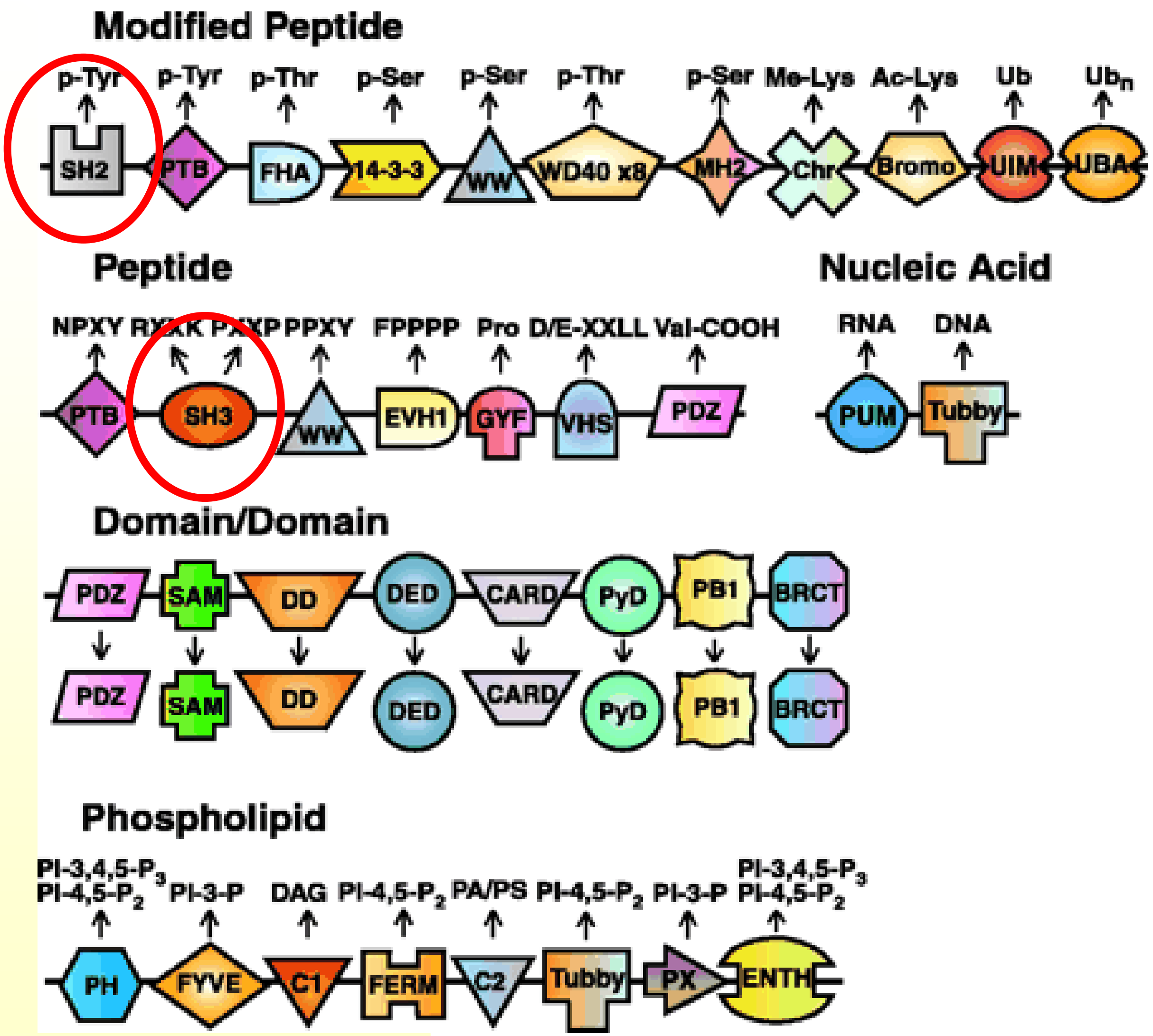
Computational study of the binding specificities of SH2 and SH3 domains

Wei Wang

Department of Chemistry and Biochemistry
Center for Theoretical Biological Physics

UCSD

Modular design of protein-protein interactions:



Pawson and Nash, Science, 300, 445, 2003.

Identification of protein-peptide interactions is challenging:

1. Issues with peptide library screening

   A. still challenging to identify interacting partners given the binding motif.

   B. may be biased by the artifacts of fixing peptides on the surface and/or the strong binding peptides not existing in the genome.

2. Issues with high throughput studies

   Domain-peptide interactions are under-represented as the interactions are weak and transient.

3. Calculation of binding free energy for domain-peptide complex is time consuming.

The first approach:

1. Roughly estimate the binding affinities of thousands of peptides selected from the human proteome.

2. Classify these peptides into binder and non-binder categories based on sequence and binding affinity.

3. Build a Hidden Markov Model (HMM) from binders and search the human proteome.

4. Remove false positives using conservation

5. Estimate the binding affinities of the top 100 candidates

6. Repeat 1-5.

Bill McLaughlin

Advantages of the first approach:

1. **Do not require very accurate binding affinity calculation and only need to separate two distributions.**

2. **Not biased by non-physiological strong binders**

   All peptides present in the human proteome.

