DegSampler: Collapsed Gibbs sampler for detecting E3 binding sites

Osamu Maruyama

Faculty of Design

Kyushu University

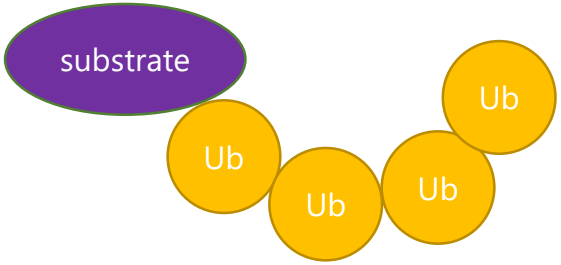
Collaborator

Dr. Fumiko Matsuzaki
Medical Institute of Bioregulation
Kyushu University

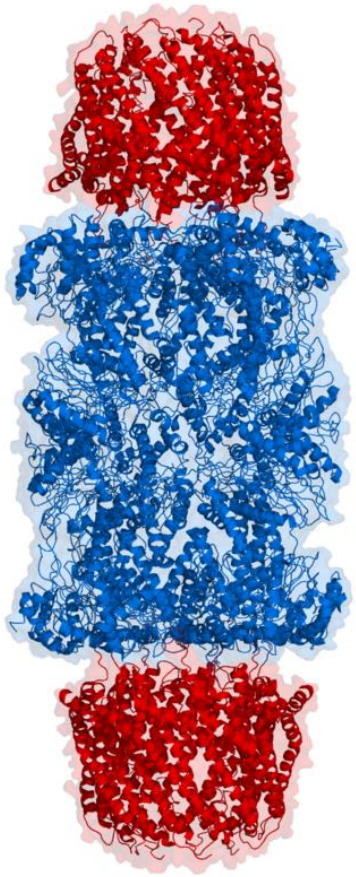
Contents

Motivation

Unneeded and damaged proteins are degraded and recycled

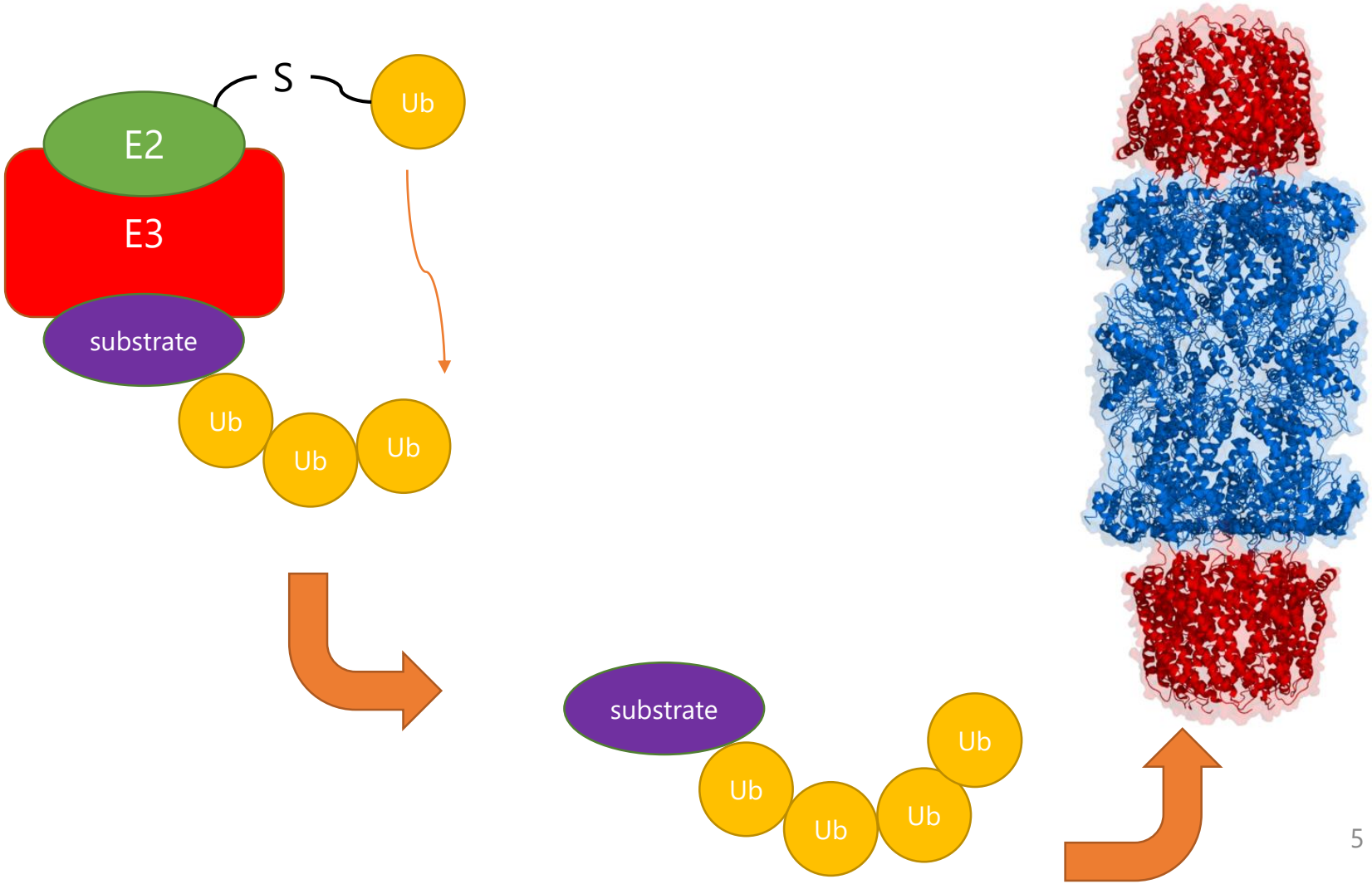


Proteasome (Wikipedia)



