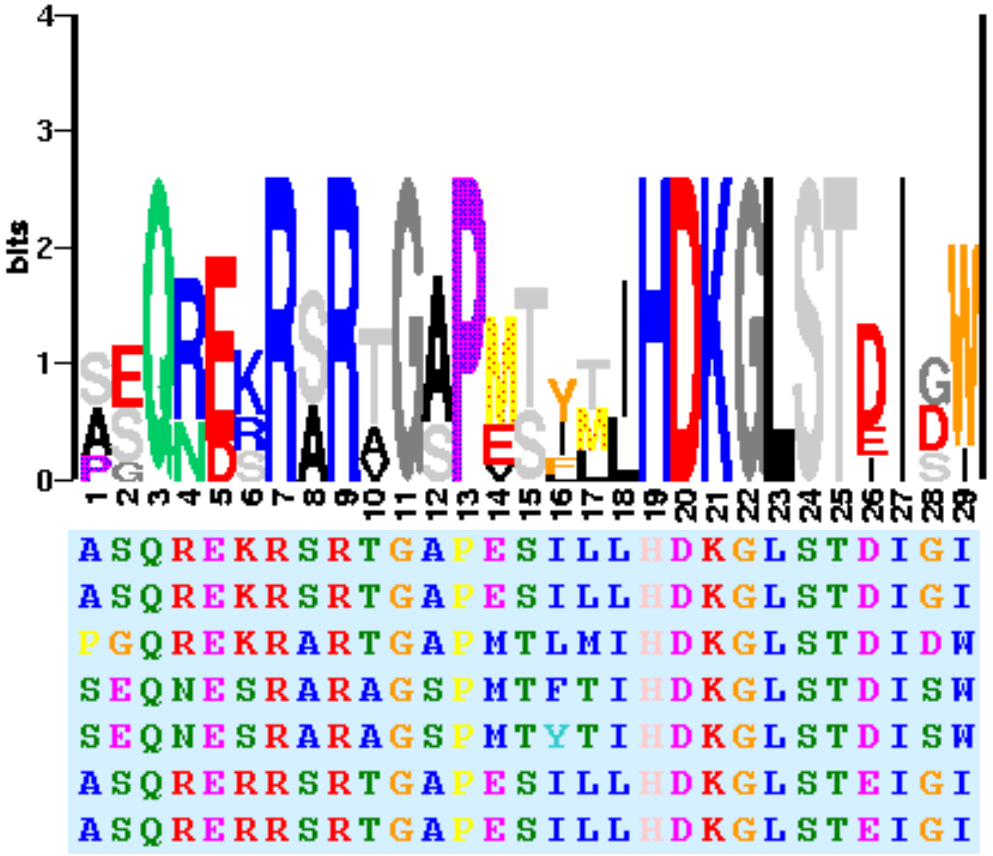
Methods and tools for exploring functional genomics data

William Stafford Noble
Department of Genome Sciences
Department of Computer Science and Engineering
University of Washington

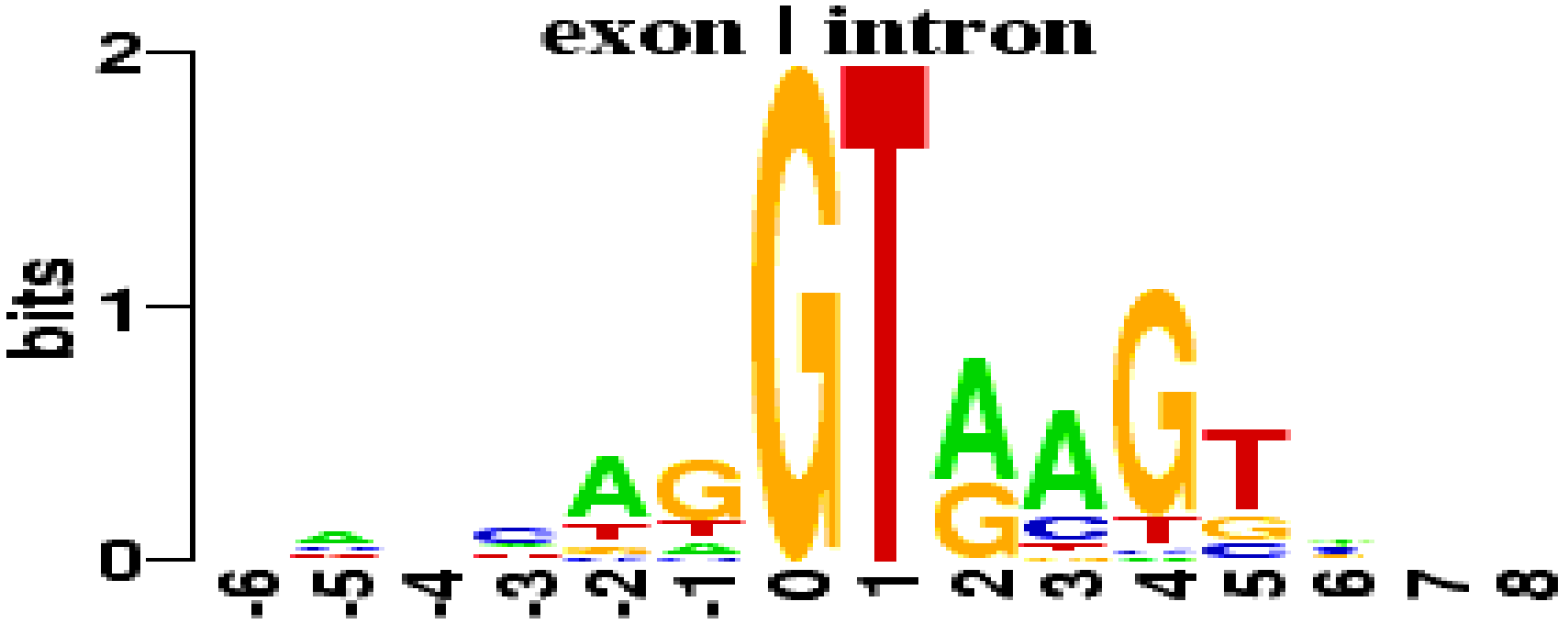
Outline

- Searching for occurrences of a given motif
- High-resolution models of transcription factor binding to DNA
- An embedding approach to remote protein homology detection

Motif in Logo Format

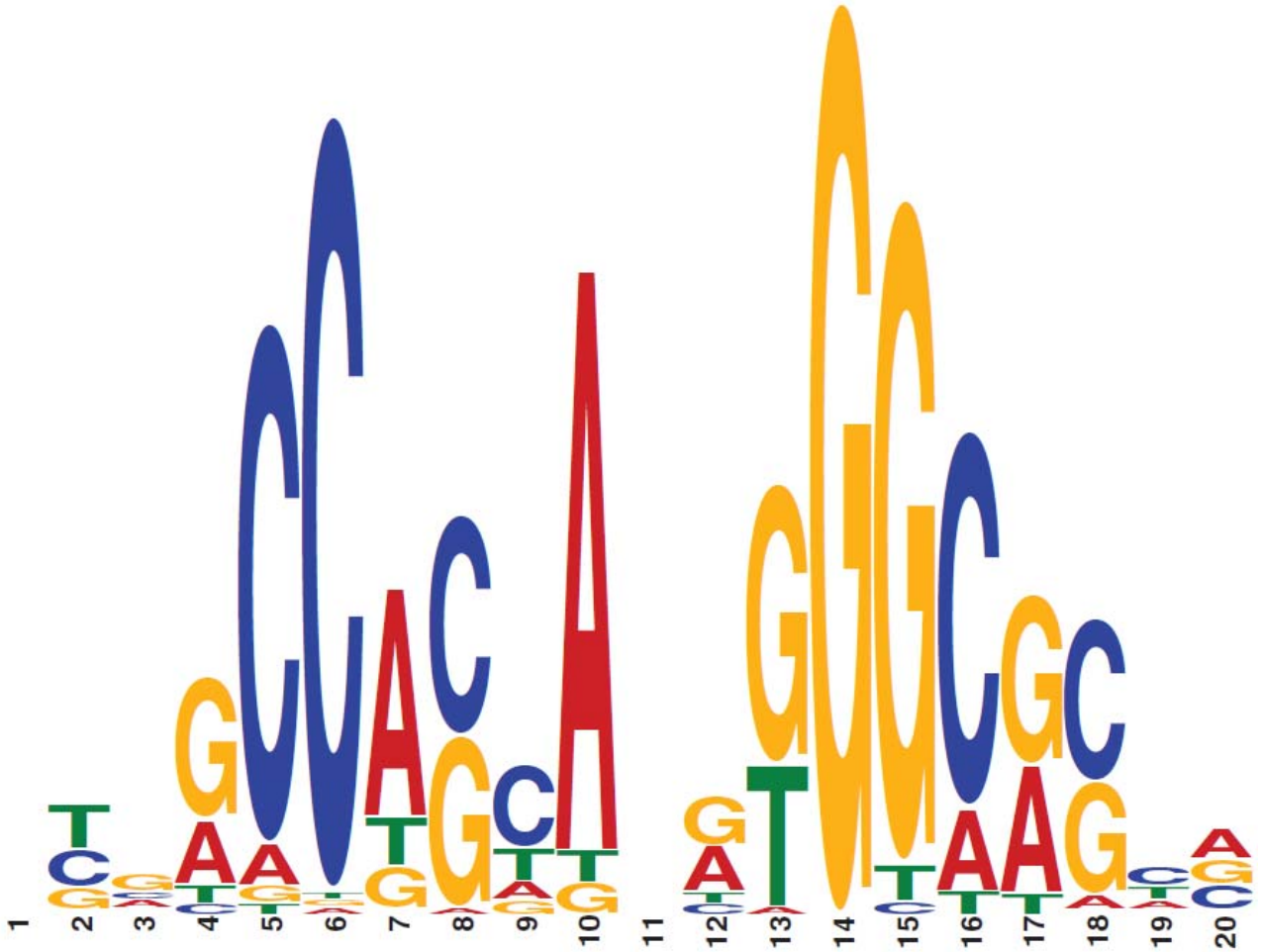


Splice site motif in logo format



weblogo.berkeley.edu

CTCF binding motif



Scoring with a motif

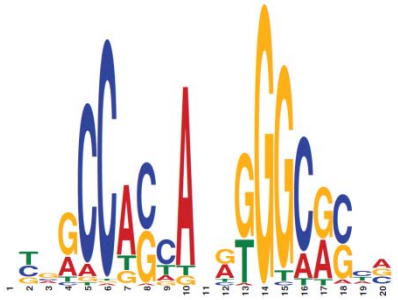
- This motif assigns the sequence NMFWAFGH a score of 0 + -2 + -3 + -2 + -1 + 6 + 6 + 8 = 12.

A	-1	-2	-1	0	-1	-2	0	-2
R	5	0	5	-2	1	-3	-2	0
N	0	6	0	0	0	-3	0	1
D	-2	1	-2	-1	0	-3	-1	-1
C	-3	-3	-3	-3	-3	-2	-3	-3
Q	1	0	1	-2	5	-3	-2	0
E	0	0	0	-2	2	-3	-2	0
G	-2	0	-2	6	-2	-3	6	-2
H	0	1	0	-2	0	-1	-2	8
I	-3	-3	-3	-4	-3	0	-4	-3
L	-2	-3	-2	-4	-2	0	-4	-3
K	2	0	2	-2	1	-3	-2	-1
M	-1	-2	-1	-3	0	0	-3	-2
F	-3	-3	-3	-3	-3	6	-3	-1
P	-2	-2	-2	-2	-1	-4	-2	-2
S	-1	1	-1	0	0	-2	0	-1
T	-1	0	-1	-2	-1	-2	-2	-2
W	-3	-4	-3	-2	-2	1	-2	-2
Y	-2	-2	-2	-3	-1	3	-3	2
V	-3	-3	-3	-3	-2	-1	-3	-3

Searching human chromosome 21 with the CTCF motif

Position	Str	Sequence	Score
19390631	+	TTGACCAGCAGGGGGCGCCG	26.30
32420105	+	CTGGCCAGCAGAGGGCAGCA	26.30
27910537	–	CGGTGCCCCCTGCTGGTCAG	26.18
21968106	+	GTGACCACCAGGGGGCAGCA	25.81
31409358	+	CGGGCCTCCAGGGGGCGCTC	25.56
19129218	–	TGGCGCCACCTGCTGGTCAC	25.44
21854623	+	CTGGCCAGCAGAGGGCAGGG	24.95
12364895	+	CCCGCCAGCAGAGGGAGCCG	24.71
13406383	+	CTAGCCACCAGGTGGCGGTG	24.71
18613020	+	CCCGCCAGCAGAGGGAGCCG	24.71
31980801	+	ACGCCCAGCAGGGGGCGCCG	24.71
32909754	–	TGGCTCCCCCTGGCGGCCGG	24.71
25683654	+	TCGGCCACTAGGGGGCACTA	24.58
31116990	–	GGCCGCCACCTTGTGGCCAG	24.58
29615421	–	CTCTGCCCTCTGGTGGCTGC	24.46
6024389	+	GTTGCCACCAGAGGGCACTA	24.46
26610753	–	CACTGCCCTCTGCTGGCCCA	24.34
26912791	–	GGGCGCCACCTGGCGGTCAC	24.34
20446267	+	CTGCCCACCAGGGGGCAGCG	24.22
21872506	–	TGGCGCCACCTGGCGGCAGC	24.22