3. **Take the structural information into account .**

The first approach:

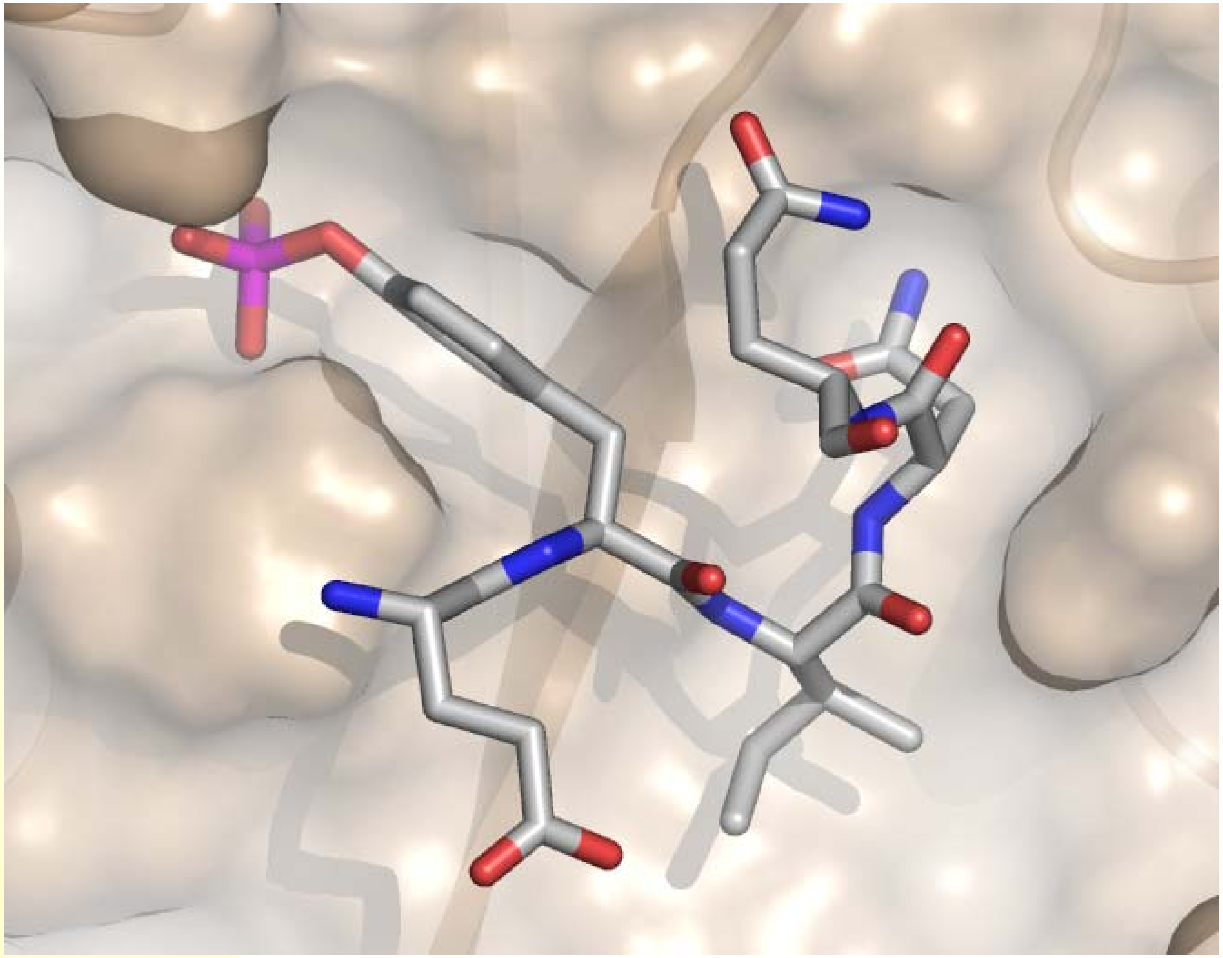
1. Roughly estimate the binding affinities of thousands of peptides selected from the human proteome.

   A. Model the complex structure from a known complex structure using a rotamer library;

   B. Optimize the complex structure using AMBER;

   C. Estimate binding free energy using MM/PBSA.

Known binder to the Grb2 SH2 domain



Sequence: Glu pTyr Ile Asn Gln

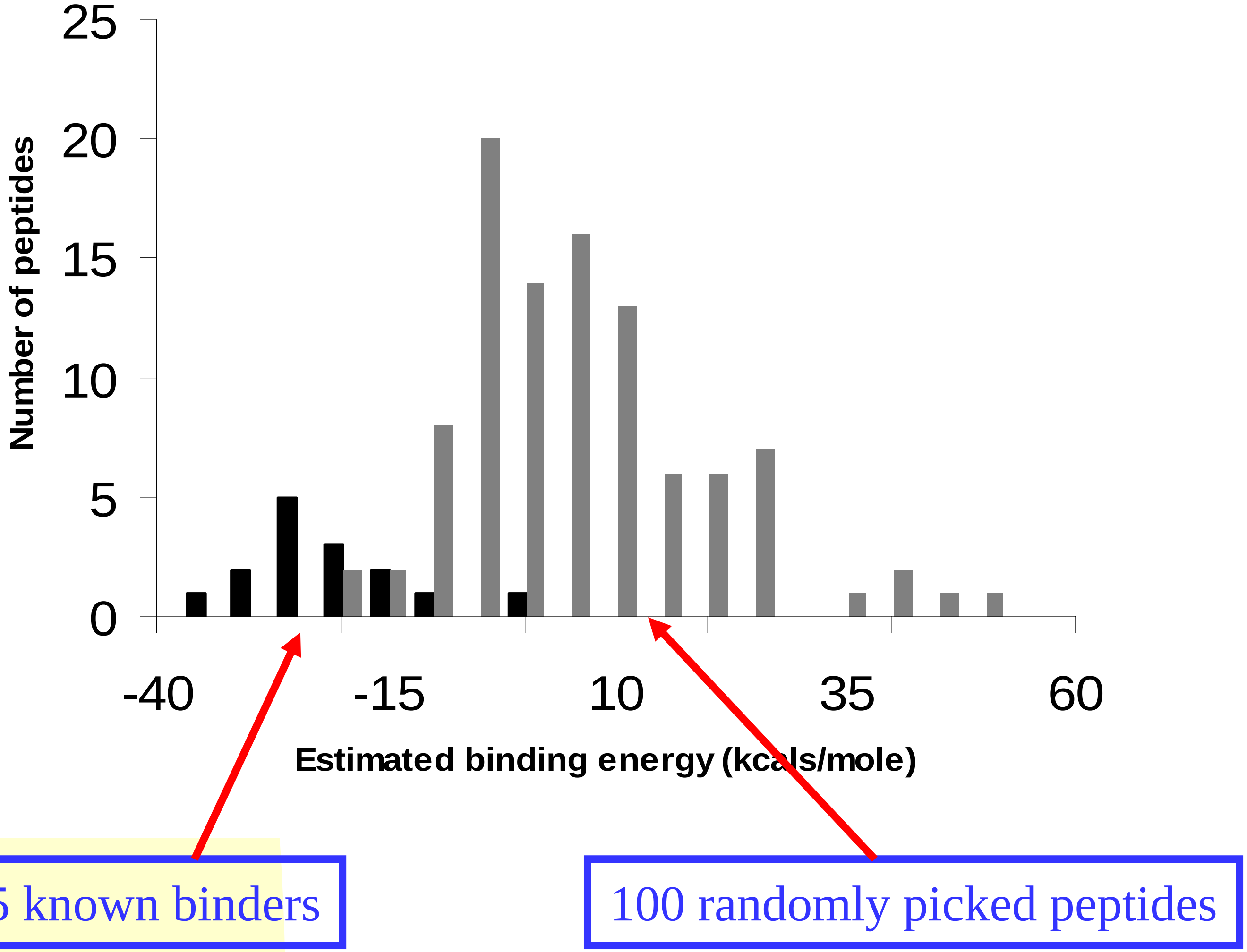
The MM/PBSA (Molecular Mechanism/Poisson-Boltzmann Solvent Area) method