Ubiquitin-proseasome system

E3 ubiquitin ligase selectively recognizes a substrate protein

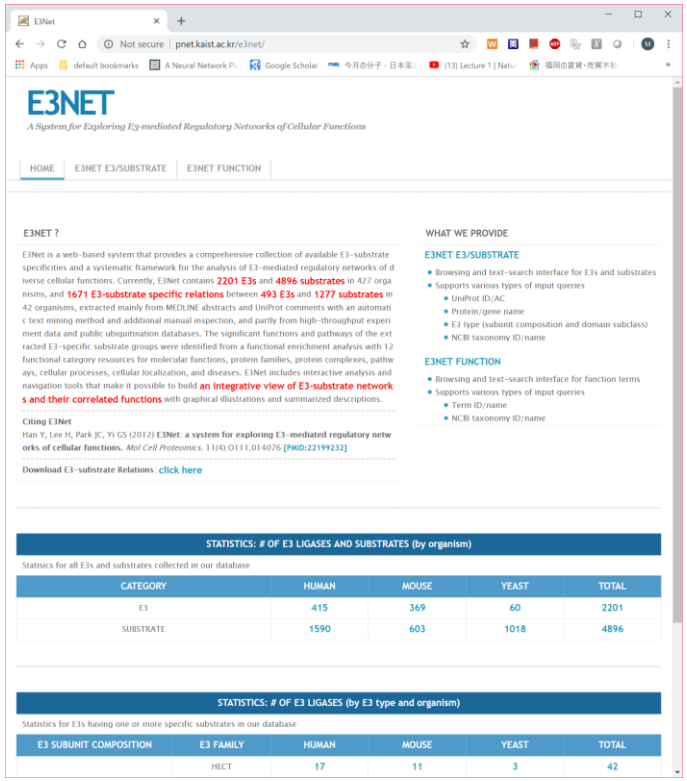


Contents

Available data sets

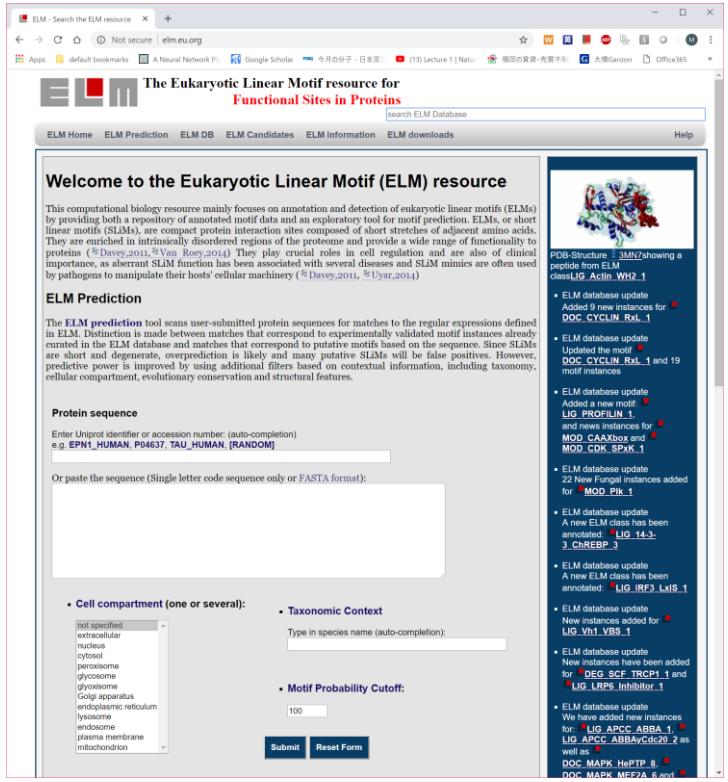
E3NET:

E3-substrate interactions



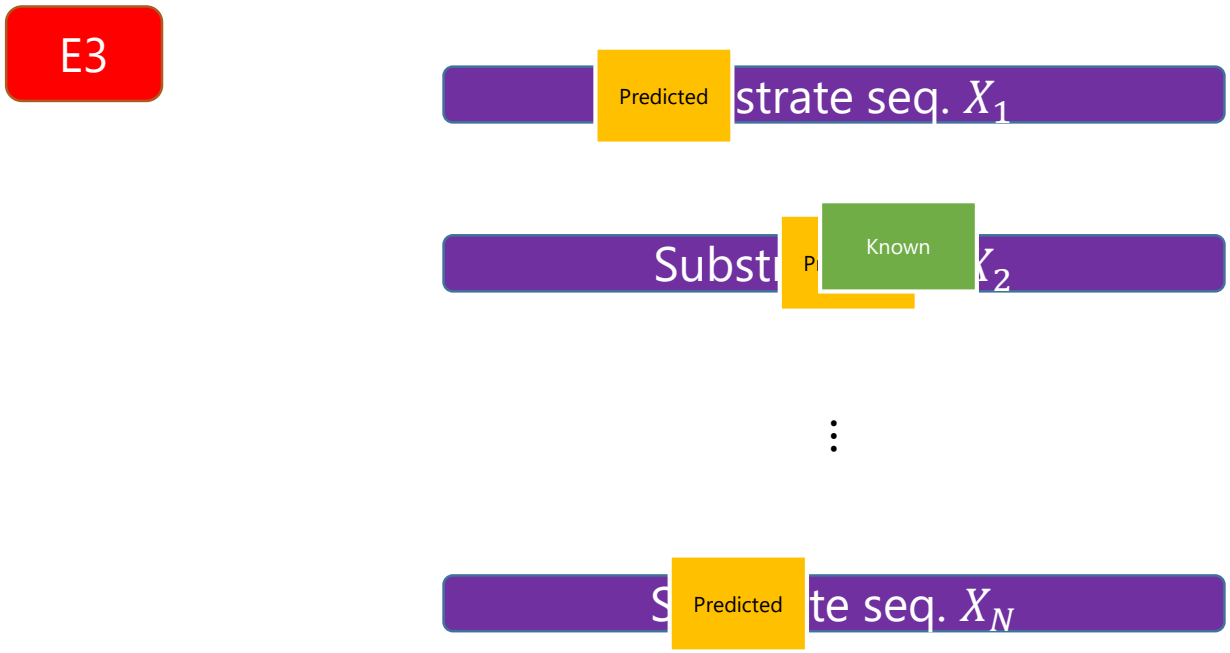
ELM: DEG class motifs:

E3 binding sites of protein sequences



Contents

Problem: Estimate the motif of binding site of an E3



- Performance measure =#position covered by known and predicted motifs simultaneously / W (motif width)
- 36 E3-specific sets of substrate proteins, some of which have known motifs.