Significance of scores



Motif Scanning algorithm

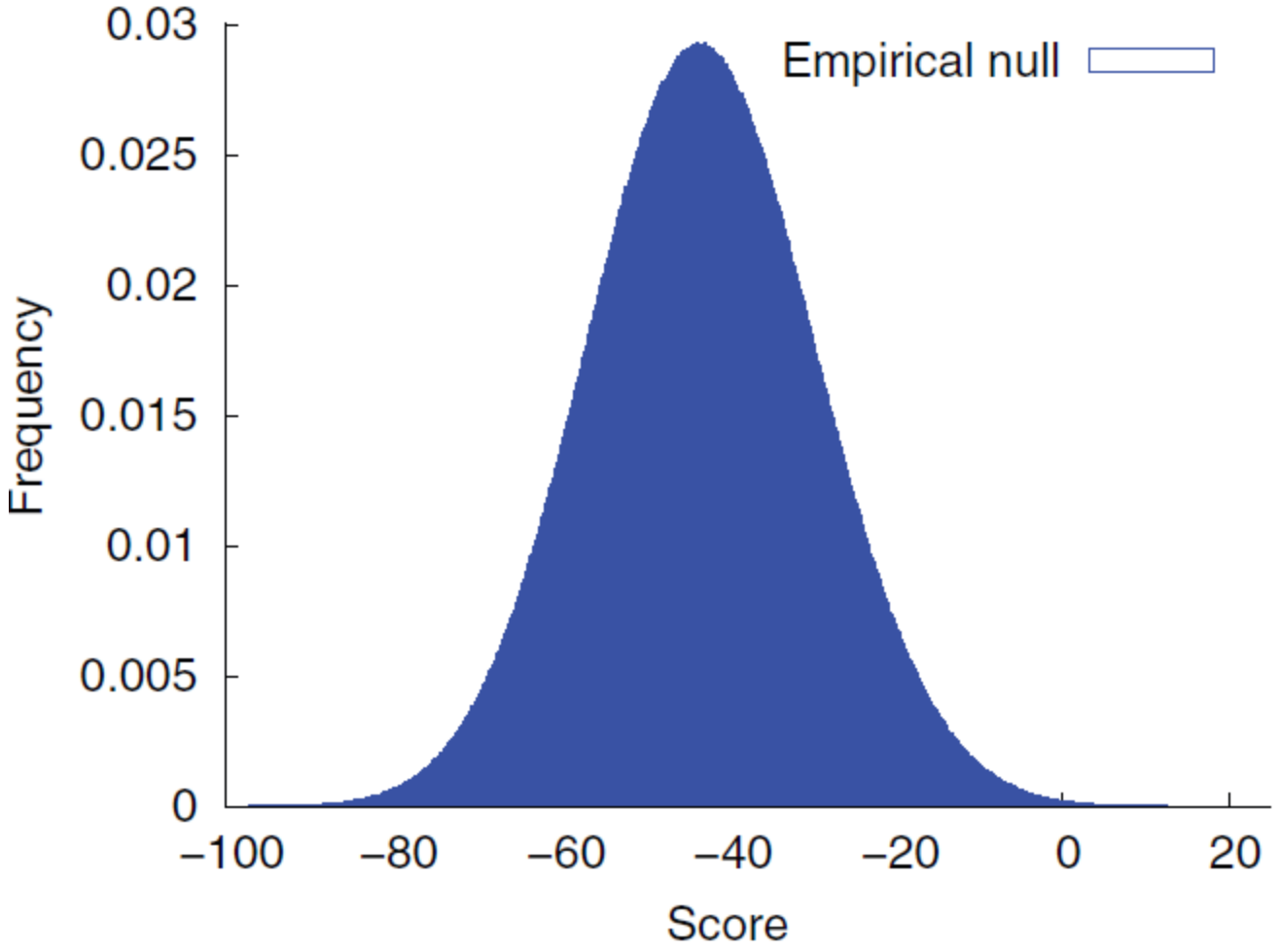
26.30

TTGACCAGCAGGGGGCGCCG

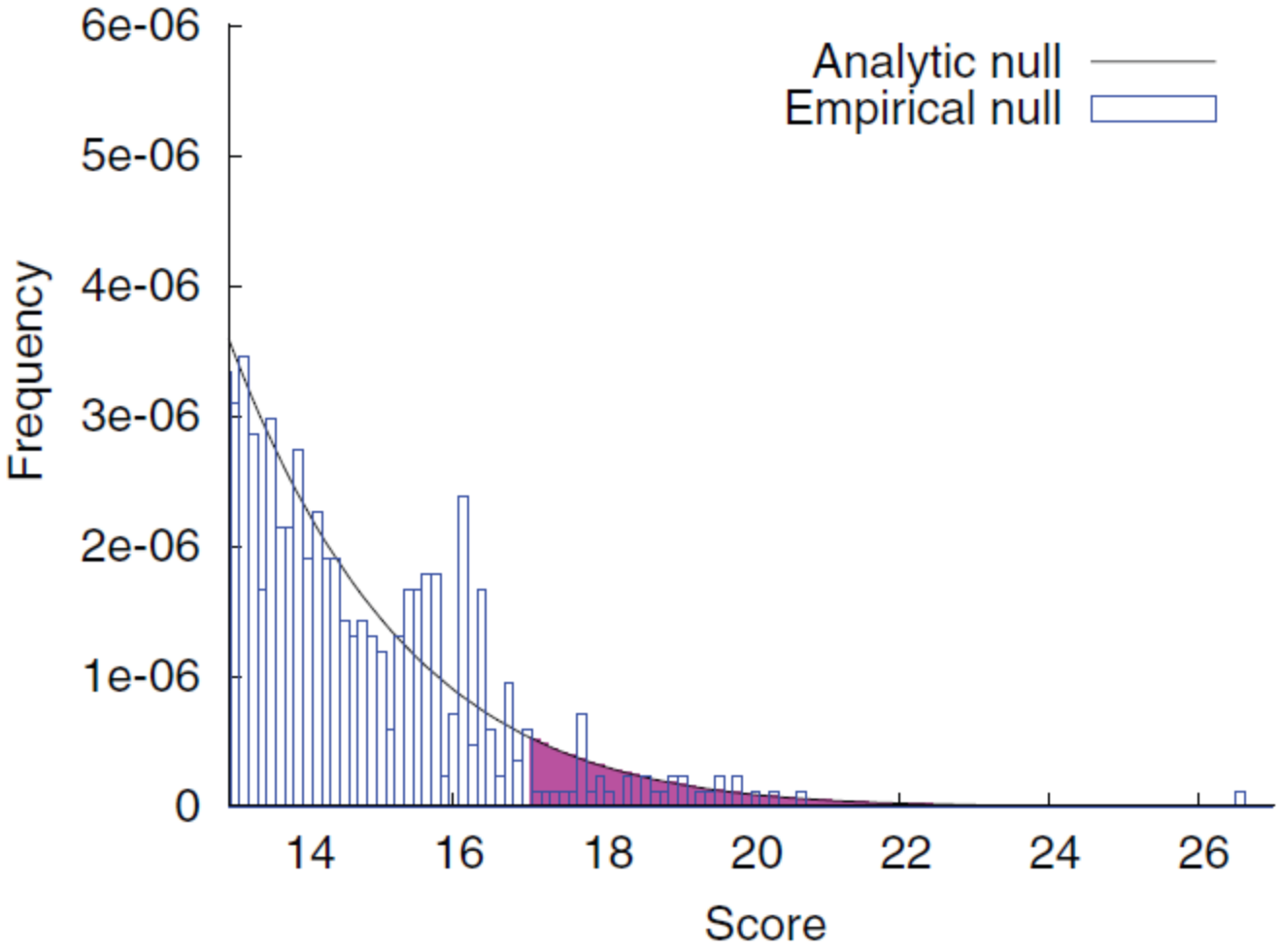
Low score = not a motif occurrence
High score = motif occurrence

How high is high enough?

CTCF empirical null distribution



Poor precision in the tail



Converting scores to p-values

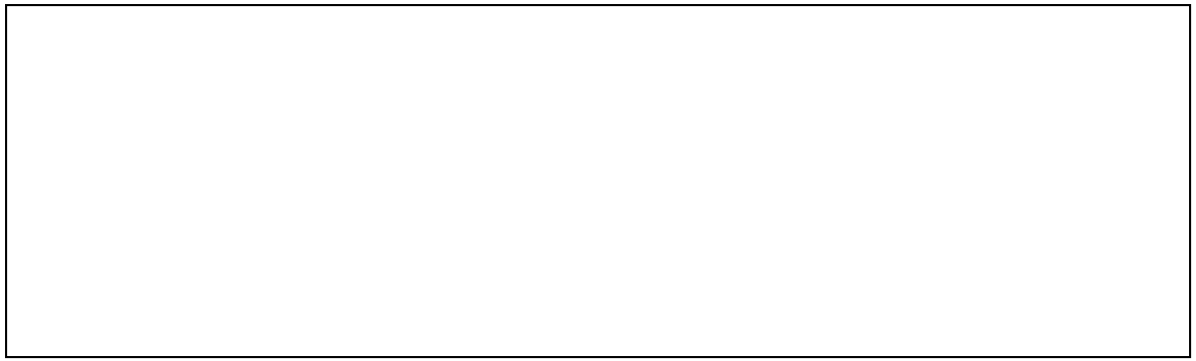
```
A -2.3  1.7  1.1  0.1
C  1.2 -0.3  0.4 -1.0
G -3.0  2.0  0.5  0.8
T  4.0  0.0 -2.1  1.5
```

```
A  10 67 59 44
C  60 39 49 29
G   0 71 50 54
T 100 43 13 64
```

- Linearly rescale the matrix values to the range [0,100] and integerize.

Converting scores to p-values

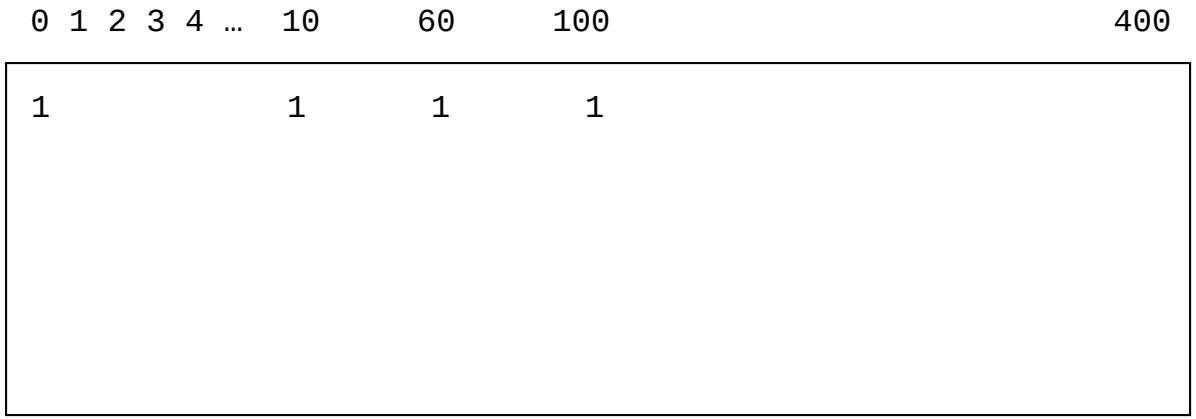
```
  0 1 2 3 4 …                                  400
A  10 67 59 44
C  60 39 49 29
G   0 71 50 54
T 100 43 13 64
```



- Say that your motif has N rows. Create a matrix that is N rows and 100N columns.
- The entry in row i, column j is the number of different sequences of length i that can have a score of j.

Converting scores to p-values

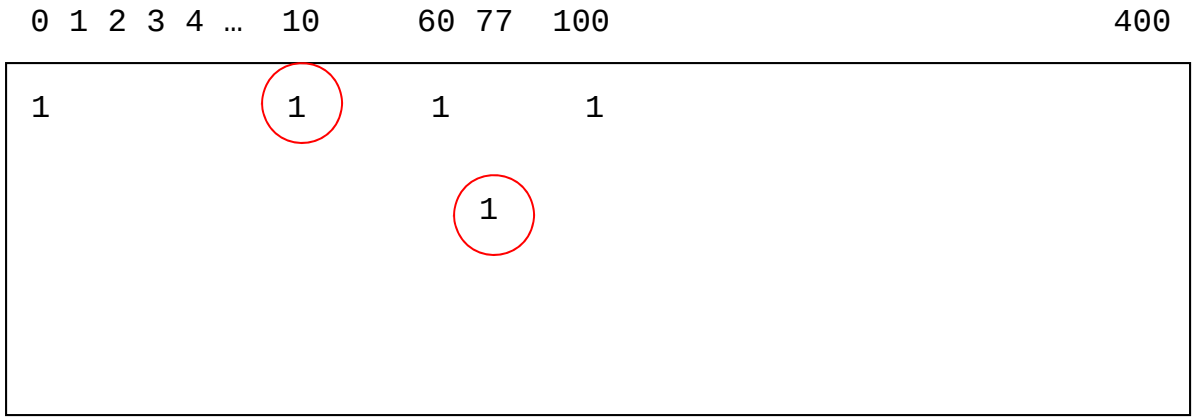
```
A  10 67 59 44
C  60 39 49 29
G   0 71 50 54
T 100 43 13 64
```



- For each value in the first column of your motif, put a 1 in the corresponding entry in the first row of the matrix.
- There are only 4 possible sequences of length 1.

Converting scores to p-values

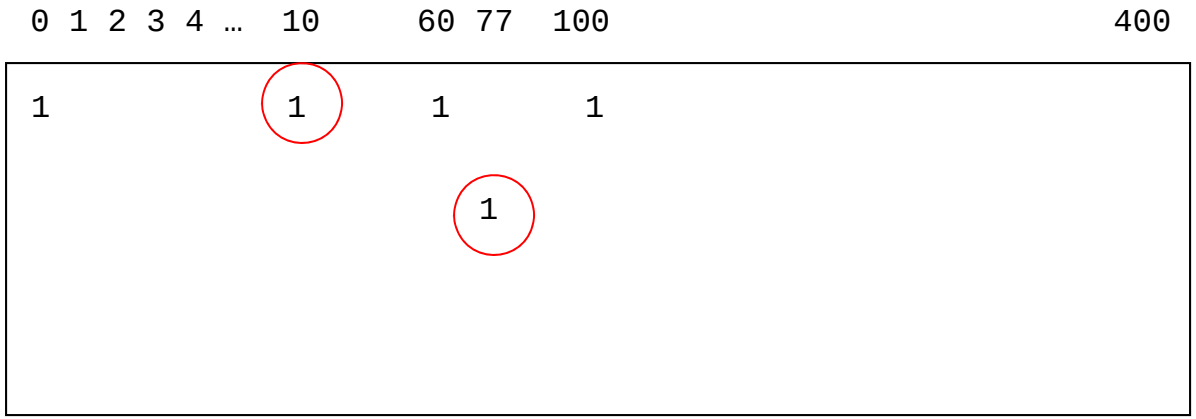
A	10	67	59	44
C	60	39	49	29
G	0	71	50	54
T	100	43	13	64



- For each value x in the second column of your motif, consider each value y in the zth column of the first row of the matrix.
- Add y to the x+zth column of the matrix.