$$G = E_{MM} + G_{polar} + G_{non-polar} - TS$$

$$\Delta G_{bind} = G_{complex} - G_{protein} - G_{ligand}$$

$$= \Delta E_{MM} + \Delta G_{polar} + \Delta G_{nonpolar} - T\Delta S$$

The binders and non-binders have distinct distributions. (Select 1400 peptides from the human proteome + 15 known binders)



15 known binders

100 randomly picked peptides

The first approach:

1. Roughly estimate the binding affinities of thousands of peptides selected from the human proteome.

2. Classify these peptides into binder and non-binder categories based on sequence and binding affinity.

Clusters created using both sequence and energy for the Grb2 dataset of peptides. Cluster “4” labeled as the binding cluster.

1	2	3	4	5	6	Contents
357	13	262	118	425	225	Random Peptides
1	0	0	14	0	0	Known binders

Sequence only

1	
1400	Random
15	Known

Energy only

1	2	3	4	
509	13	218	660	Random
14	0	0	1	Known

The first approach:

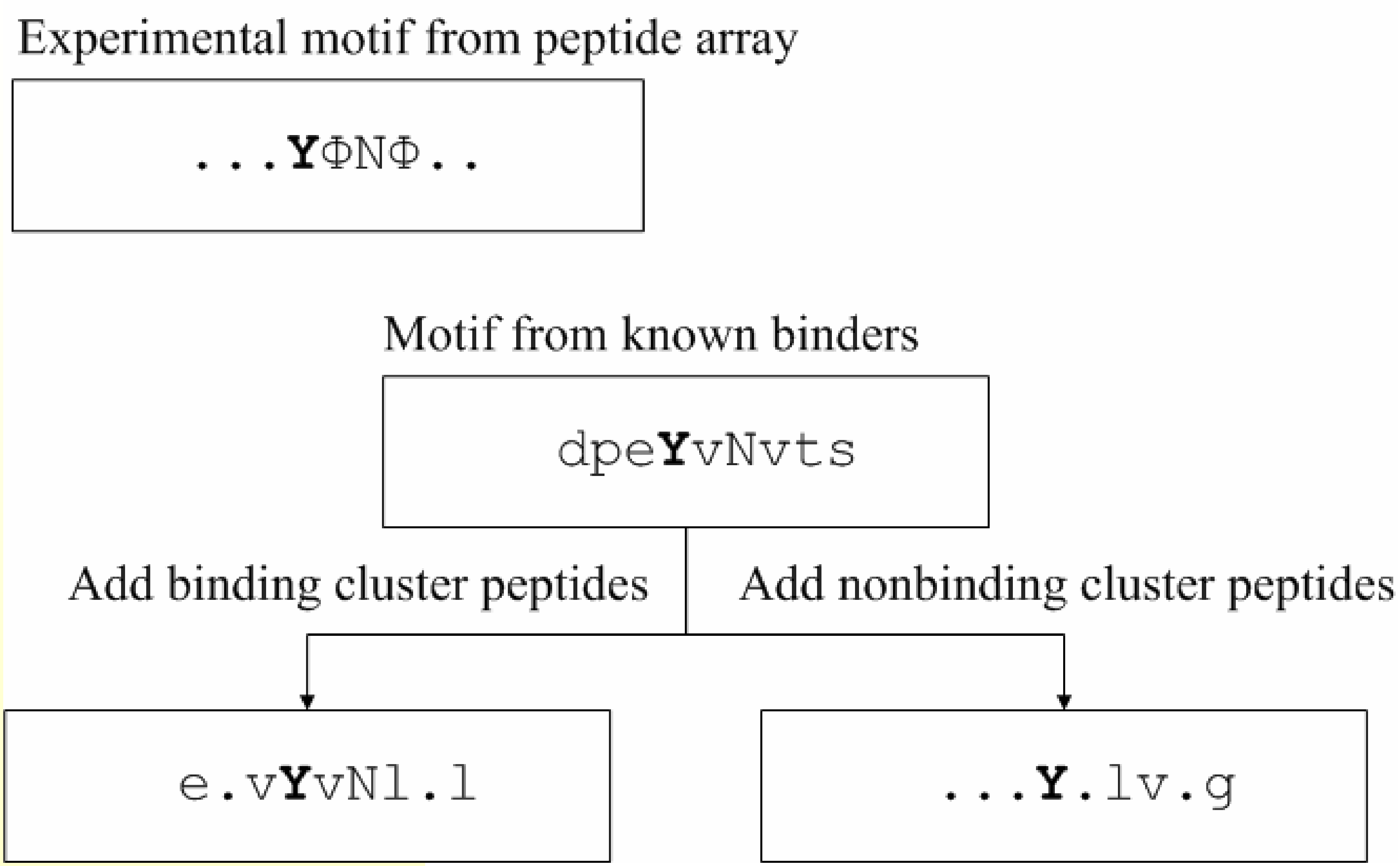
1. Roughly estimate the binding affinities of thousands of peptides selected from the human proteome.