Contents

Posterior probability distribution for motif identification

- $\boldsymbol{X} = (X_1, \ldots, X_N)$: N given sequences.
- $\boldsymbol{Z} = (z_1, \ldots, z_N)$: starting positions of motif occurrences.
- $\boldsymbol{\theta}_{0:W} = (\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_W)$: $W + 1$ categorical distributions of letters.
- Posterior:

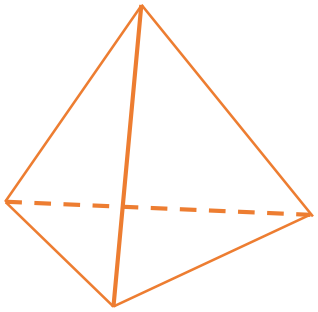
$$p(\boldsymbol{Z}, \boldsymbol{\theta}_{0:W} | \boldsymbol{X}) \propto p(\boldsymbol{X} | \boldsymbol{Z}, \boldsymbol{\theta}_{0:W}) \cdot p(\boldsymbol{Z}) \cdot p(\boldsymbol{\theta}_{0:W})$$

MEME: meme-suite.org

Gibbs Motif Sampler:

Likelihood function $L_{pssm}(\boldsymbol{Z}, \boldsymbol{\theta}_{0:W} | \boldsymbol{X})$

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



Symbol	Probability
A	1/10
C	2/10
G	3/10
T	4/10

$$\frac{2}{10} \cdot \frac{2}{10} \cdot \frac{3}{10}$$

Categorical distribution

Prior distributions and collapsing

$p(\boldsymbol{Z})$:

Uniform distribution

$p(\boldsymbol{\theta}_{0:W})$:

Dirichlet distribution

∵ conjugate prior of the categorical distribution.

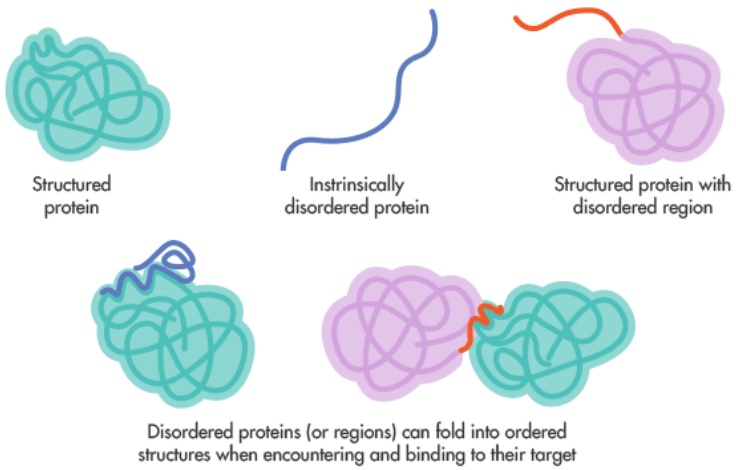
- Recall posterior:

  $p(\boldsymbol{Z}, \boldsymbol{\theta}_{0:W}|\boldsymbol{X}) \propto p(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{\theta}_{0:W}) \cdot p(\boldsymbol{Z}) \cdot p(\boldsymbol{\theta}_{0:W})$

- Collapsed posterior:

$$p(\boldsymbol{Z}\,|\boldsymbol{X}) = \int p(\boldsymbol{Z}, \boldsymbol{\theta}_{0:W}|\boldsymbol{X}) d\boldsymbol{\theta}_{0:W}$$

Modeling: Prior knowledge



E3 binding sites are often located in disordered regions of substrates [Guharoy *et al.*, 2016]

https://www.quantamagazine.org/how-disordered-proteins-are-upending-molecular-biology-20170118/

Prior of starting position z_i

$disorder_{i,j}$: disorderness of $X_{i,j}$

$$p(z_i) \propto \left(\prod_{w=1}^{W} disorder_{i,z_i+w-1} \right)^{\frac{1}{W}}$$

$$p(\boldsymbol{Z}) \propto \prod_{i=1}^{N} p(z_i)$$

Modeling: 2nd likelihood based on amino acid indexing $L_{aai}(\boldsymbol{Z}\,|\boldsymbol{X})$