Converting scores to p-values

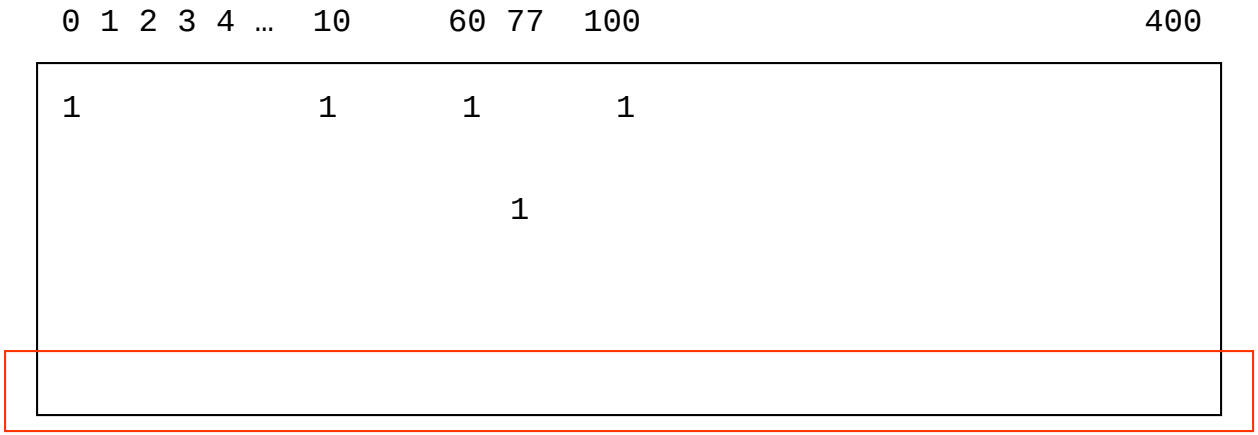
A	10	67	59	44
C	60	39	49	29
G	0	71	50	54
T	100	43	13	64



- For each value x in the second column of your motif, consider each value y in the zth column of the first row of the matrix.
- Add y to the x+zth column of the matrix.
- What values will go in row 2?
  - 10+67, 10+39, 10+71, 10+43, 60+67, …, 100+43
- These 16 values correspond to all 16 strings of length 2.

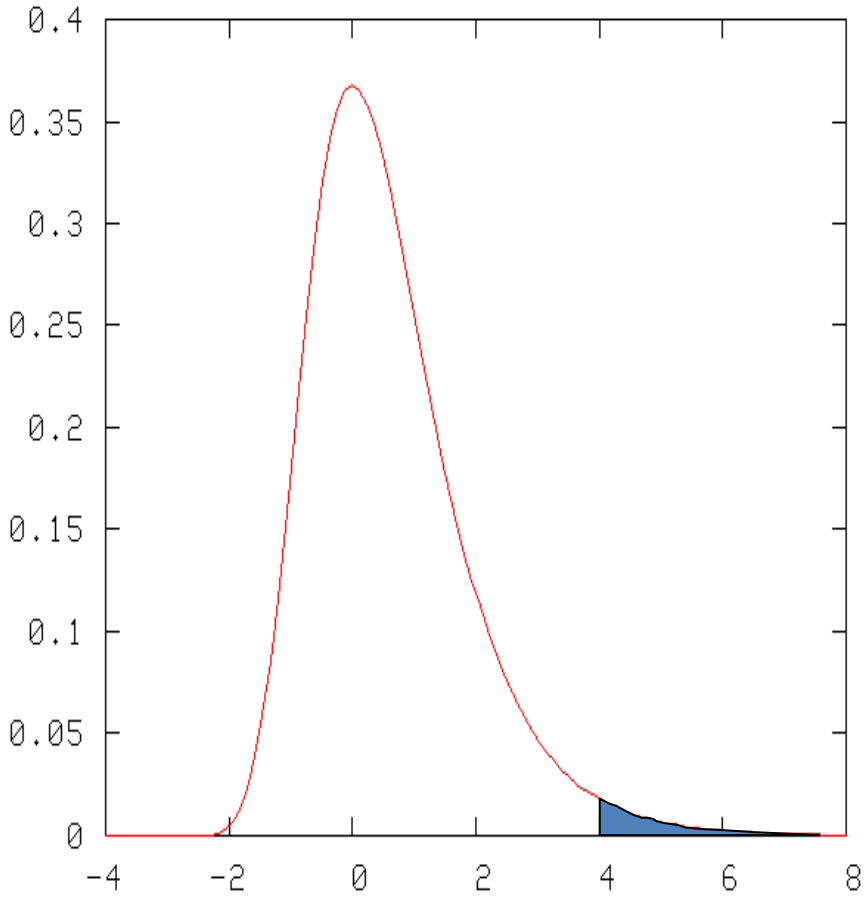
Converting scores to p-values

```
A  10 67 59 44
C  60 39 49 29
G   0 71 50 54
T 100 43 13 64
```



- In the end, the bottom row contains the scores for all possible sequences of length N.
- Use these scores to compute a p-value.

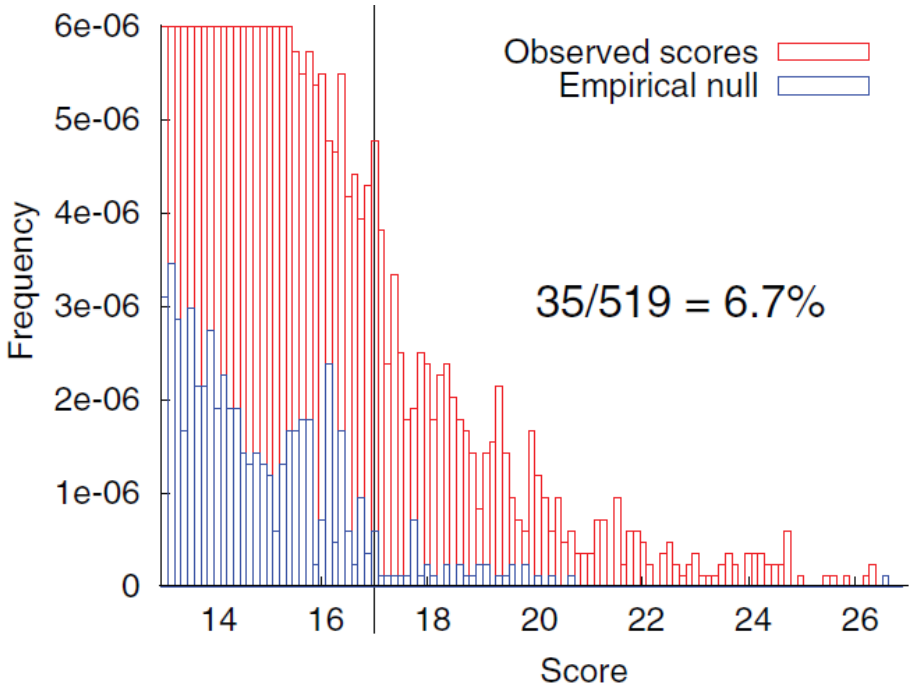
Computing a p-value



- The probability of observing a score >4 is the area under the curve to the right of 4.
- This probability is called a p-value.
- p-value = Pr(data|null)

Multiple testing correction

- Finally, apply standard methods to correct for the number of tests.



FIMO

Find Individual Motif Occurences

Version 4.1.0

Use this form to submit motifs to **FIMO** to be used in searching a sequence database.

Data Submission Form

Required

Your e-mail address:

Re-enter e-mail address:

Your motif file:

Browse...

Sample DNA minimal motif format file.

(Use your browser to save the sample to a file and then use "Browse" to upload it to test FIMO.)

Sequence database to search--select **one** of the following:

A supported database:

or

Your FASTA sequence file (1000000 sequence characters maximum):

Browse...

Sample DNA database.

Optional

Description of your motifs:

p-value output theshold

1e-4

Scan given strand only

Start search Clear Input

Outline

- Searching for occurrences of a given motif
- High-resolution models of transcription factor binding to DNA
- An embedding approach to remote protein homology detection

OPEN ACCESS Freely available online PLoS COMPUTATIONAL BIOLOGY

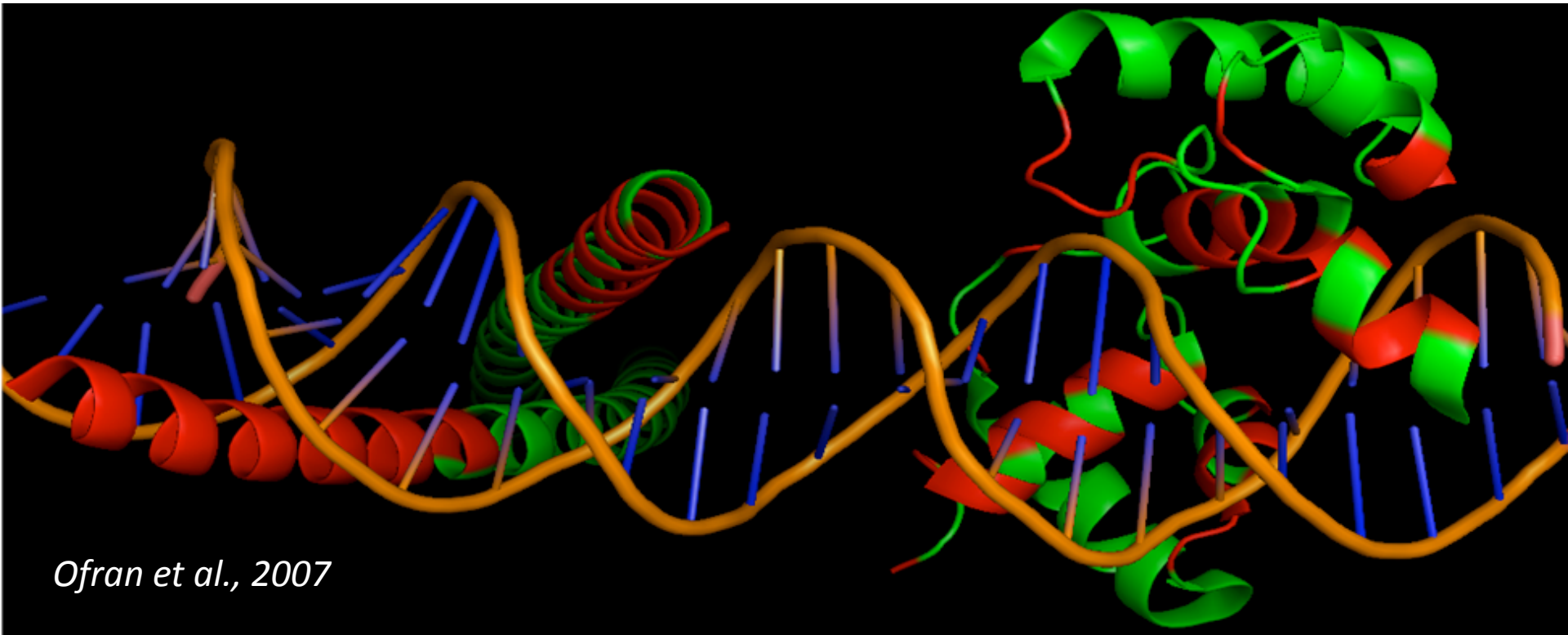
High Resolution Models of Transcription Factor-DNA Affinities Improve *In Vitro* and *In Vivo* Binding Predictions

Phaedra Agius[1], Aaron Arvey[1], William Chang[1], William Stafford Noble[2], Christina Leslie[1]*

1 Computational Biology Program, Memorial Sloan-Kettering Cancer Center, New York, New York, United States of America, **2** Department of Genome Sciences, University of Washington, Seattle, Washington, United States of America

Transcription factor binding sites

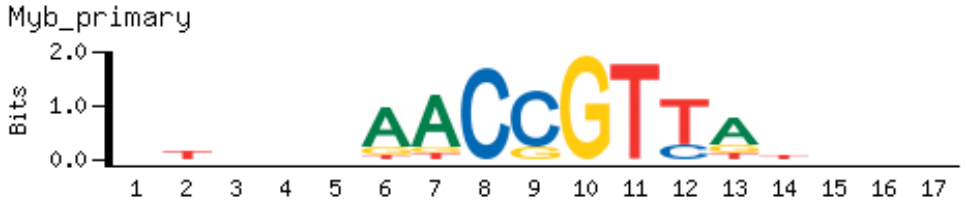
- Predicting genome-wide *occupancy* of TFs, key part of unraveling regulatory code
- > 500 sequence-specific human TFs, vast non-coding regions
- Need to model sequence preferences of TFs