2. Classify these peptides into binder and non-binder categories based on sequence and binding affinity.

3. Build a Hidden Markov Model (HMM) from binders and search the human proteome.

   - Using only the 15 known binders
   - Using peptides in the binding cluster
   - Using known binders plus peptide sequences from the nonbinding clusters

Grb2 SH2 binding sequence motifs (majority rule given by the HMMs)



Experimental motif from peptide array

...**Y**ΦNΦ..

Motif from known binders

dpe**Y**vNvts

Add binding cluster peptides

Add nonbinding cluster peptides

e.v**Y**vNl.l

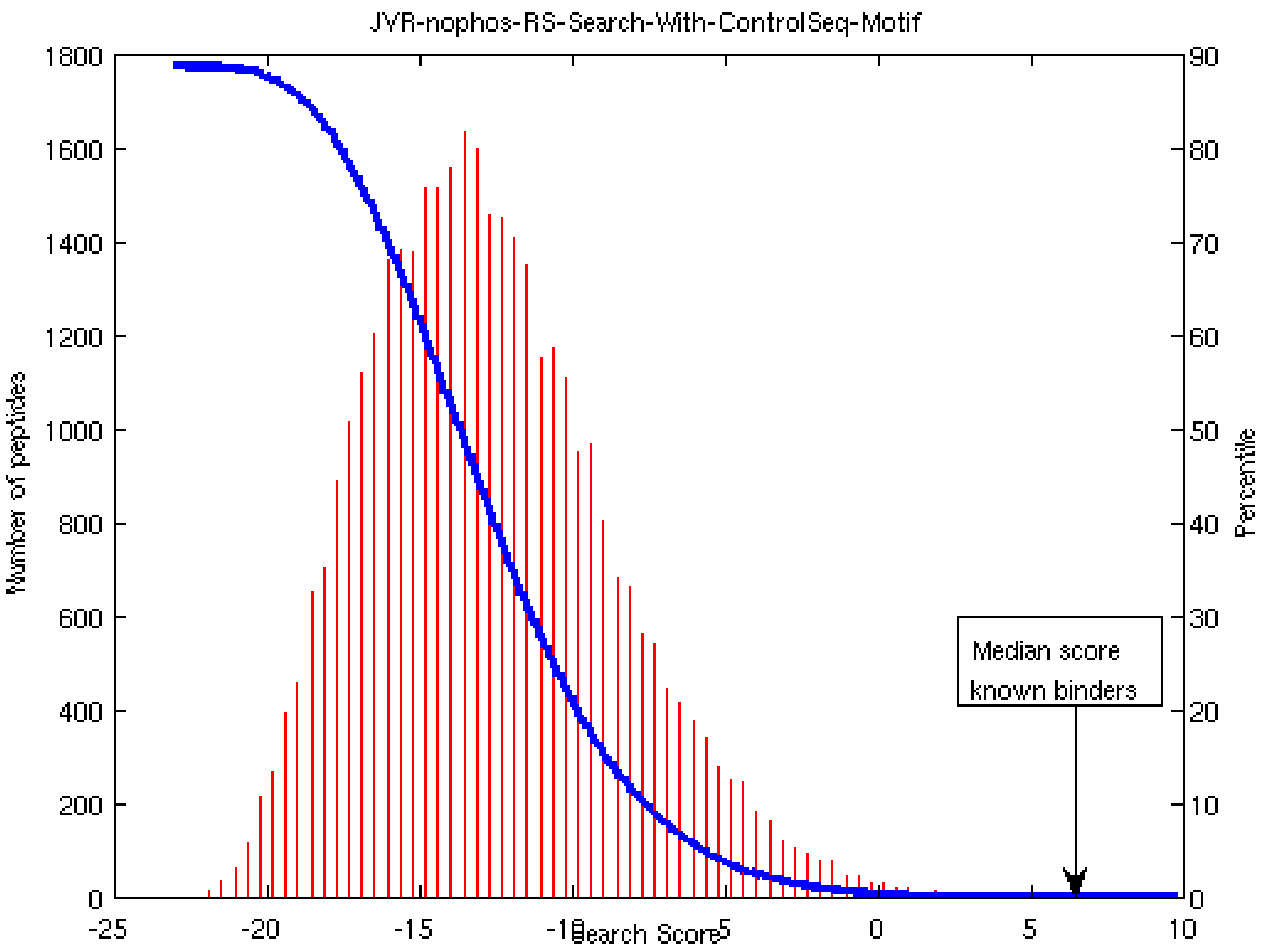
...**Y**.lv.g

Database screening:

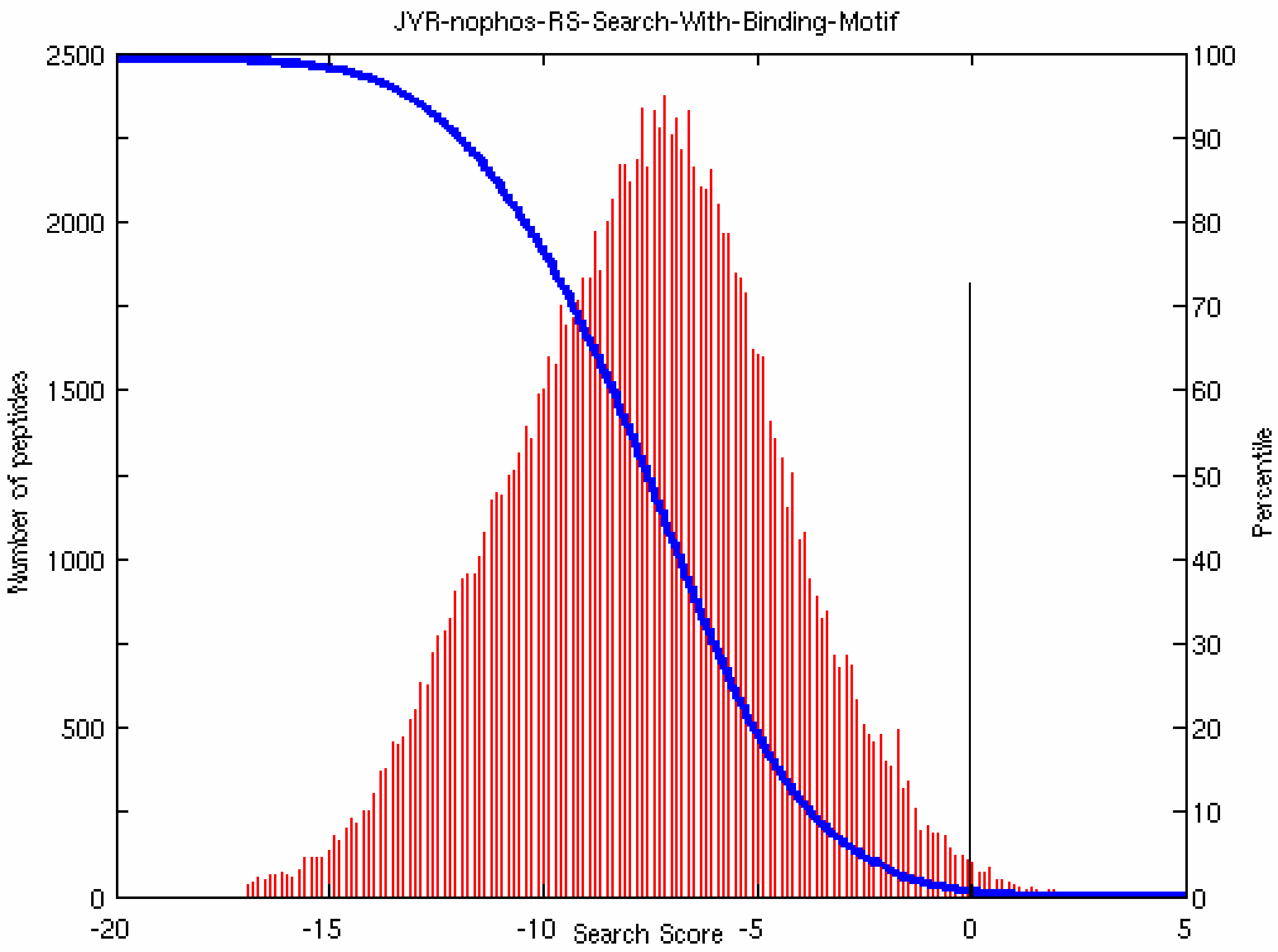
- Extract 174,604 peptides with xxxYxxxx sequences from the human proteins in SWISS-PROT

