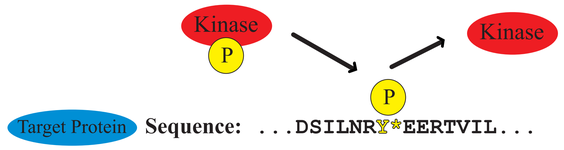
A Minimum Description Length Approach to the Multiple Motif Finding Problem

Anna Ritz[1], Gregory Shakhnarovich[1], and Benjamin J. Raphael[1,2]

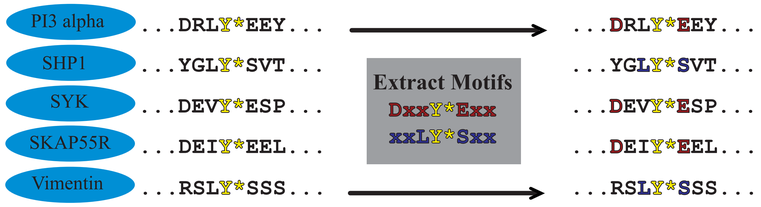
[1]Department of Computer Science, Brown University
[2]Center for Computational Molecular Biology, Brown University

1 Motivation

There are many interactions in the cell between proteins. Phosphorylation is an interaction where a kinase attaches a phosphate group to a specific target protein. Kinases identify their target proteins partly by the amino acid sequence around the interaction site. A motif is a pattern that approximates these binding preferences.



New technology produces datasets with a mixture of targets, presenting a new problem.



2 The Multiple Motif Finding Problem

Given: $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ foreground protein sequences of a fixed length l that are known to interact with a kinase; $Y = \{\mathbf{y}_1, \ldots, \mathbf{y}_m\}$ background protein sequences of length l over the entire proteome, where $n << m$.

Objective: A motif model $\mathcal{M} = \{M_1, \ldots M_k\}$ that best describes the underlying structure of X given Y, where $k \leq K$, the number of motifs allowed. Motifs can have wildcard positions ('x') and shared-letter positions (brackets). $\mathcal{M} = \{\}$ if the distributions determined by the BG describe the FG better than any set of patterns.

Problem: How can we evaluate motif models? We use description length (DL) as a metric, which computes the number of bits necessary to encode X.

3 Description Length

To compute DL, add the cost of encoding the model and the cost of encoding the data:

$$DL(X, \mathcal{M}) = DL_{\mathcal{M}} + DL(X|\mathcal{M}) = DL_{\mathcal{M}} + \sum_{i=1}^{n} DL(\mathbf{x}_i|\mathcal{M})$$

What is the best single motif for X = ABCD / ACCC / AADB ?

Axxx

Ax[CD]x

	1	2	3	4
A	1	1/3	0	0
B	0	1/3	0	1/3
C	0	1/3	2/3	1/3
D	0	0	1/3	1/3

Simplistic ⟷ **Complex**

Small $DL_{\mathcal{M}}$, Large $DL(X|\mathcal{M})$

Large $DL_{\mathcal{M}}$, Small $DL(X|\mathcal{M})$

4 Calculating Description Length

4.1 Intuition

Toy Example:
X =ABCD
$\mathcal{M}$ ={Ax[CD]x}
Alphabet Σ ={A,B,C,D}
Y distributions $p(A) = p(B) = p(C) = p(D) = 0.25$

Motif Representation:

$$\texttt{Ax[CD]x} \rightarrow \begin{cases} Z = \{1, 0, 1, 0\} \\ S = \{\{\underbrace{A}_{\mathbf{S}_1}\}, \emptyset, \{\underbrace{C, D}_{\mathbf{S}_3}\}, \emptyset\} \end{cases}$$

$DL_{\mathcal{M}}$

Item to Encode	Bits
# of motifs	1
encode Z	4
$\mathbf{S}_1$ (1 letter, A)	$2\lceil \log_2 4 \rceil$
$\mathbf{S}_3$ (2 letters, CD)	$3\lceil \log_2 4 \rceil$