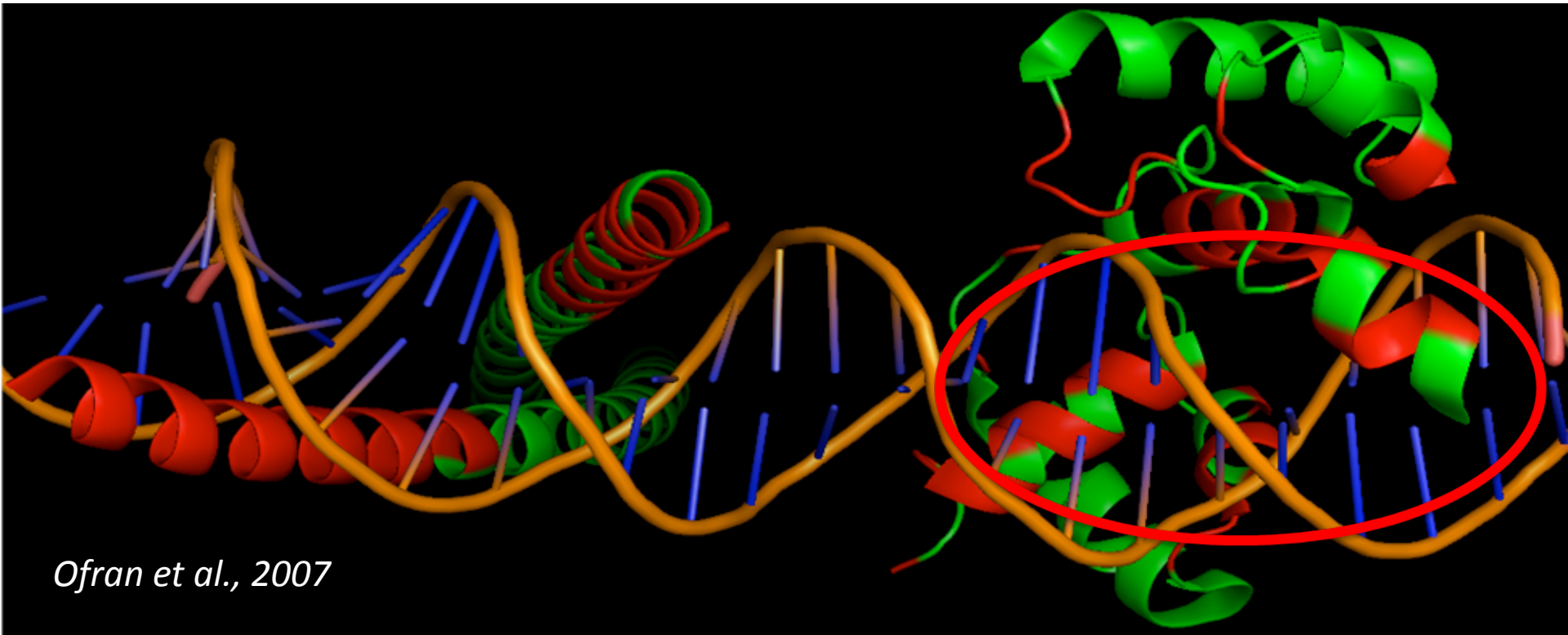


Ofran et al., 2007

Transcription factor binding sites

- Usually described as a *position specific scoring matrix (PSSM)* giving position-specific nucleotide frequencies
- Yields many false positives



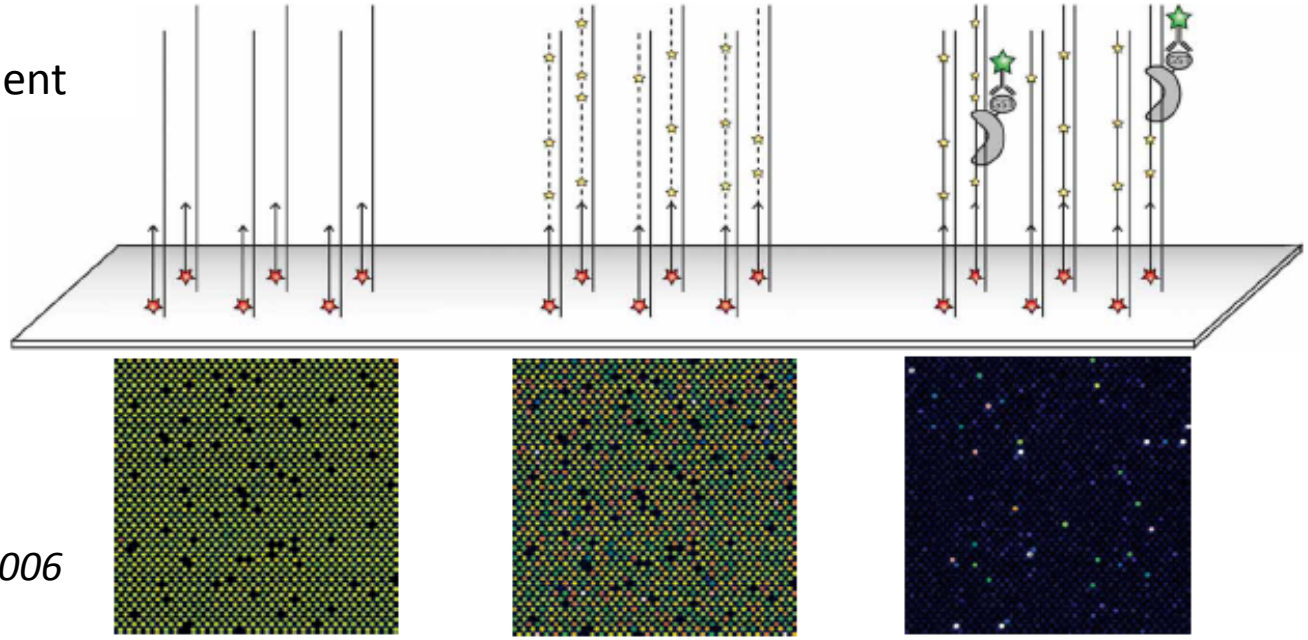


Ofran et al., 2007

Can we learn better TF binding preferences from high resolution data?

- Protein binding microarrays (PBMs): array with ~40K double-stranded probes (36-mers), designed to cover all 10-mers
- Measure *in vitro* DNA preferences of fluorescently tagged TF
- Good statistics on 8-mer patterns

PBM experiment

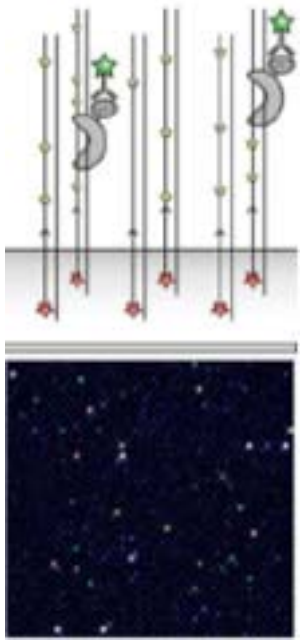


Berger et al., 2006

Standard analysis of PBM data

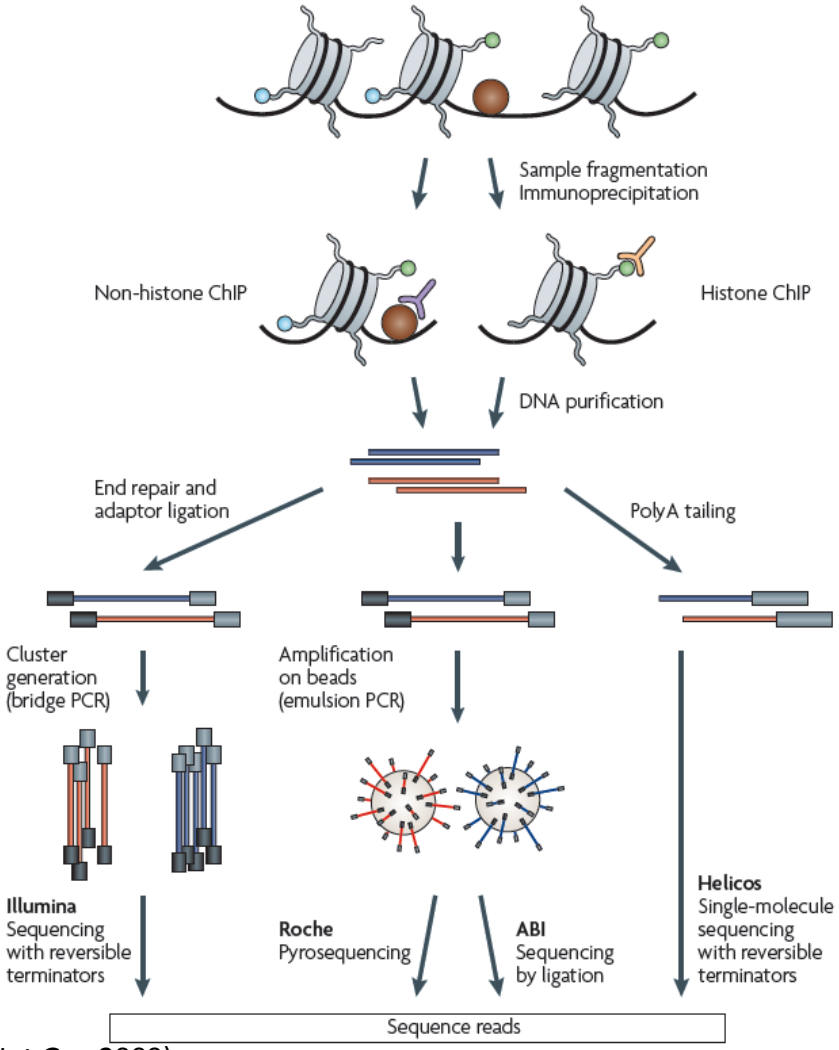
- Compute an E-score (enrichment score) for *every* 8-mer pattern, using rank statistic on probe intensities
- Obtain list of hundreds of 8-mer patterns with significant E-scores
- Too unwieldy, too much noise?

AACCGTTA, AACCGTCA, ACCGTTAT, AACGGTTA, AAC_GTTAT, ATCCGTTA, A_CCGTTAT, …

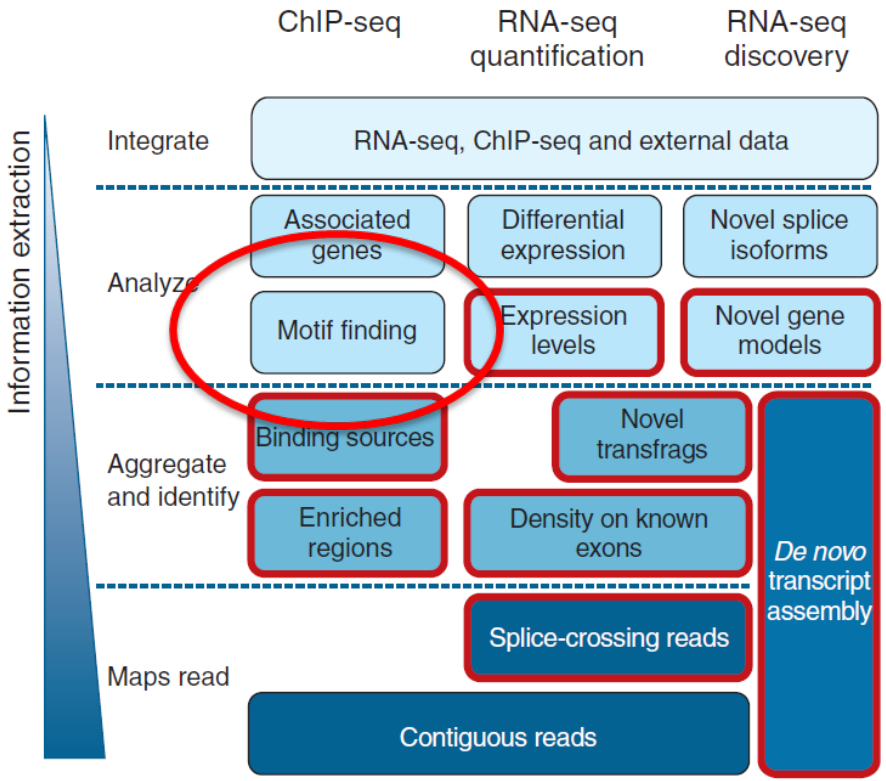


- Derive position weight matrix
- E.g., Seed-and-wobble (Bulyk lab): seed at top-scoring 8-mer, pull in additional patterns
- Too compact, underfitting?

Can we learn better *TF* binding preferences from high resolution data?

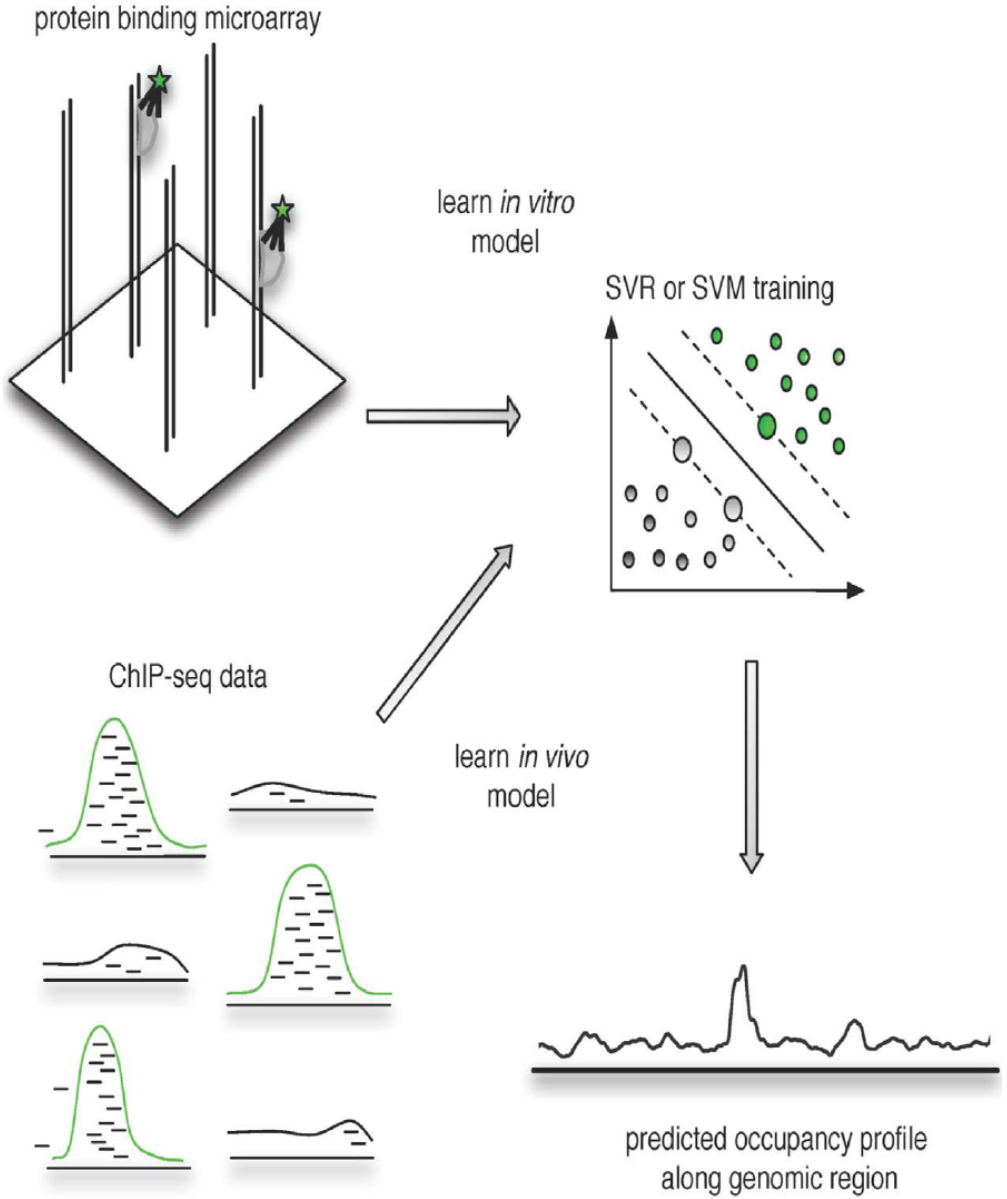


(Park *Nat Gen* 2009)