Resulting likelihood: $p(\boldsymbol{X}|\boldsymbol{Z},\boldsymbol{\theta}_{0:W}) \propto L_{pssm}(\boldsymbol{Z},\boldsymbol{\theta}_{0:W}|\boldsymbol{X}) \cdot L_{aai}(\boldsymbol{Z}\,|\boldsymbol{X})$

⋯ A P A P P V ⋯
⋯ A P P A A V ⋯
⋯ G P R A G V ⋯

Symbol	Probability
A	
R	
N	
D	
C	
Q	
E	
G	
H	
I	
L	
K	
M	
F	
P	
S	
T	
W	
Y	
V	

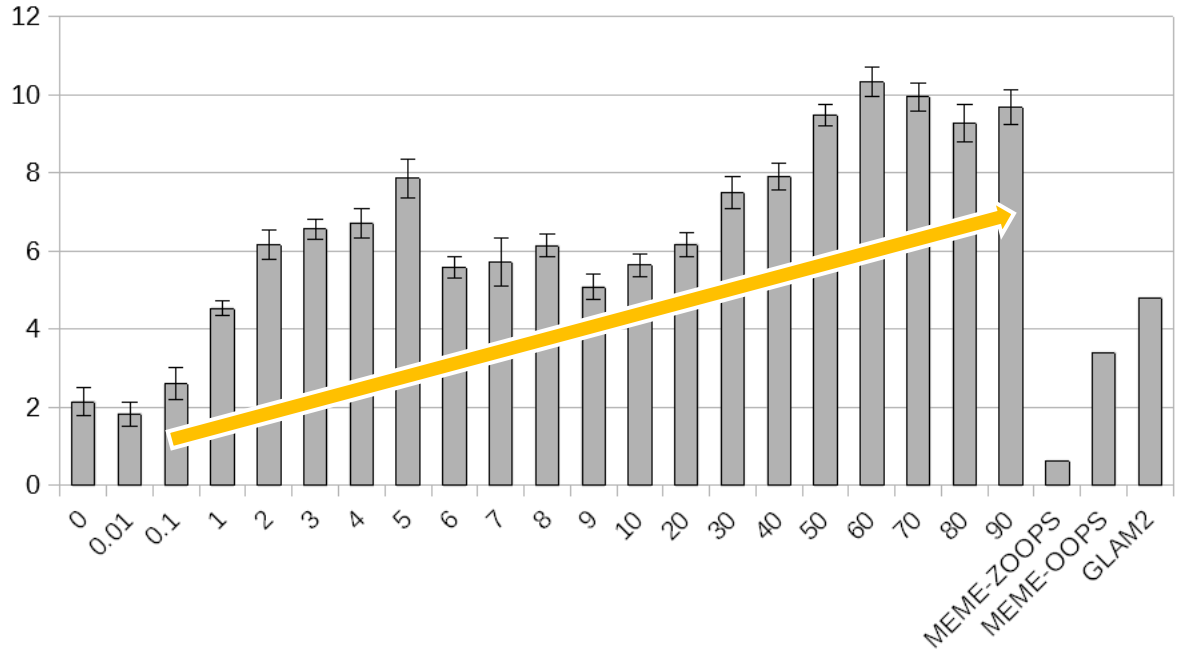
Symbol	Probability
A	a (non-polar)
G	
I	
L	
M	
F	
P	
W	
V	
N	b (polar)
C	
Q	
S	
T	
Y	
D	c (negatively charged)
E	
R	d (positively charged)
H	
K	

ELM DB motif:
DEG_APCC_DBOX_1
.R..L..[LIVM].

