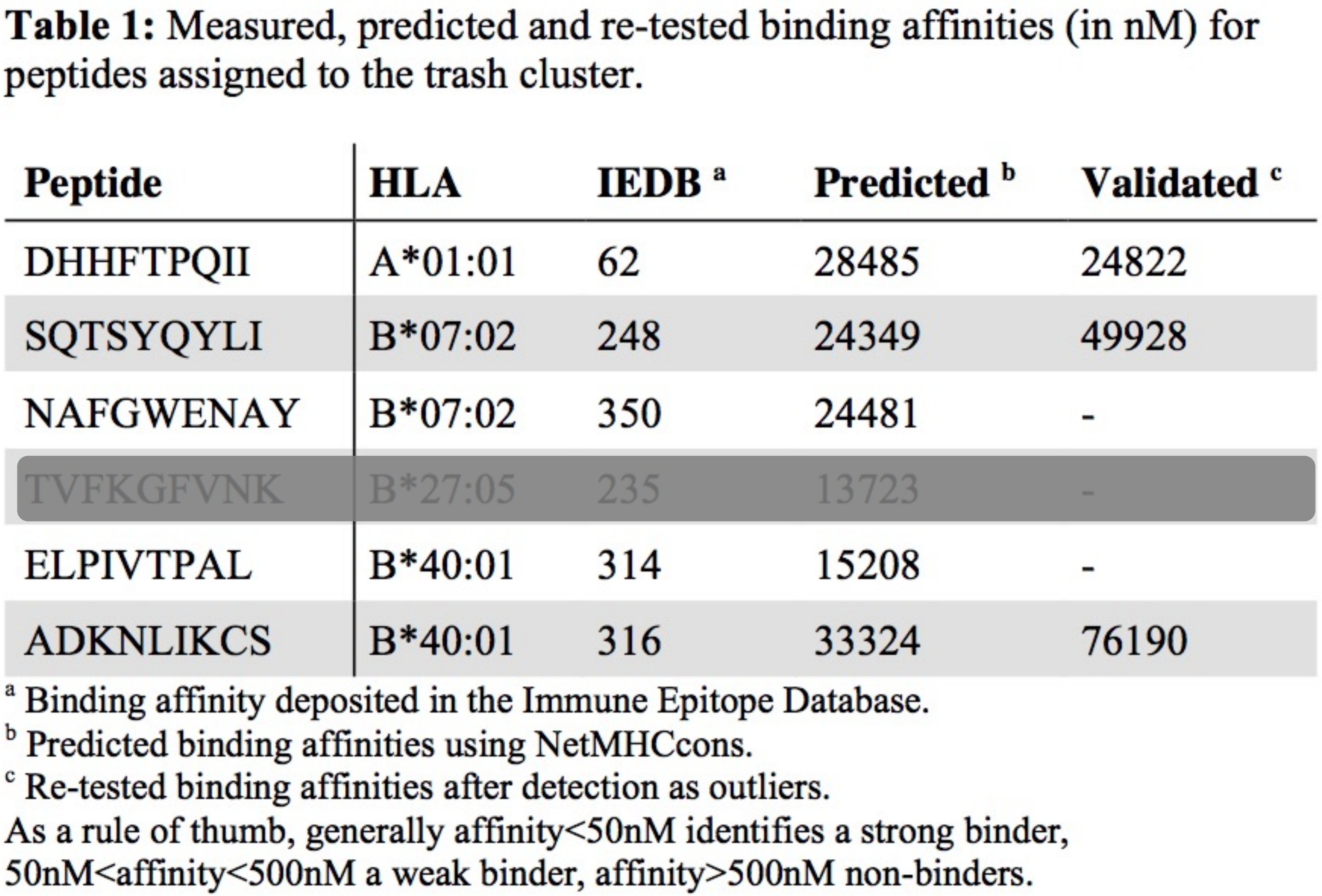


Table 1: Measured, predicted and re-tested binding affinities (in nM) for peptides assigned to the trash cluster.