(Pepke et al. *Nat Meth* 2009)

Our approach: discriminative learning



Modeling Sequences With Kernels

Spectrum kernel (k=3)

CCACAGGCGGCTAGCTCGCTGCACGATATTATACAT
CCTTCGCTCGATAGTAGTTCTCGGCGGTATTATTTC
ATGGTTATCCGCTTTATTGCCGCCAGAATACTACTG
GTTACATCAACCAATAGCCGCTGGCAAGTTCTCACA
TAGCACATTGATATCCCAATTAGCCGCCTAGCACAG
GCTAGACTAGGGGACATCCGGCGGCTACTTCCAAAT
CGTTCTGCCCGTTGATCACATGCCGCTTATAAAACT
ACTGCGTTCGTTTAGGTTTTTGTGCGCGCTTAACCT

TTA	GCT	CCA	TTT	ATG	...
1	2	1	0	0	...
1	1	0	1	0	...
0	1	1	0	1	...
0	1	1	0	0	...
1	0	1	0	0	...
0	2	1	0	0	...
1	1	0	0	0	...
1	1	0	4	0	...

Mismatch Kernel (k=5, m=1)

CCACAGGCGGCTAGCTCGCTGCACGATATTATACAT
CCTTCGCTCGATAGTAGTTCTCGGCGGTATTATTTC
ATGGTTATCCGCTTTATTGCCGCCAGAATACTACTG
GTTACATCAACCAATAGCCGCTGGCAAGTTCTCACA
TAGCACATTGATATCCCAATTAGCCGCCTAGCACAG
GCTAGACTAGGGGACATCCGGCGGCTACTTCCAAAT
CGTTCTGCCCGTTGATCACATGCCGCTTATAAAACT
ACTGCGTTCGTTTAGGTTTTTGTGCGCGCTTAACCT

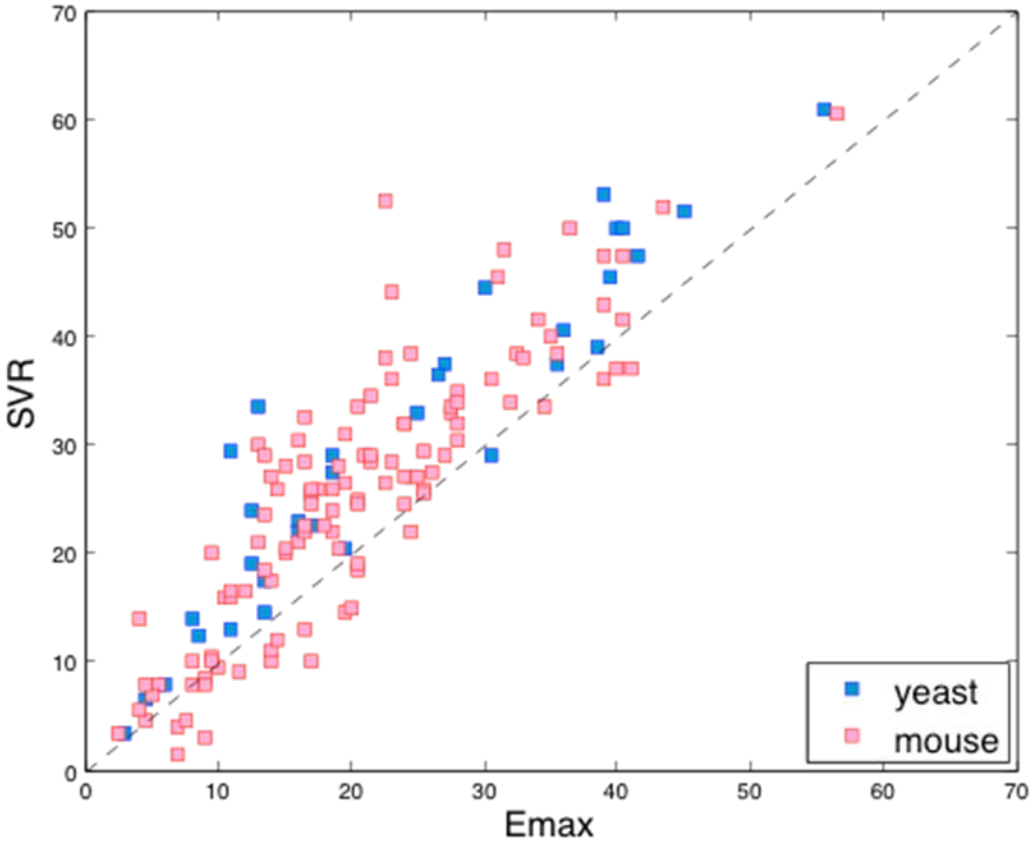
GCTTA	TATTA	TCACA	...
0	1	0	...
0	1	0	...
0	1	0	...
0	0	1	...
1	1	0	...
0	0	0	...
1	0	0	...
1	0	0	...

Kernel versus PSSM

- A kernel can
  - model dependencies between positions,
  - model background accessibility sequence signal, and
  - capture multiple and mutually disjoint cofactors of a given transcription factor.

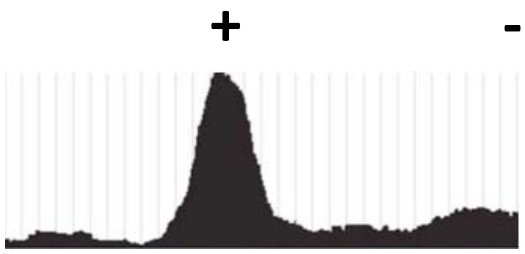
Better test accuracy on *in vitro* data

- Two PBM probe designs, train on one, test on other
- 3 large yeast and mouse data sets
- Detection of top 100 probes: improves over E-max (max E-score in test probe) in > 80% of experiments
- Similar performance improvement over PBM-derived PWMs

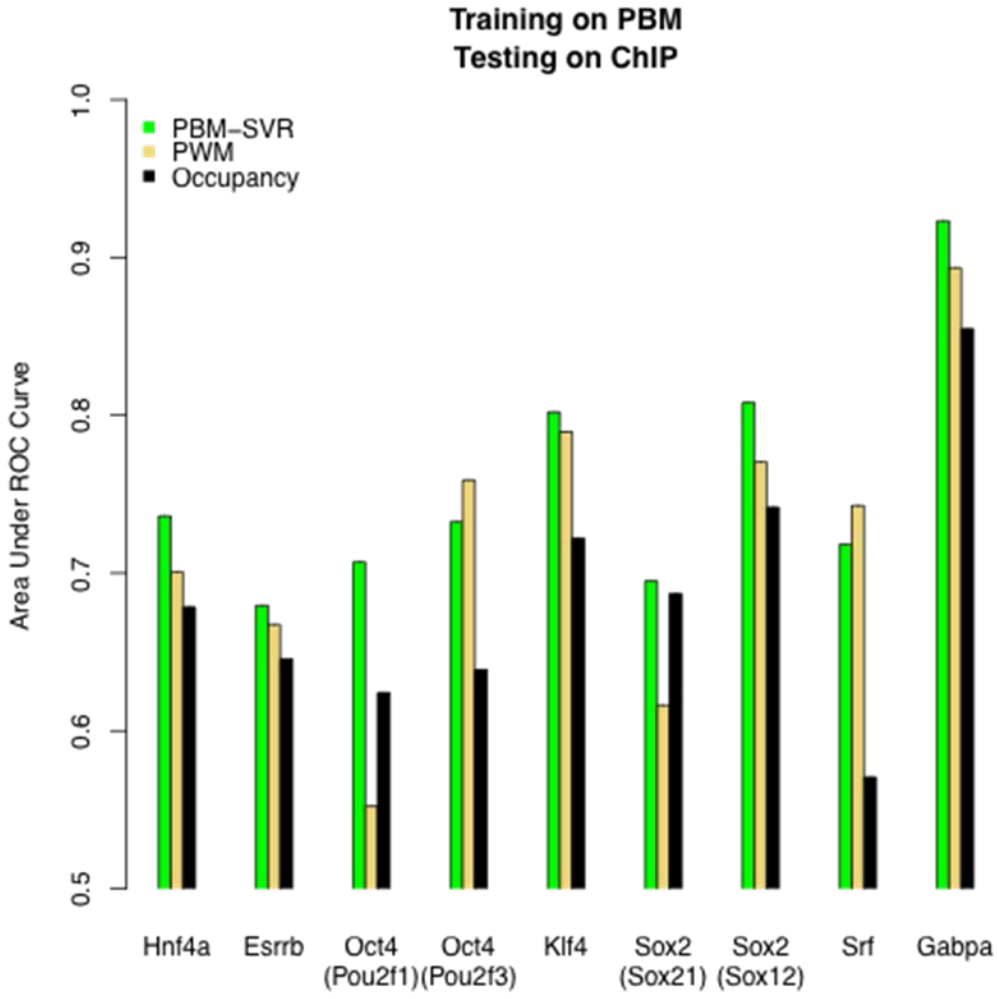


PBM-derived SVRs improve prediction of ChIP-seq peaks

- Use AUC to assess detection of mouse/ human ChIP-seq peaks (+'s) or nearby non-peak regions (-'s)

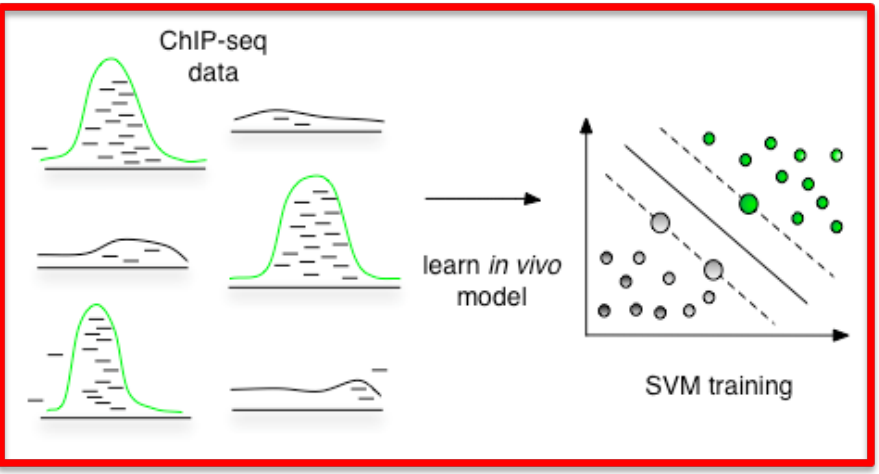


- Take PBM for "nearest neighbor" TF if needed
- Good improvement on high-resolution data