- Score all of the peptides using each of the HMMs

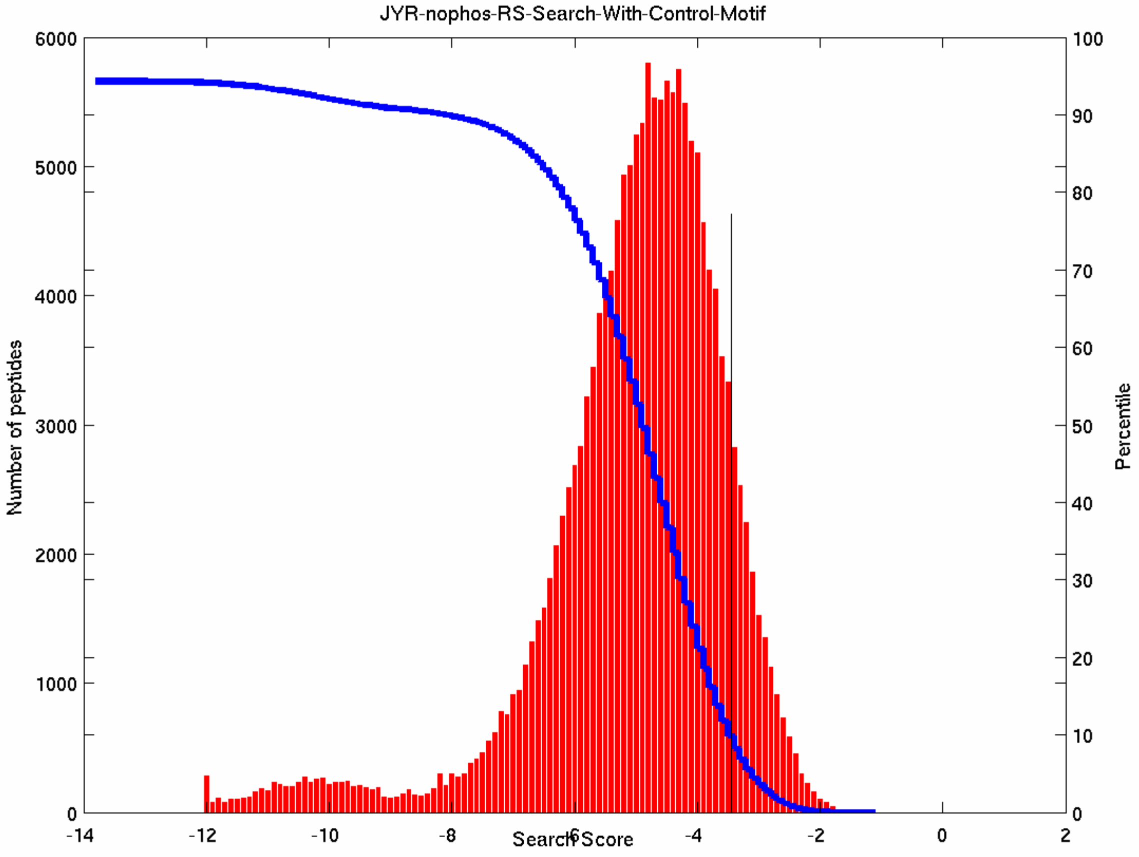
Known binders motif search



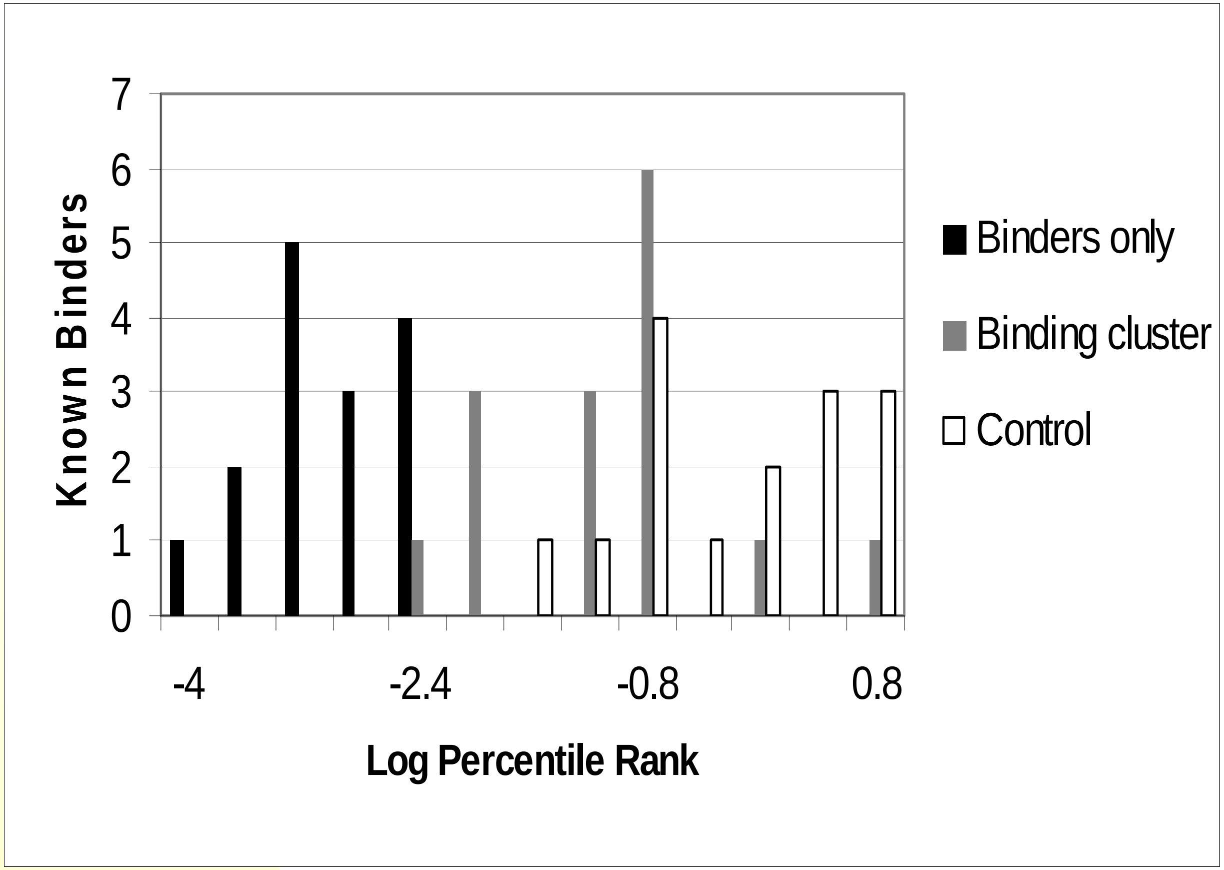
Search with binding motif (HMM created with binding cluster peptide sequences)



HMM of known binders plus sequences from the non binding clusters



Grb2 HMMs search results summary



P-value of t-test comparing known binding ranks using binding cluster HMM and the control HMM = 0.032

Examine the top 100 hits of each search

1. Search with known binder HMM retrieved the known binders with little more.

2. Search with binding cluster HMM retrieved many possible binders and one documented case (UFO, ranked 46 in our prediction but only 227 in Scansite output).