Peptide	HLA	IEDB [a]	Predicted [b]	Validated [c]
DHHFTPQII	A*01:01	62	28485	24822
SQTSYQYLI	B*07:02	248	24349	49928
NAFGWENAY	B*07:02	350	24481	-
TVFKGFVNK	B*27:05	235	13723	-
ELPIVTPAL	B*40:01	314	15208	-
ADKNLIKCS	B*40:01	316	33324	76190

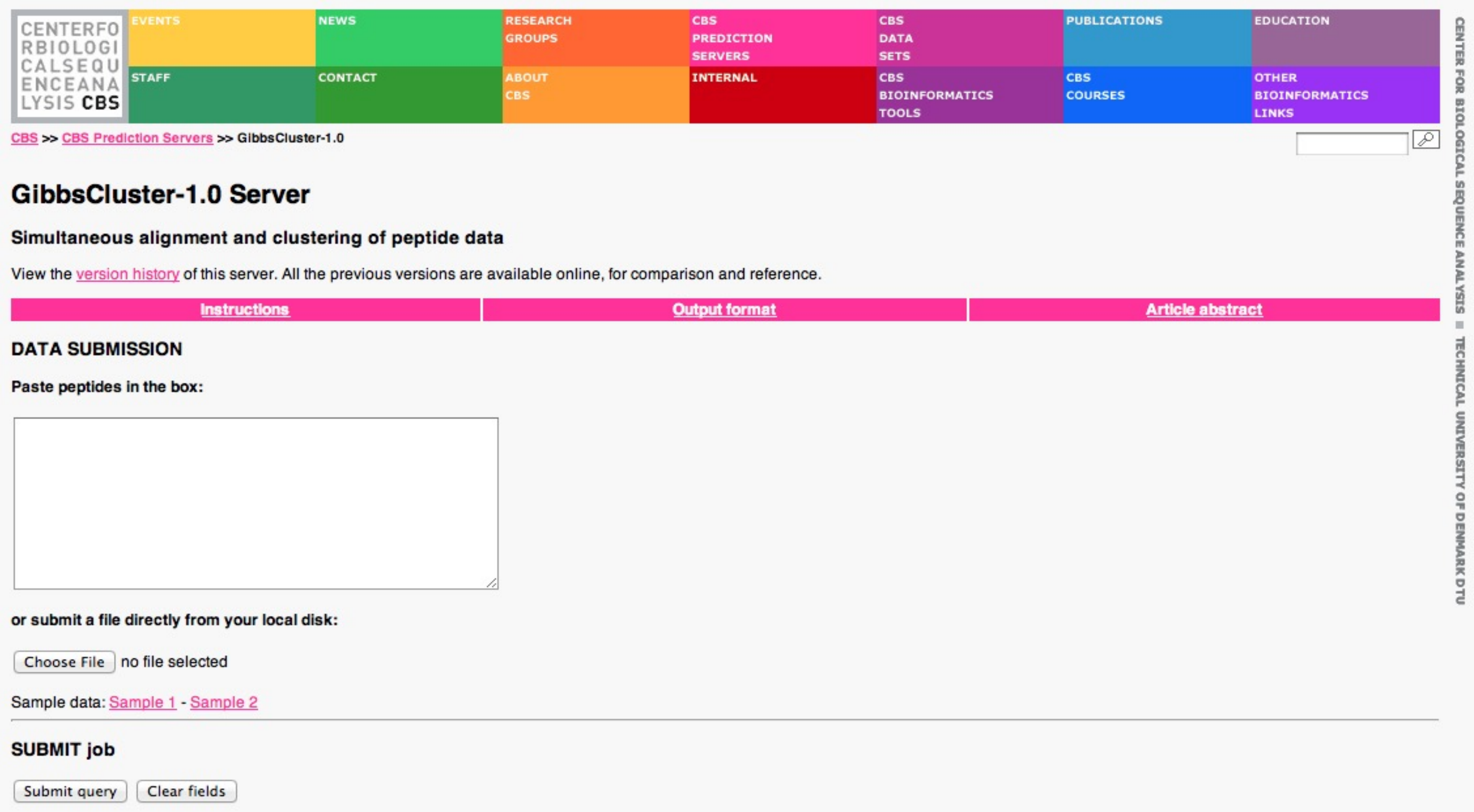
[a] Binding affinity deposited in the Immune Epitope Database.

[b] Predicted binding affinities using NetMHCcons.

[c] Re-tested binding affinities after detection as outliers.

As a rule of thumb, generally affinity<50nM identifies a strong binder, 50nM<affinity<500nM a weak binder, affinity>500nM non-binders.

http://www.cbs.dtu.dk/services/GibbsCluster



CBS >> CBS Prediction Servers >> GibbsCluster-1.0

GibbsCluster-1.0 Server

Simultaneous alignment and clustering of peptide data

View the version history of this server. All the previous versions are available online, for comparison and reference.

DATA SUBMISSION

Paste peptides in the box:

or submit a file directly from your local disk:

no file selected

Sample data: Sample 1 - Sample 2

SUBMIT job

CENTER FOR BIOLOGICAL SEQUENCE ANALYSIS