$DL(X|\mathcal{M})$

Item to Encode	Bits
motif or background	1
encode col. 1	0
encode col. 2	$-\log p(B)$
encode col. 3	1
encode col. 4	$-\log p(D)$

Sequences can have multiple motifs: $Q_{im} = \begin{cases} 1 \text{ if } \mathbf{x}_i \text{ contains motif } m \\ 0 \text{ otherwise} \end{cases}$

4.2 Formal Definition

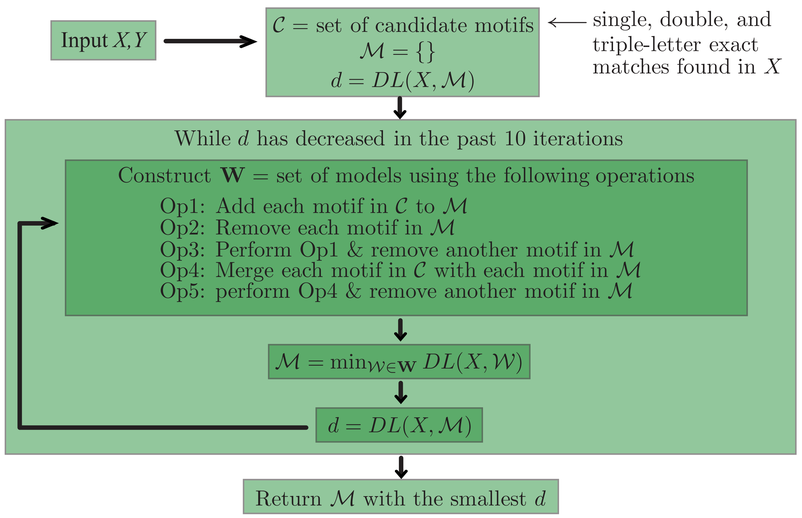
$$DL_{\mathcal{M}} = \lceil \log_2 K \rceil + \sum_{m=1}^{k} \left(L + \lceil \log_2 |\Sigma| \rceil \sum_{j:Z_{mj}=1} |\mathbf{S}_{mj}| \lceil \log_2 |\Sigma| \rceil \right)$$

$$DL(\mathbf{x}_i|\mathcal{M}) = k + \underbrace{\sum_{m=1}^{k} \left(Q_{im} \sum_{j=1}^{L} Z_{mj} \lceil \log_2 |\mathbf{S}_{mj}| \rceil \right)}_{\text{Explained by motif } m} - \underbrace{\sum_{j=1}^{L} (1 - Q_{im} Z_{mj}) \log p(x_{ij})}_{\text{Not explained by any motif}}$$

4.3 Generalizations

- Flexible Encoding Scheme for $\mathbf{S}_{mj}$:
  If $\lceil \log_2 |\mathbf{S}_{mj}| \rceil < |\Sigma|$, store $\mathbf{S}_{mj}$ as a $|\Sigma|$-bit binary vector.
- Flexible Encoding Scheme for Specifying Motifs for a Sequence:
  Background sequences are not penalized for adding a new motif.

5 Greedy Algorithm



6 Results

Other Algorithms:

Motif-X considers significant letters according to a binomially distributed model.

Semi-Exhaustive finds the best pair of motifs that have up to two active columns and share up to two letters per column.

Datasets:

Cao et. al.: Mouse cells enriched for tyrosine phosphorylation (Y-centered).

Wolf-Yadlin et. al.: Human cells enriched for tyrosine phosphorylation.

Olsen et. al.: Human cells *not* enriched, producing 3 datasets (S,T&Y).

