Comparison Between m-SVM and TargetLoc on Protein Subcellular Localization Problem

Fanchao Meng
fcmeng@cis.udel.edu

Computer and Information Sciences Department
University of Delaware

May 19th, 2014

Outline

- **Motivation**
- **Method Comparison**
- **Experimental Results**
- **Future Work**
- **Conclusion**

Motivation

- **What have we seen in TargetLoc** [1]**?**
  - Multi-layer prediction system
  - N-terminal targeting sequence
  - Overall amino acid composition
  - Protein specific motifs

- **Is there an alternative way?**
  - To deal with multiple kernels
  - To use amino acid composition and motifs better

Method Comparison

- **TargetLoc**

➢ General Framework [1]

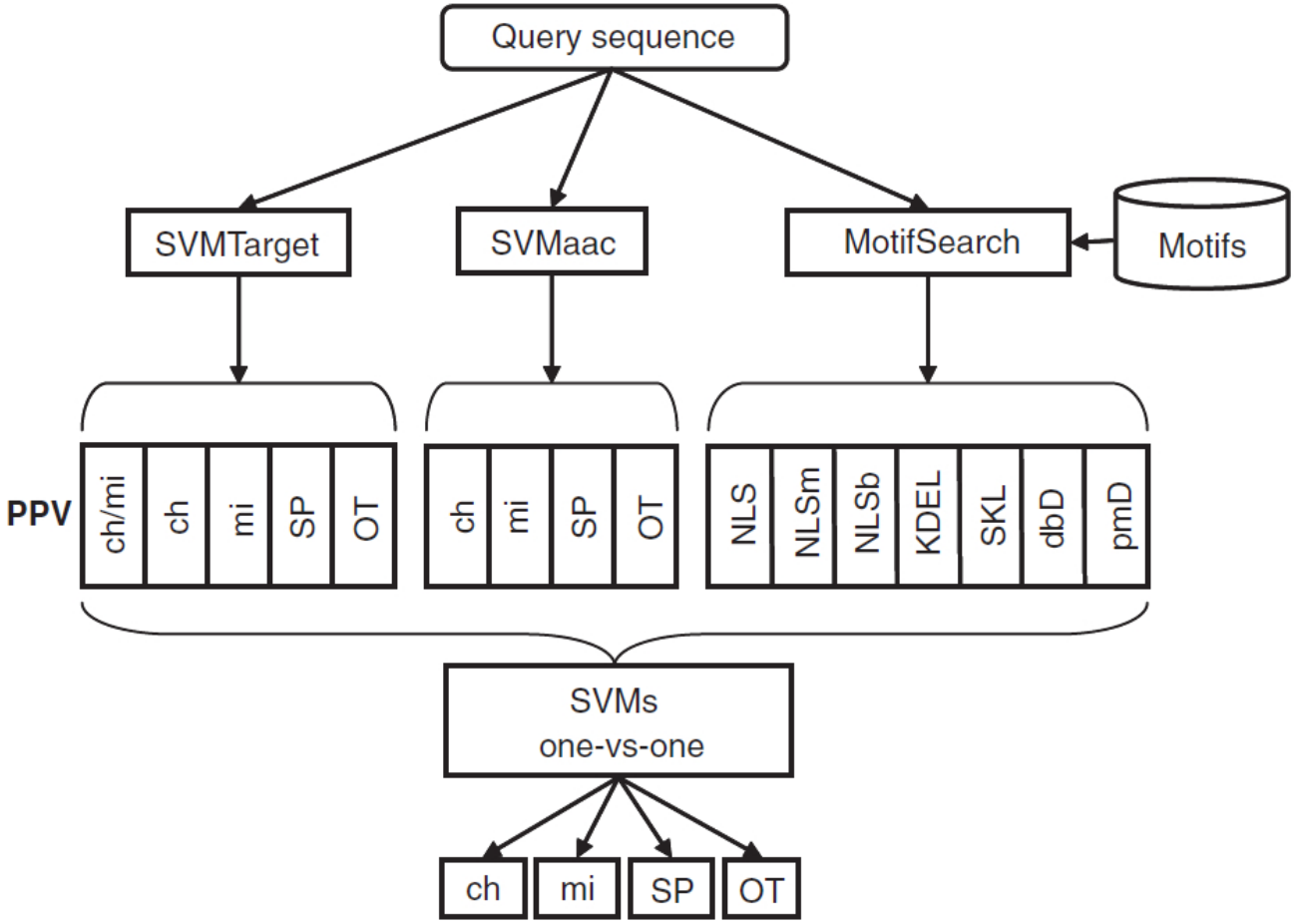


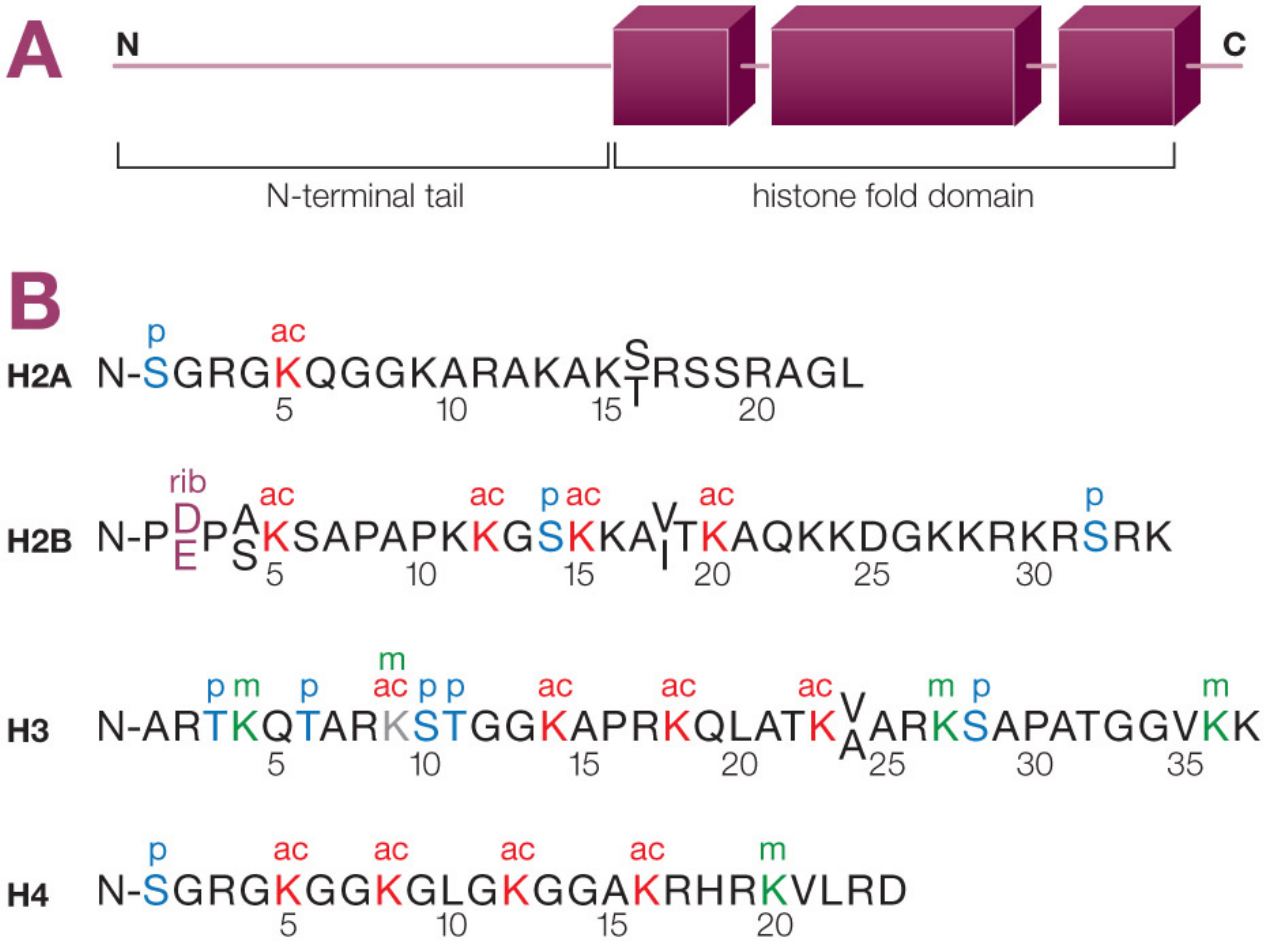
Fig. 1. The architecture of the TargetLoc prediction system. A query sequence enters a first layer of prediction methods; SVMTarget, SVMaac and MotifSearch. The information is collected in the protein profile vector (PPV). A set of one-versus-one SVMs are used by TargetLoc for the final classification according to the highest score using probability estimates.