3. Search with control HMM retrieved no viable candidates

The first approach:

1. Roughly estimate the binding affinities of thousands of peptides selected from the human proteome.

2. Classify these peptides into binder and non-binder categories based on sequence and binding affinity.

3. Build a Hidden Markov Model (HMM) from binders and search the human proteome.

4. Remove false positives using conservation

Examples of conserved peptides: UFO

```
Blast alignment for an example top hit: UFO protein

Human: 780 ELNPQDRPSFTELREDLENTLKALPPAQEPDEILYVNMDEGGGYPEPPGAAGGADPPTQP 839
           ELNP+DRPSF ELREDLENTLKALPPAQEPDEILYVNMDEGG + EP GAAGGADPPTQP
Mouse: 781 ELNPRDRPSFAELREDLENTLKALPPAQEPDEILYVNMDEGGSHLEPRGAAGGADPPTQP 840

Comparison to the Grb2 binding motif

Grb2 binding motif  *->e.vYvNl.l<-*
                       E +Y N+
  UFO             1    EILYVNMDE    9
```

Examples of conserved peptides: Nebulin

```
Blast alignment for an example top hit: Nebulin protein

Human : 2356 KFSSPVDMLGVVLAKKCQELVSDVDYKNYLHQWTCLPDQNDVVQAKKVYELQSENLYKSD 2415
             K++SPVDMLGVVLAKKCQ LVSD DY+NYLHQWTCLPDQNDV+QAKKVYELQSEN+YKSD
Mouse : 241  KYTSPVDMLGVVLAKKCQALVSDADYRNYLHQWTCLPDQNDVIQAKKVYELQSENMYKSD 300
```

```
Comparison to the Grb2 binding motif

Score = 2.1

Grb2 binding motif  *->e.vYvNl.l<-*
                       + +Y N+ +
  Nebulin 2380    1    DaDYRNYlH    9
```

The first approach:

1. Roughly estimate the binding affinities of thousands of peptides selected from the human proteome.

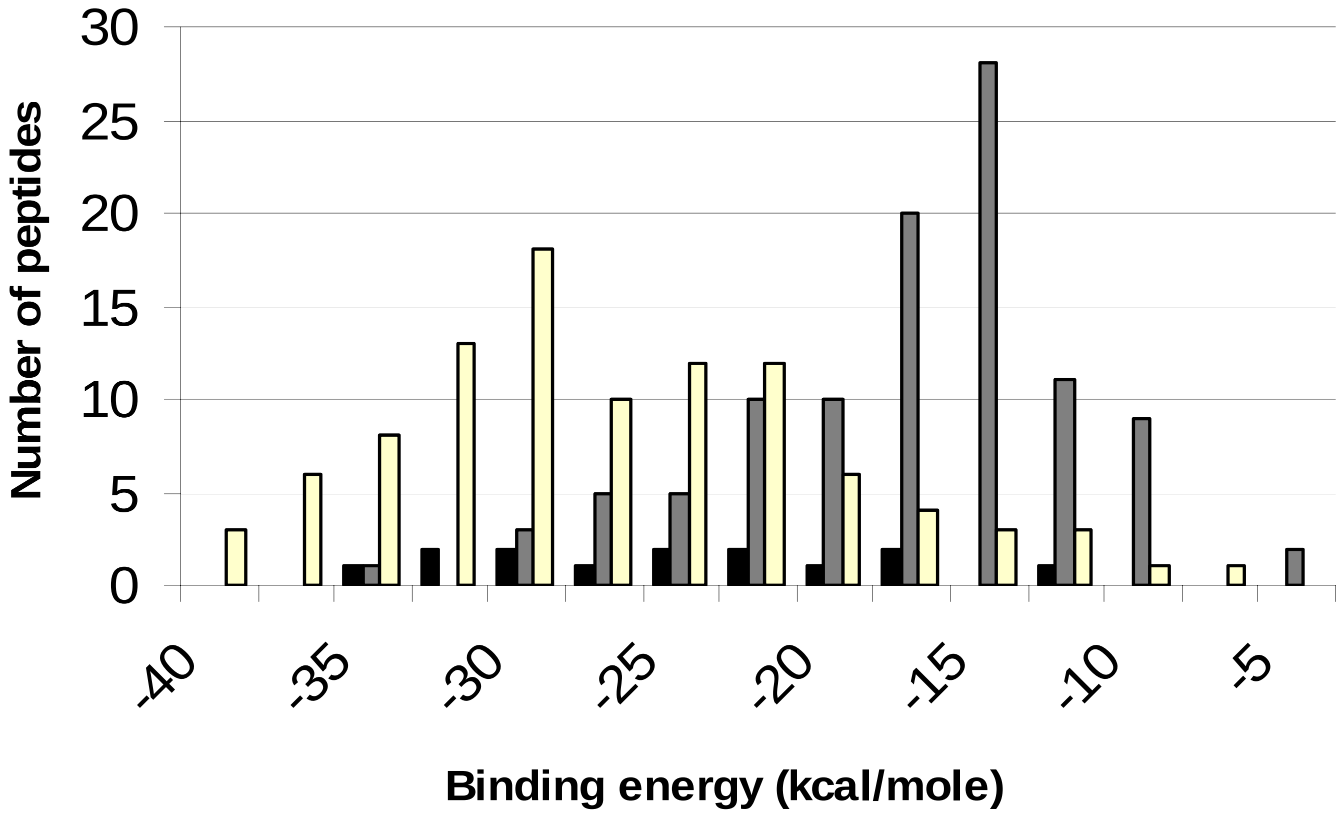
2. Classify these peptides into binder and non-binder categories based on sequence and binding affinity.