Dataset	Greedy MDL Model	FG%	BG%	Score[2]	Motif-X[1] Model	FG%	BG%	Score[2]	Semi-Exhaustive Model
Cao et. al., Y	[DILV]Y*[ES]	28	4.0	52.51	DxxY*E	8	0.3	30.14	DxxY*
144 seq.	DxxY*	24	5.0	31.15	IY*	19	5.3	20.00	[IL]Y*[DE]
					ExxY*	19	6.8	13.35	
DL Reduction:	**144.88**				**56.88**				**131.06**
W-Y. et. al., Y	Y*[DESY]x[LP]	31	3.2	28.07	Y*xxP	25	5.4	14.14	Y*xxP
59 seq.									ExxxY*
DL Reduction:	**24.29**				**24.33**				**21.86**
Olsen et. al., S	S*[DEP]x[DER]	27	3.4	1,026.40	S*P	33	7.3	820.75	S*P
2,958 seq.	S*P	33	7.3	820.75	S*xxE	21	9.0	312.07	S*[DE]x[DE]
					SxxxS*	16	10.9	41.34	
					SxS*	20	10.9	24.46	
DL Reduction:	**3,937**				**1,427**				**3,843**
Olsen et. al., T	T*P	27	6.9	126.75	T*P	27	6.9	126.75	T*P
652 seq.	[DPRS][DPS]xT*	22	5.8	91.90	SxT*	22	8.9	55.14	[ES][EP]T*
					SxxT*	17	9.0	25.73	
					SxxxT*	17	9.2	23.27	
DL Reduction:	**614.41**				**203.61**				**580.40**
Olsen et. al., Y	[EFS]xx[DENP]Y*	27	4.0	45.79	Y*xxP	18	5.7	15.36	[ES]xxxY*xx[DE]
137 seq.	Y*xxP	18	5.7	15.36					Y*[ES]x[DP]
DL Reduction:	**92.36**				**54.90**				**73.76**

[1]The threshold parameter was 10^{-6} and the min. # of occurrences were 5, 5, 5, 200, and 50 respectively.
[2]The score is $-\log(p)$, where p is the p-value according to a binomially distributed model.

The Greedy MDL model saves the greatest number of bits (denoted DL Reduction) in all datasets except the W-Y et. al.; we argue that the descriptiveness of the Greedy MDL motif for this dataset compensates for the 0.04 bits that are not saved..

There is one motif in the Greedy MDL model that describes more letters for more sequences in the data.

The score that Motif-X tries to optimize is as at least as good in every Greedy MDL motif.

The motif models for the Semi-Exhaustive method are submotifs of the final model the the greedy MDL algorithm returns.

The difference between the smallest and largest datasets (in terms of number of sequences) are 2,899, showing that this method scales

7 Acknowledgements

We thank Arthur Salomon, professor of Biology in the Molecular Biology, Cell Biology, and Biochemistry Department at Brown University for his contributions, as well as Lulu Cao and Kebing Yu from his lab.

8 References

Lulu Cao, Cidy Banh, Vinh Nguyen, Anna Ritz, Benjamin J. Raphael, Yuko Kawakami, Toshiaki Kawakami, and Arthur R. Salomon. Quantitaitve time-resolved phosphoproteomic analysis of mast cell signaling. *Journal of Immunology*, 2007.

Jesper V Olsen, Blagoy Blagoev, Florian Gnad, Boris Macek, Chanchal Kumar, Peter Mortensen, and Matthias Mann. Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. *Cell*, 127(3):365-648, Nov. 2006.

Daniel Schwartz and Steven P Gygi. An iterative statistical approach to the identification of protein phosphorylation motifs from large-scale datasets. *Nat Biotechnol*, 23(11):1391-1398, 2005.

Alejandro Wolf-Yadlin, Neil Kumar, Yi Zhang, Sampsa Hautaniemi, Muhammad Zaman, Hyung-Do Kim, Viara Grantcharova, Douglas A Lauffenburger, and Forest M White. Effects of HER2 overexpression on cell signaling networks governing proliferation and migration. *Mol Syst Biol*, 2:54, 2006.