Method Comparison (cont.)

- **TargetLoc**

➢ SVMTarget [1]

- Predicts localization categories based on N-terminal targeting sequences.
- **N-terminal targeting sequence**: *(In fact a kind of partial amino acid composition)*



A. Positioning of the histone tail relative to the C-terminal folded region.

B. B. Amino acid sequences of core histone N-terminal tails, indicating sites of phosphorylation (p), acetylation (ac), ADP ribosylation (rib), and methylation (m).

Method Comparison (cont.)

- **TargetLoc**

➢ SVMTarget [1]

o Architecture:

Binary SVM + Multi-layer

(In fact a Multi-class SVM)

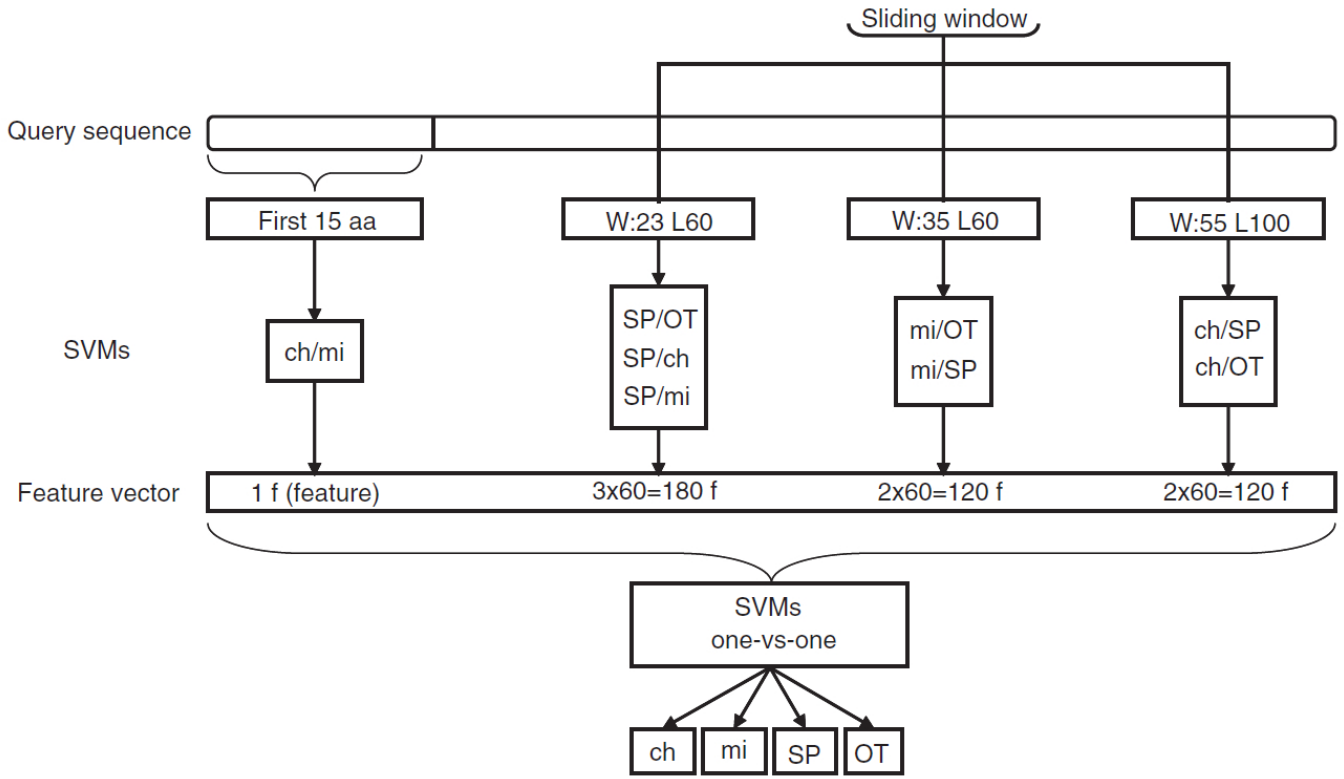


Fig. 3. The architecture of the SVMTarget plant version is illustrated here. Sliding windows of width *W* over the first *L* N-terminal amino acid residues create the partial amino acid composition vectors, which