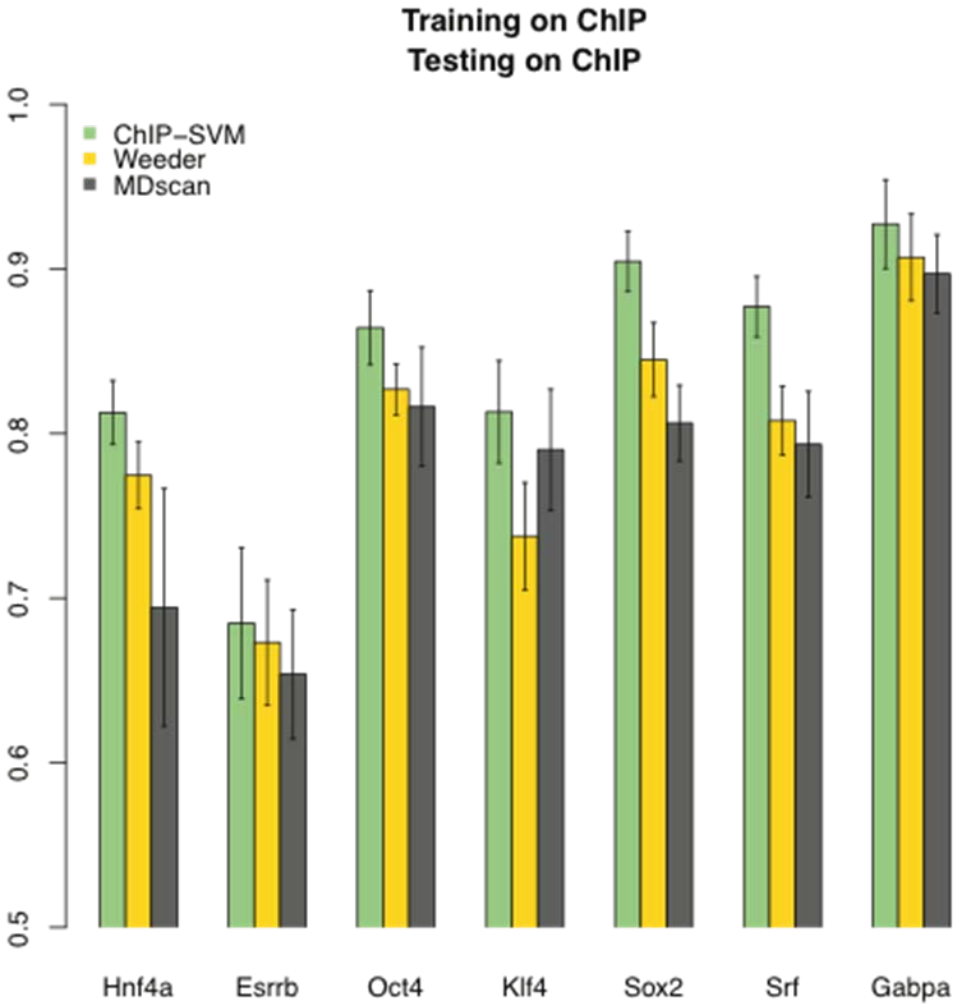


Training SVRs directly on genomic occupancy data increases accuracy

- Same kernel, train SVMs on peaks vs. flanking non-peaks

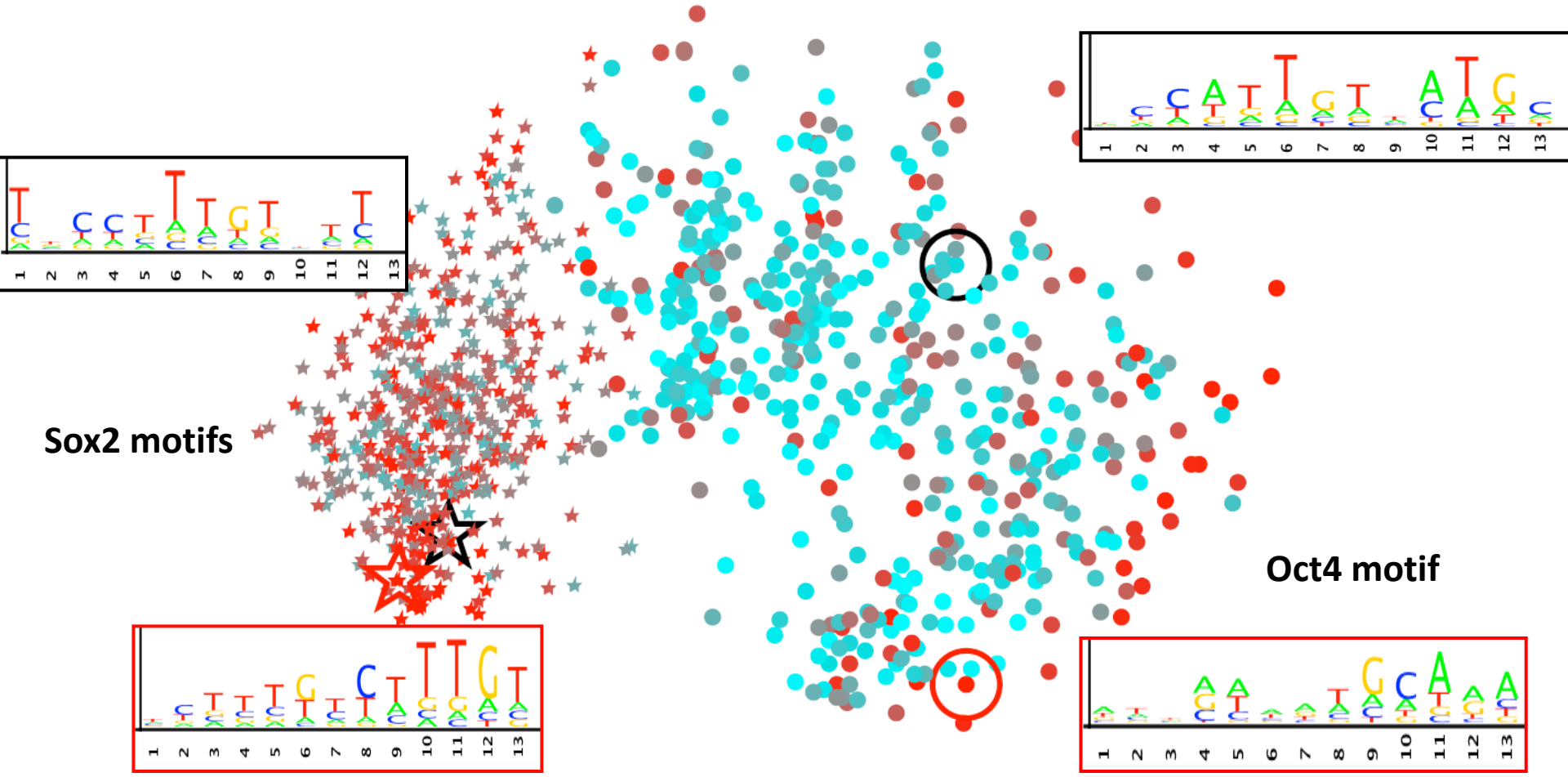


- Comparison to motif-discovery methods: Weeder (*k*-mers), MDscan (PWMs)



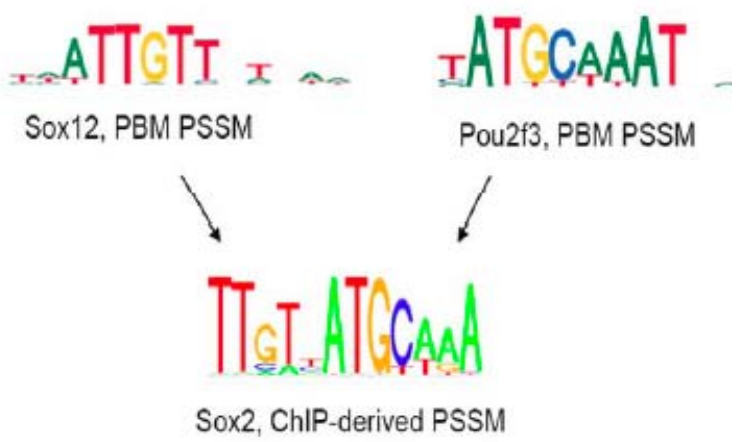
Inside the box: ChIP-derived SVRs find cofactor motifs

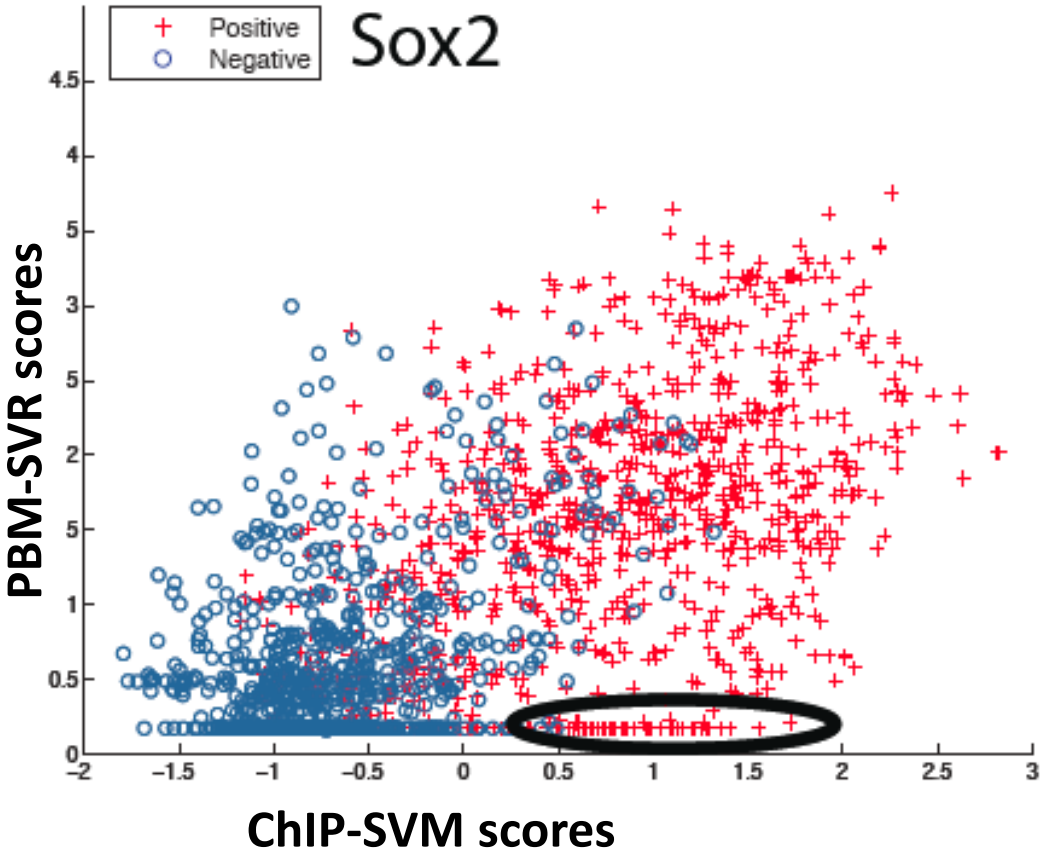
- ChIP-derived model of Sox2 (mouse ES cells) finds Oct4 motif



PBM-derived vs. ChIP-derived binding models

- 33 strong peaks missed by Sox2 PBM model but detected by ChIP model
- Oct4 PBM model detects 32/33 peaks
- ChIP model learning cofactor motifs





Outline

- Searching for occurrences of a given motif
- High-resolution models of transcription factor binding to DNA
- An embedding approach to remote protein homology detection

OPEN ACCESS Freely available online PLoS COMPUTATIONAL BIOLOGY

Detecting Remote Evolutionary Relationships among Proteins by Large-Scale Semantic Embedding

Iain Melvin[1], Jason Weston[2], William Stafford Noble[3]*, Christina Leslie[4]*

1 NEC Laboratories America, Princeton, New Jersey, United States of America, **2** Google, New York, New York, United States of America, **3** Department of Genome Sciences, University of Washington, Seattle, Washington, United States of America, **4** Computational Biology Program, Memorial Sloan-Kettering Cancer Center, New York, New York, United States of America

Iain Melvin

Jason Weston

Christina Leslie

NCBI

protein–protein **BLAST**

Nucleotide | Protein | Translations | Retrieve results for an RID

Search

Set subsequence From: To:

Choose database nr

Do CD-Search

Now: BLAST! or Reset query Reset all

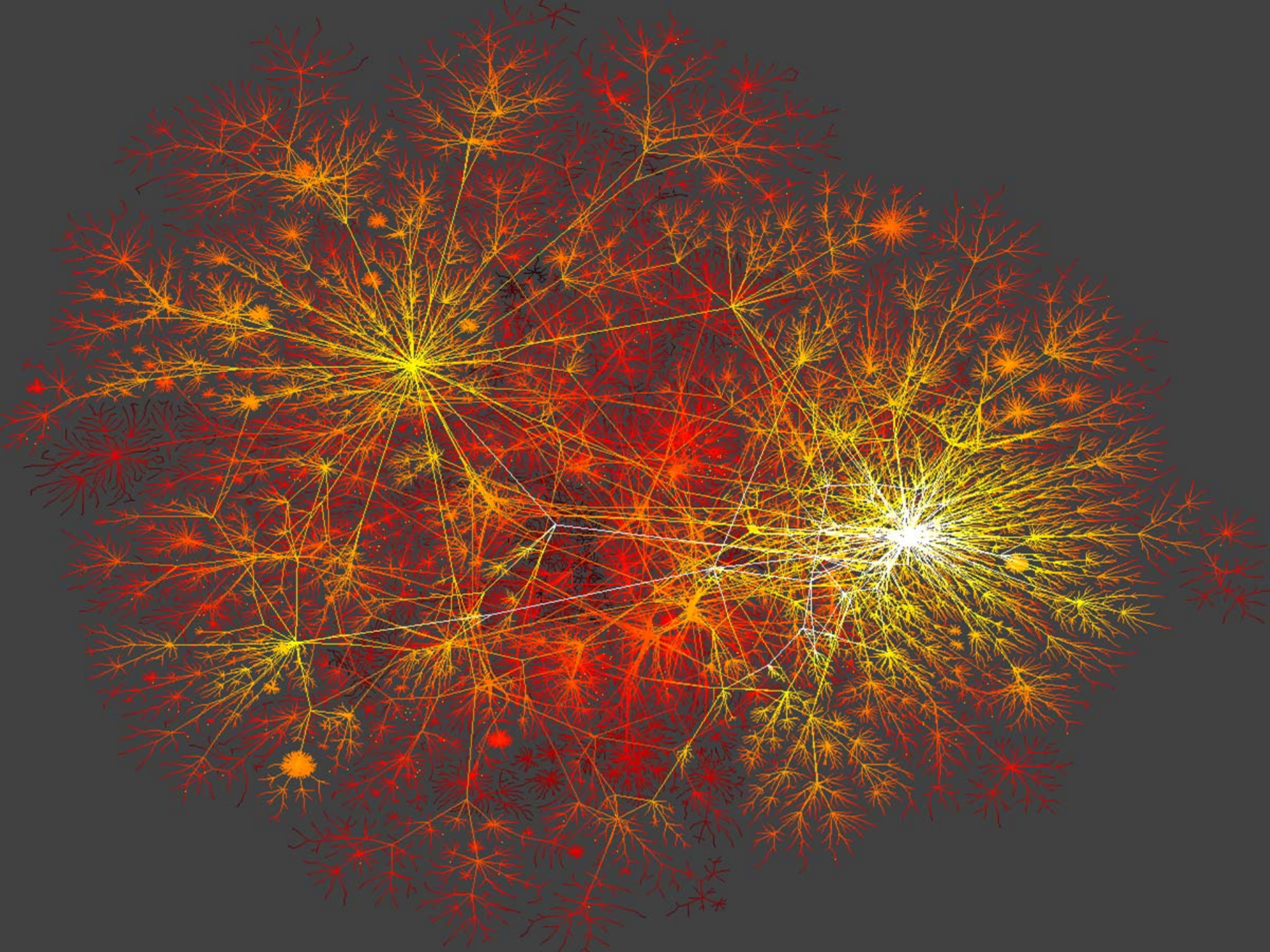
Napredno iskanje

Nastavitve

Jezikovna orodja

Iskanje Google | Klik na srečo

Išči po: celotnem spletu straneh v državi Slovenija



History

- Smith-Waterman (1981)
  - Optimal pairwise local alignment via dynamic programming
- BLAST (1990)
  - Heuristic approximation of Smith-Waterman
- PSI-BLAST (1997)
  - Iterative local search using profiles
- Rankprop (2004)
  - Diffusion over a network of protein similarities
- HHSearch (2005)
  - Pairwise alignment of profile hidden Markov models

Supervised semantic indexing

- Data: 1.8 million Wikipedia documents
- Goal: given a query, rank linked documents above unlinked documents
- Training labels: linked versus unlinked pairs
- Method: ranking SVM (essentially)
  - Margin ranking loss function
  - Low rank embedding
  - Highly scalable optimizer

(Bai et al., ECIR 2009)

Key idea

- Learn an embedding of proteins into a low-dimensional space such that homologous proteins are close to one another.
- Retrieve homologs of a query protein by retrieving nearby proteins in the learned space.