3. Build a Hidden Markov Model (HMM) from binders and search the human proteome.

4. Remove false positives using conservation

5. Estimate the binding affinities of the top 100 candidates

The HMM captures both sequence and energy features



known binders (black), 100 random peptides in binding cluster (gray), top 100 predictions (white)

The first approach:

1. Roughly estimate the binding affinities of thousands of peptides selected from the human proteome.

2. Classify these peptides into binder and non-binder categories based on sequence and binding affinity.

3. Build a Hidden Markov Model (HMM) from binders and search the human proteome.

4. Remove false positives using conservation

5. Estimate the binding affinities of the top 100 candidates

6. Repeat 1-5.

Evaluation of the top hits using sequence and energy (binding cluster is Cluster 5)

```
Clustering of top one hundred candidates plus original dataset

   1   2   3   4   5  <-- assigned to cluster
 270  13 510 503 104 | Random peptides
   0   0   0   1  14 | Known binding
   0   0   0   0 100 | Top 100 from search

Cluster probabilities for the top ten candidates

Instance Clus1     Clus2      Clus3    Clus4    Clus5
0                   0        0        0        0        1
1                   0        0        0        0        1
2                   0        0        0        0        1
3                   0        0        0        0        1
4                   0        0        0        0        1
5                   0        0        0        0        1
6                   0        0        0        0        1
7                   0        0        0        0        1
8                   0        0        0.00001  0        0.99999
9                   0        0        0        0        1
10                  0        0        0        0        1
```

The second approach:

1. Computational point mutation to generate a Position Specific Scoring Matrix (PSSM)

   Better consideration of conformational flexibility

2. Scan the database using this PSSM.

The second approach:

1. Computational point mutation to generate a Position Specific Scoring Matrix (PSSM)
   - A. mutate each residue to other 19 amino acids
   - B. calculate the binding free energy using MM/PBSA
   - C. take the free energy difference between the mutated and the template peptides as the entry in the PSSM

2. Scan the database using this PSSM.

Residue	Position									
	P_{-6}*	P_{-5}	P_{-4}	P_{-3}	P_{-2}	P_{-1}	P_0	P_1	P_2	P_3
A	0.00	6.43	0.00	6.70	0.00	5.44	6.30	-0.37	2.56	1.75
R	4.00	7.44	13.62	9.14	3.50	7.39	17.18	8.60	8.80	15.02
N	2.00	2.90	1.40	2.32	0.79	5.16	6.15	6.57	4.98	3.29
D	4.00	20.85	4.15	16.36	12.31	18.78	11.43	-0.50	2.30	14.99
C	0.00	4.41	-0.38	1.59	-0.46	3.41	6.26	-0.10	1.92	3.45
Q	0.00	18.36	0.69	3.29	13.63	2.77	11.28	1.18	1.32	1.34
E	4.00	20.69	8.56	11.99	13.68	27.11	14.18	6.55	6.62	10.71
G	0.00	9.27	0.21	3.32	0.59	7.70	6.19	0.89	6.56	11.63
H	0.00	10.73	1.14	3.30	0.65	5.15	3.98	1.94	6.46	1.39
I	0.00	10.84	1.13	1.53	4.28	6.23	6.88	5.40	0.33	5.50
L	0.00	3.52	1.59	1.47	7.57	6.08	7.52	0.98	1.35	2.98
K	4.00	5.26	11.44	9.14	8.14	5.63	20.98	8.58	13.71	15.21
M	0.00	3.27	3.26	0.27	3.08	6.42	3.96	3.84	0.83	0.48
F	0.00	3.32	2.70	-1.52	8.64	7.29	7.56	5.44	0.77	0.71
S	0.00	8.80	1.09	5.01	0.12	-3.3	9.36	-1.51	5.25	6.45
T	0.00	5.22	-0.18	3.48	-0.94	7.53	4.57	1.02	2.30	5.24
W	0.00	13.58	0.14	-2.12	7.39	3.63	1.39	1.87	1.56	1.63
Y	0.00	3.44	3.98	0.00	3.37	-2.36	5.82	3.96	1.38	2.51
V	0.00	2.00	0.52	3.26	2.44	-2.77	8.46	-0.18	1.71	3.80
P	0.00	0.00	-0.60	0.93	-0.21	0.00	0.00	0.00	0.00	0.00

Rank	Protein	Protein name	Start position	End position	Peptide	Score	Scansite Rank
1	RW1	RW1 protein [Fragment]	1521	1530	SPTPASPSPP	-4.06	Not in the top 2000
2	WASF4 (SCAR2)	Wiskott-Aldrich syndrome protein family member 4	475	484	PPPPSSPSFP	-3.59	Not in the top 2000
3	TREX1	Three prime repair exonuclease 1	107	116	GPPPTVPPPP	-3.38	1194
4	ACRO (ACR, ACRS)	Acrosin [Precursor]	344	353	PPPPPSPPPP	-3.18	40
5	LRRN5 (GAC1)	Leucine-rich repeats neuronal protein 5 [Precursor]	22	31	VVPWHVPCPP	-2.94	Not in the top 2000
6	**SEM6A (SEMA6A)**	**Semaphorin 6A [Precursor]**	**791**	**800**	**MPPMGSPVIP**	**-2.89**	**Not in the top 2000**
7	HDAC4 (HD4)	Histone deacetylase 4	343	352	LPLYTSPSLP	-2.81	Not in the top 2000
8	**EVL (RNB6)**	**Ena/vasodilator stimulated phosphoprotein-like protein**	**185**	**194**	**PPPPPVPPPP**	**-2.65**	**83**
9	**WASF1 (WAVE1, WAVE-1)**	**Wiskott-Aldrich syndrome protein family member 1**	**347**	**356**	**TPPPPVPPPP**	**-2.65**	**132**
10	YLPM1 (ZAP3, ZAP113)	YLP motif containing protein 1	14	23	YPPPPVPPPP	-2.65	115

Summary:

1. Computational approach and goal
   - A. Identify binding motifs of modular domains
   - B. Identify new physiological interacting partners

2. Readiness of the application to study biological complex

3. Bottleneck:
   - A. Domain-peptide complex structures
   - B. Experimental verification
   - C. Nomenclature (gene names different in databases)

Acknowledgement

Ken Chen
Han-Yu Chuang
Jie Liu
Tingjun Hou
Bill McLaughlin
Robert Shoemaker

Phil Bourne
Andy McCammon
Bing Ren
Yang Xu
Chanfeng Zhao (Illumina)

http://wanglab.ucsd.edu