Penalize the inconsistency among the mapped symbols

Contents

Result:
Effectiveness of disorder prior

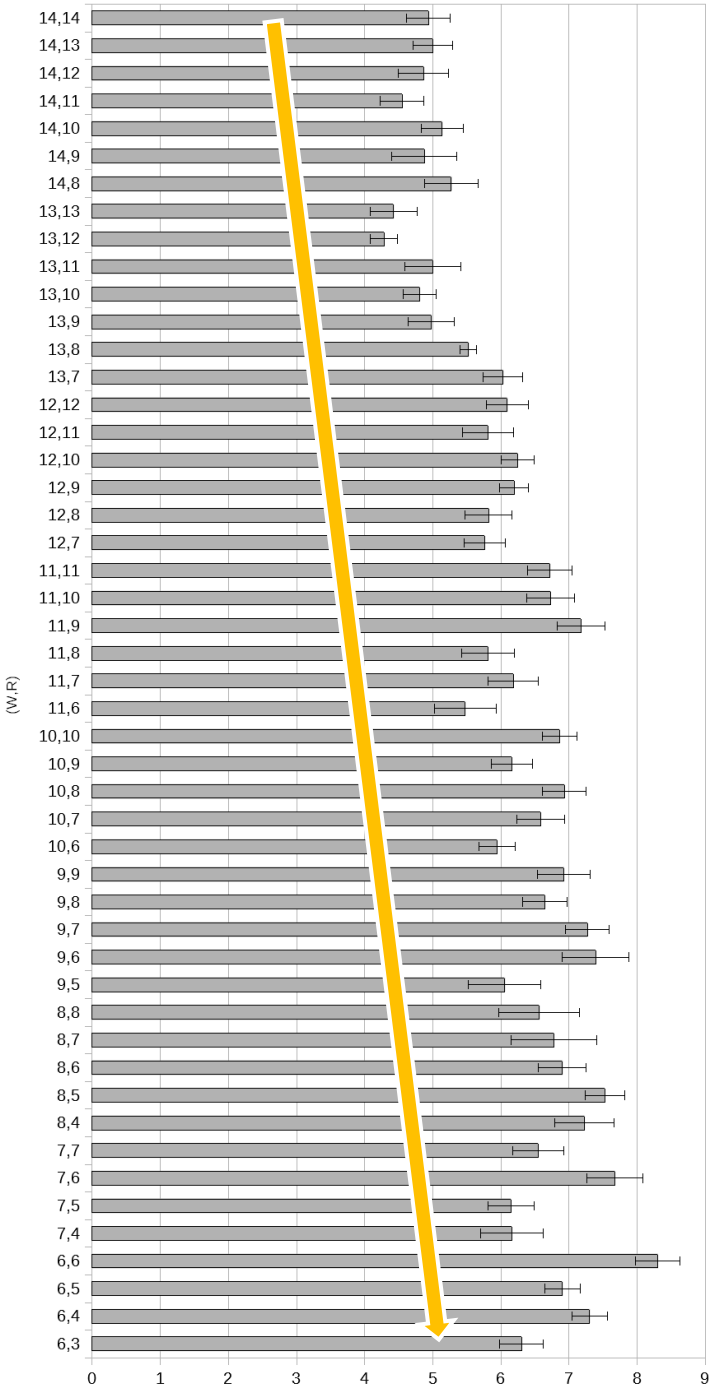


c_d: coefficient of disorder prior.
(W,R)=(9,6)

W: motif width

R: #selected columns as motif part)

Result:
Effectiveness of (W,R)



Example: E3 SIAH2_HUMAN

			.	P	.	A	.	V	.	P
Substrate	Pos	*	*	*		*		*		*
O14974	390	A	A	**P**	V	**A**	V	T	T	P
O75376	124	A	A	V	L	P	L	**V**	H	**P**
O75925	562	A	A	A	A	**A**	A	**V**	S	D
P29590	437	A	Q	**P**	M	**A**	V	**V**	Q	S
P37840	89	A	A	A	T	G	F	**V**	K	K
P43146+	1330	T	I	**P**	T	**A**	C	**V**	R	**P**
Q16633+	45	P	A	**P**	T	**A**	V	**V**	L	**P**
Q92667	284	A	A	**P**	A	P	P	**V**	A	D
Q9GZT9	77	P	A	**P**	P	**A**	A	**V**	P	**P**
Q9H6Z9	62	A	G	**P**	R	**A**	G	**V**	S	K

Contents

Future works

- Improvement of the likelihood function based on amino acid indexing.
- Simulated annealing as a postprocessing.