This method requires
- A feature representation
- A training signal
- An algorithm to learn the embedding

Protein similarity network

- Compute all-vs-all PSI-BLAST similarity network.
- Store all E-values (no threshold).
- Convert E-values to weights via transfer function (weight = $e^{-E/\sigma}$).
- Normalize edges leading into a node to sum to 1.

Sparse feature representation

Query protein

Target protein

$$\Phi(p') = (E(p', p_1), \ldots E(p', p_\ell))$$

PSI-BLAST / HHSearch
E-value for query j, target i

Hyperparameter

$$W(p', p_i) = exp(-S_j(i)/\sigma)$$

Probability that a random walk on the protein similarity network moves from protein p' to p_i.

$$E(p', p_i) = W_{p'p_i} / \sum_j W_{p'p_j}$$

Training signal

- Use PSI-BLAST or HHSearch as the teacher.
- Training examples consist of protein pairs.
- A pair *(q,p)* is positive if and only if query *q* retrieves target *p* with E-value < 0.01.
- The online training procedure randomly samples from all possible pairs.

Learning an embedding

- Goal: learn an embedding

$$g(p) = W\Phi(p)$$

where *W* is an *n*-by- rℓatrix, resulting in an *n*-dimensional embedding.

- Rank the database with respect to *q* using

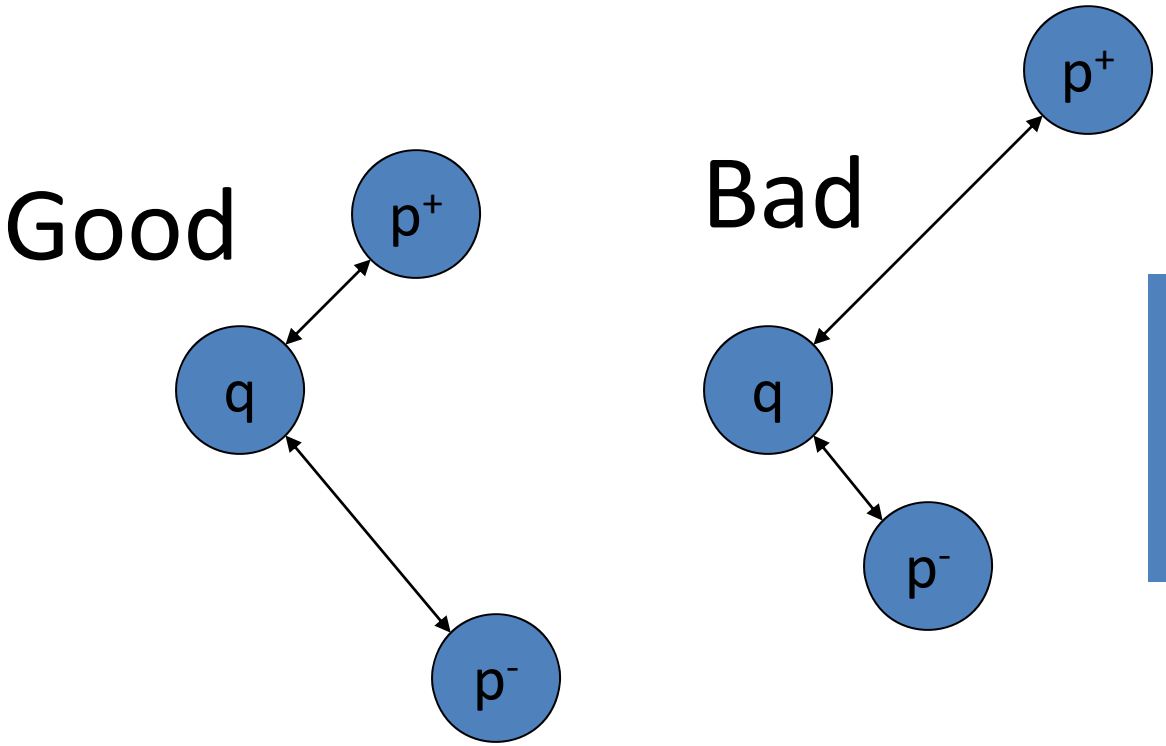
$$f(q, p_i) = ||g(q) - g(p_i)||_1 = ||W\Phi(q) - W\Phi(p_i)||_1$$

where small values are more highly ranked.

- Choose *W* such that for any tuple

$$f(q, p^+) < f(q, p^-)$$

Learning an embedding



Negative examples should be further from the query than positive examples by a margin of at least 1.

- Minimize the margin ranking loss with respect to tuples (q, p^+, p^-):

$$\sum_{(q,p^+,p^-)\in\mathcal{R}} \max(0, 1 - f(q,p^-) + f(q,p^+))$$

Training procedure

- Minimize the margin ranking loss with respect to tuples (q, p^+, p^-):

$$\sum_{(q,p^+,p^-)\in\mathcal{R}} \max(0, 1 - f(q,p^-) + f(q,p^+))$$

- Update rules:

$$\text{if } \; 1 - f(q,p^-) + f(q,p^+) > 0$$

$$W \leftarrow W - \lambda \, \text{sign}(W\Phi(q) - W\Phi(p^-))\Phi(q)^\top,$$
$$W \leftarrow W + \lambda \, \text{sign}(W\Phi(q) - W\Phi(p^-))\Phi(p^-)^\top,$$
$$W \leftarrow W + \lambda \, \text{sign}(W\Phi(q) - W\Phi(p^+))\Phi(q)^\top,$$
$$W \leftarrow W - \lambda \, \text{sign}(W\Phi(q) - W\Phi(p^+))\Phi(p^+)^\top,$$

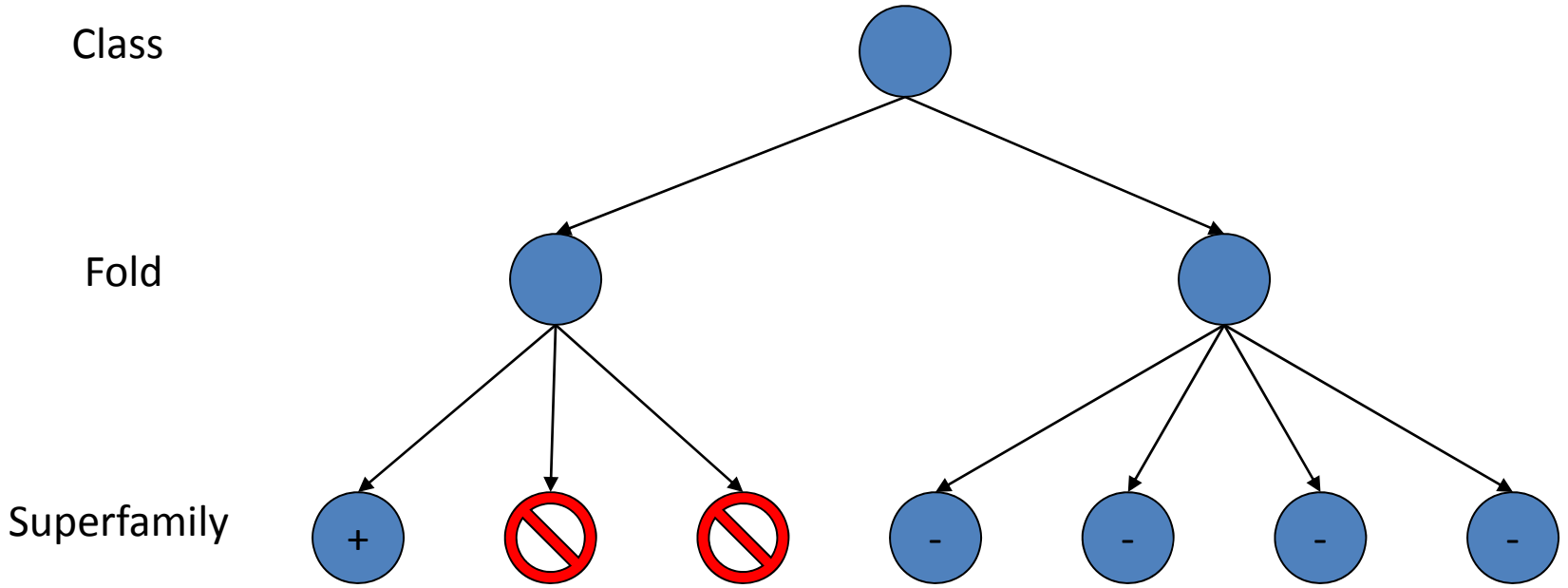
Push q away from p^-

Push p^- away from q

Push q toward p^+

Push p^+ toward q

Remote homology detection



- Semi-supervised setting: initial feature vectors are derived from a large set of unlabeled proteins.
- Performance metric: area under the ROC curve up to the 1st or 50th false positive, averaged over queries.

Results

Method	ROC_1	ROC_{50}
PSI-BLAST	0.624	0.632
Rankprop	0.647	0.707
Protembed PSI-BLAST	0.689	0.739
HHPred	0.771	0.836
Protembed HHPred	0.777	0.853

Results are averaged over 100 queries.

4-helical cytokines
Thioredoxin-like
EF-hand
Snake toxin-like
(Trans)glycosidases
Acid proteases
Scorpion toxin-like
Globin-like
Lysozyme-like
2HC zinc fingers
-binding domain

Key idea #2

- Protein structure is more informative for homology detection than sequence, but is only available for a subset of the data.

- Use *multi-task learning* to include structural information when it is available.

Structure-based labels

- Use the Structural Classification of Proteins to derive labels

$$y_i \in \{1, \ldots, C\}$$

- Introduce a centroid c_i for each SCOP category (fold, superfamily).
- Keep proteins in category *i* close to c_i:

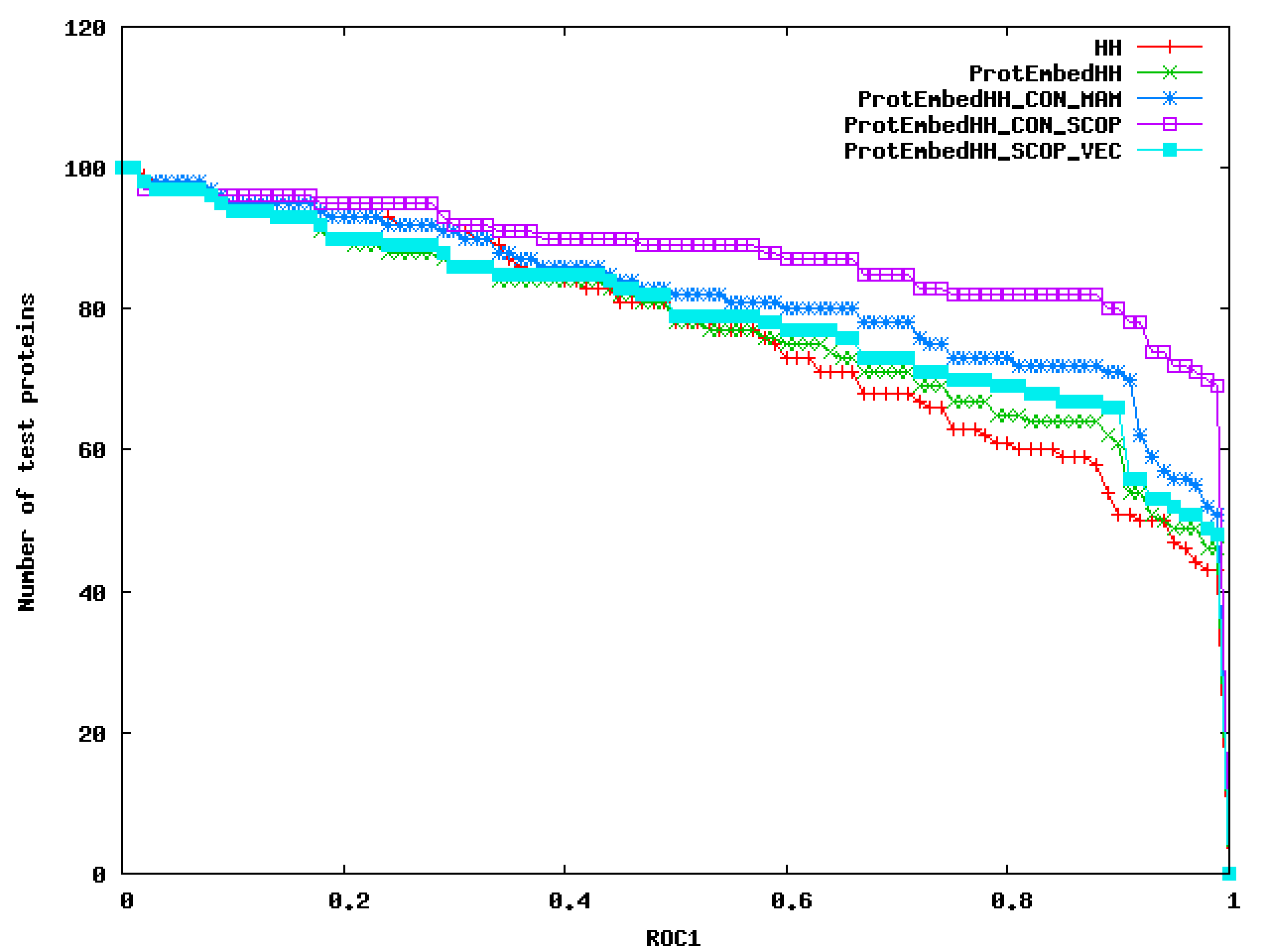
$$f(p_i, c_{y_i}) < f(p_j, c_{y_i}), \quad \forall j : y_j \neq y_i$$

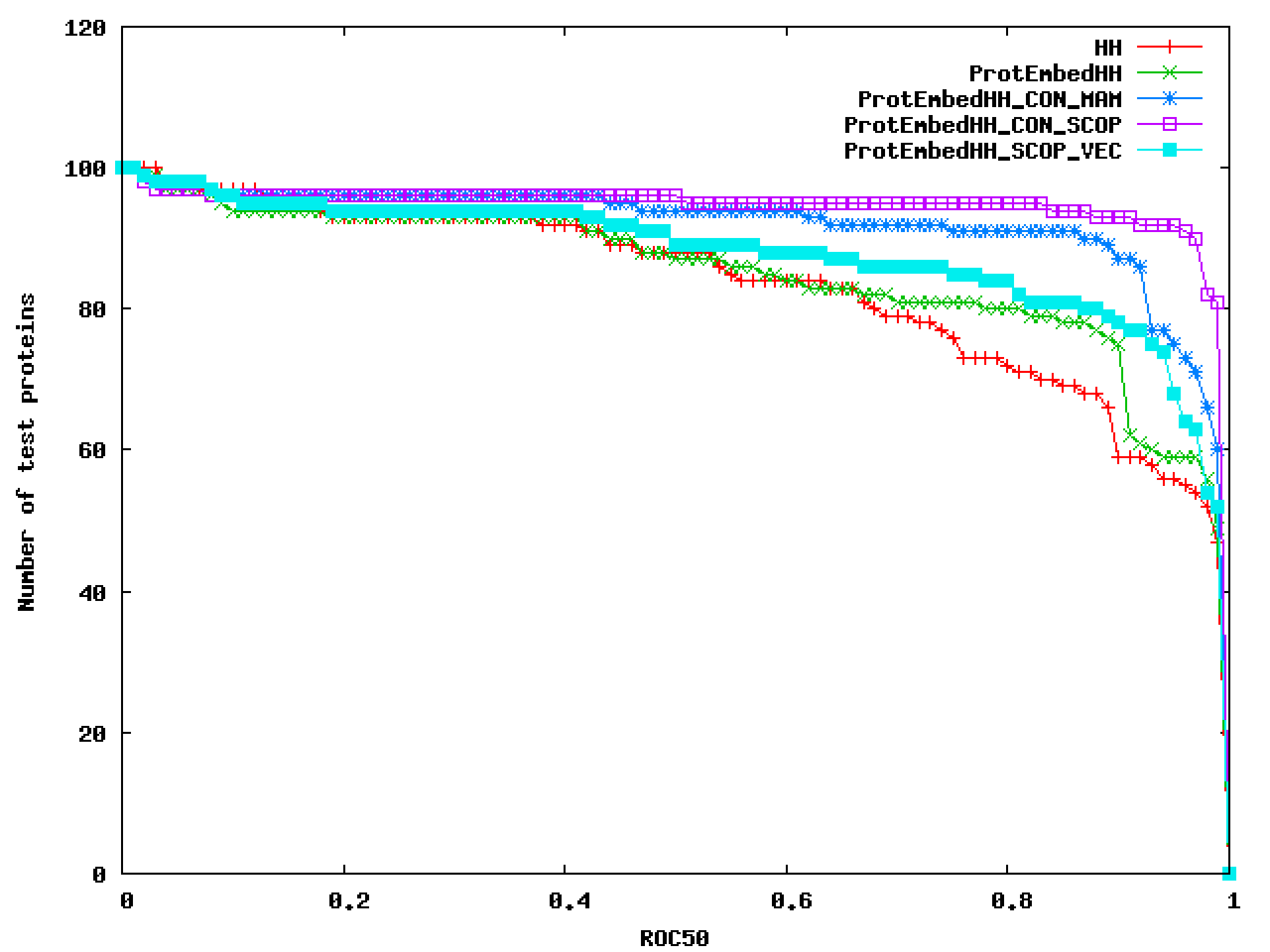
Structure-based ranks

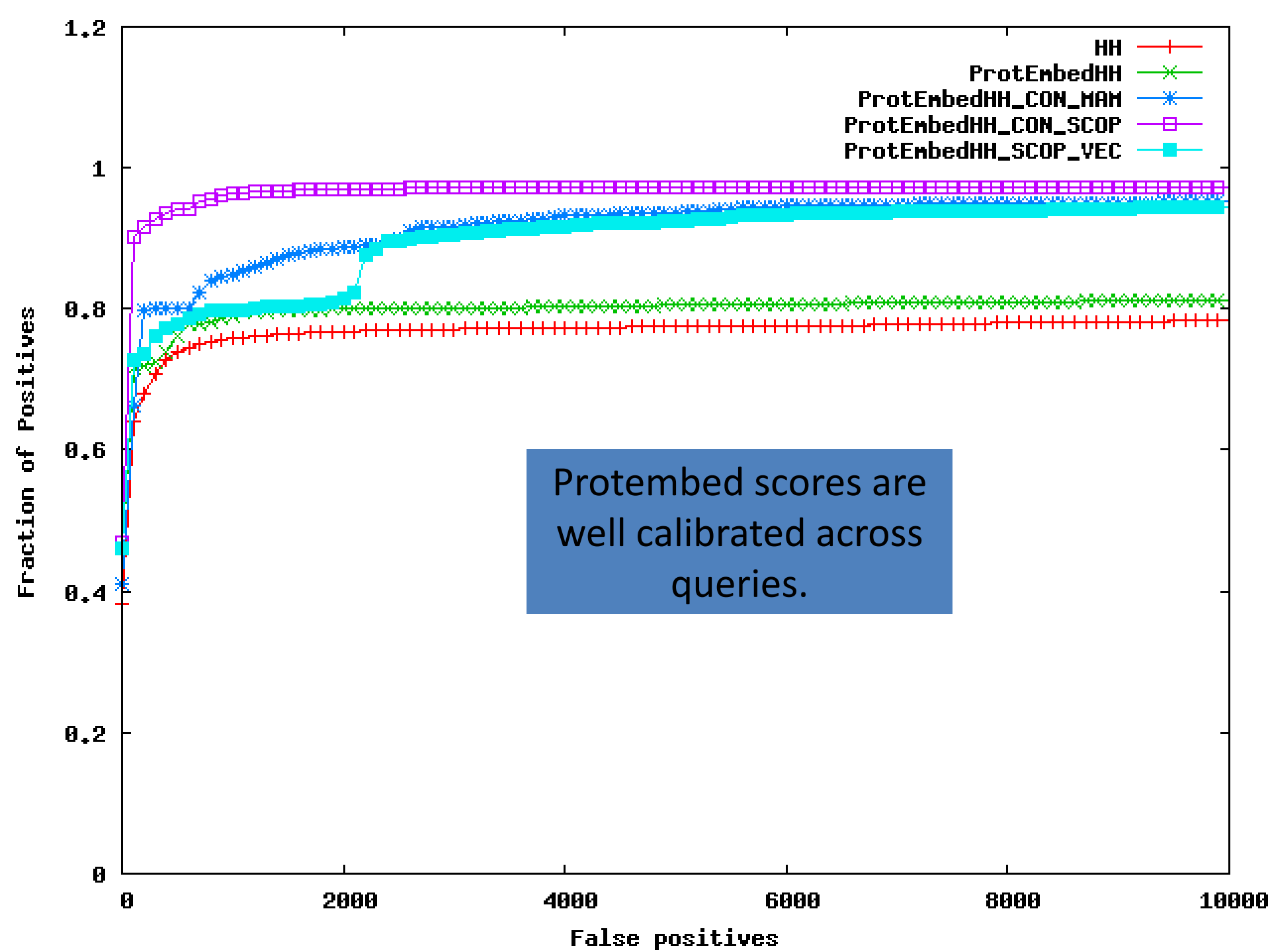
- Use a structure-based similarity algorithm (MAMMOTH) to introduce additional rank constraints.
- Divide proteins into positive and negative with respect to a query by thresholding on the MAMMOTH E-value.

$$f(q, p^+) < f(q, p^-)$$

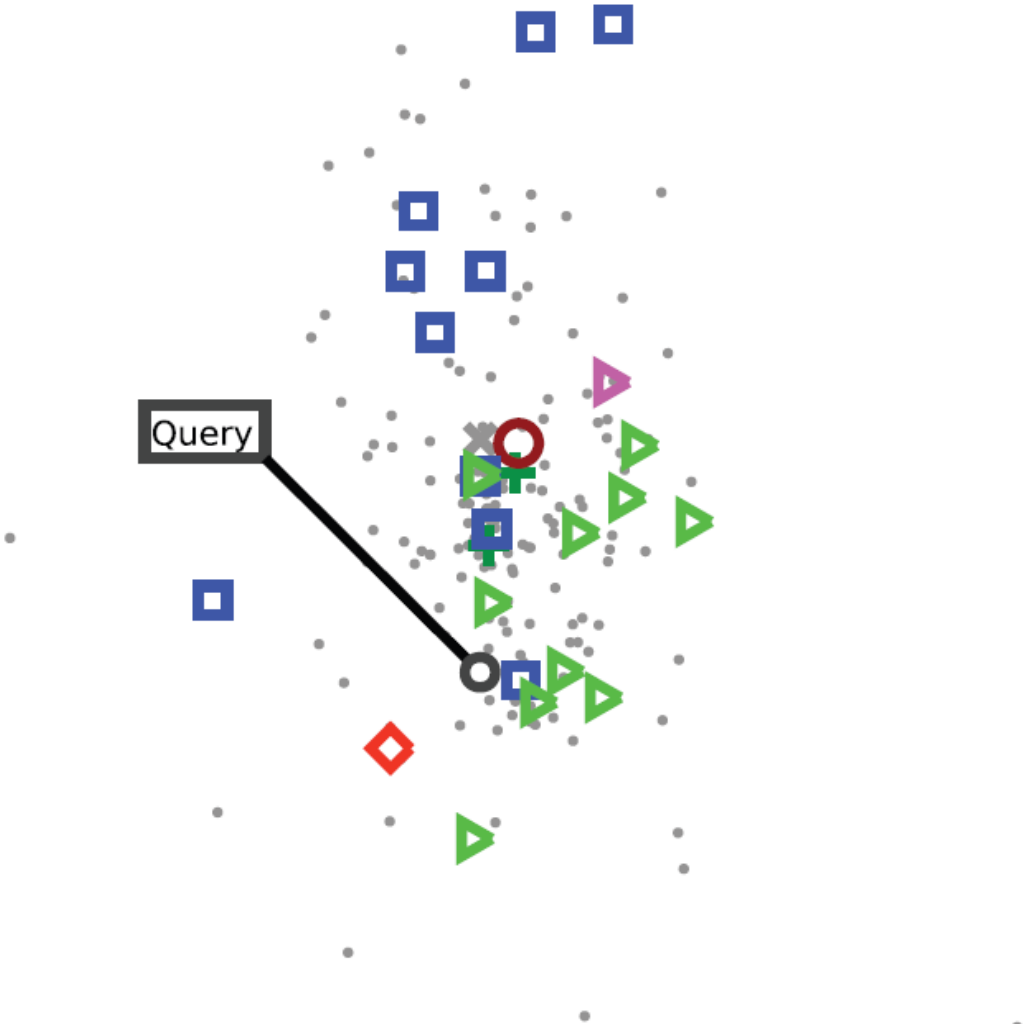
Method	ROC_1	ROC_{50}
PSI-BLAST	0.624	0.632
Rankprop	0.647	0.707
Protembed PSI-BLAST	0.689	0.739
Protembed PSI-BLAST+SCOP	0.852	0.918
Protembed PSI-BLAST+MAMMOTH	0.744	0.844
HHPred	0.771	0.836
Protembed HHPred	0.777	0.853
Protembed HHPred+MAMMOTH	0.822	0.923
Protembed HHPred+SCOP	0.881	0.949

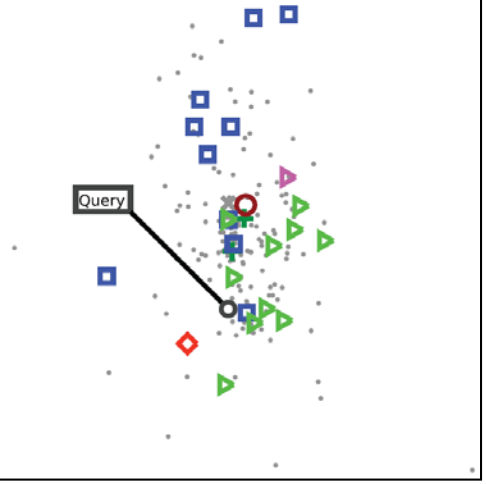

HH
ProtEmbedHH
ProtEmbedHH_CON_MAM
ProtEmbedHH_CON_SCOP
ProtEmbedHH_SCOP_VEC
Number of test proteins
ROC1
0
20
40
60
80
100
120
0
0.2
0.4
0.6
0.8
1


HH
ProtEmbedHH
ProtEmbedHH_CON_MAM
ProtEmbedHH_CON_SCOP
ProtEmbedHH_SCOP_VEC
Number of test proteins
ROC50
0
20
40
60
80
100
120
0
0.2
0.4
0.6
0.8
1


HH
ProtEmbedHH
ProtEmbedHH_CON_MAM
ProtEmbedHH_CON_SCOP
ProtEmbedHH_SCOP_VEC
Fraction of Positives
False positives
0
0.2
0.4
0.6
0.8
1
1.2
0
2000
4000
6000
8000
10000
Protembed scores are well calibrated across queries.

PSI-BLAST embedding





Learned embedding

Multiheme cytochromes

Rudiment single hybrid motif

Type II DNA topoisomerase

5' to 3' exonuclease, C-terminal subdomain

Resolvase-like

Query

Bacterial enterotoxins

Superantigen toxins, C-terminal domain

Conclusions

- Supervised semantic indexing projects proteins into a low-dimensional space where nearby proteins are homologs.
- The method bootstraps from unlabeled data and a training signal.
- The method can easily incorporate structural information as additional constraints, via multi-task learning.