are used as input for the first layer of SVMs. There are eight binary SVMs in the plant version and four in the non-plant version (not shown). The input for the ch/mi classifier is the amino acid composition of the first 15 N-terminal residues. The input for the second layer of SVMs consists of the output scores (features) from the first, where a set of one-versus-one SVMs are used for the final classification using probability estimates.

Method Comparison (cont.)

- **TargetLoc**

➢ SVMaac [1]

- A set of classifiers for locations based on overall amino acid composition.
- Example: [4]

SVMs for:

1. Amino acid composition

2. Amino acid pair composition

3. Gapped amino acid composition (1-3 intervening residues)

(In fact can be generalized to Amino acid composition + Composition pattern)

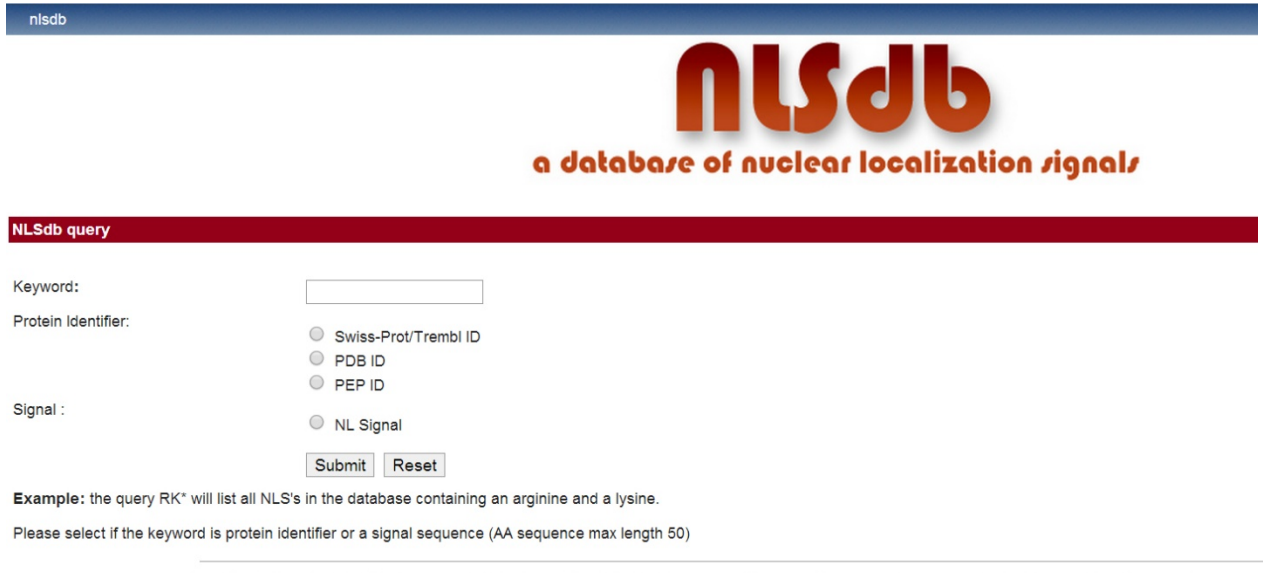
Voting Scheme:

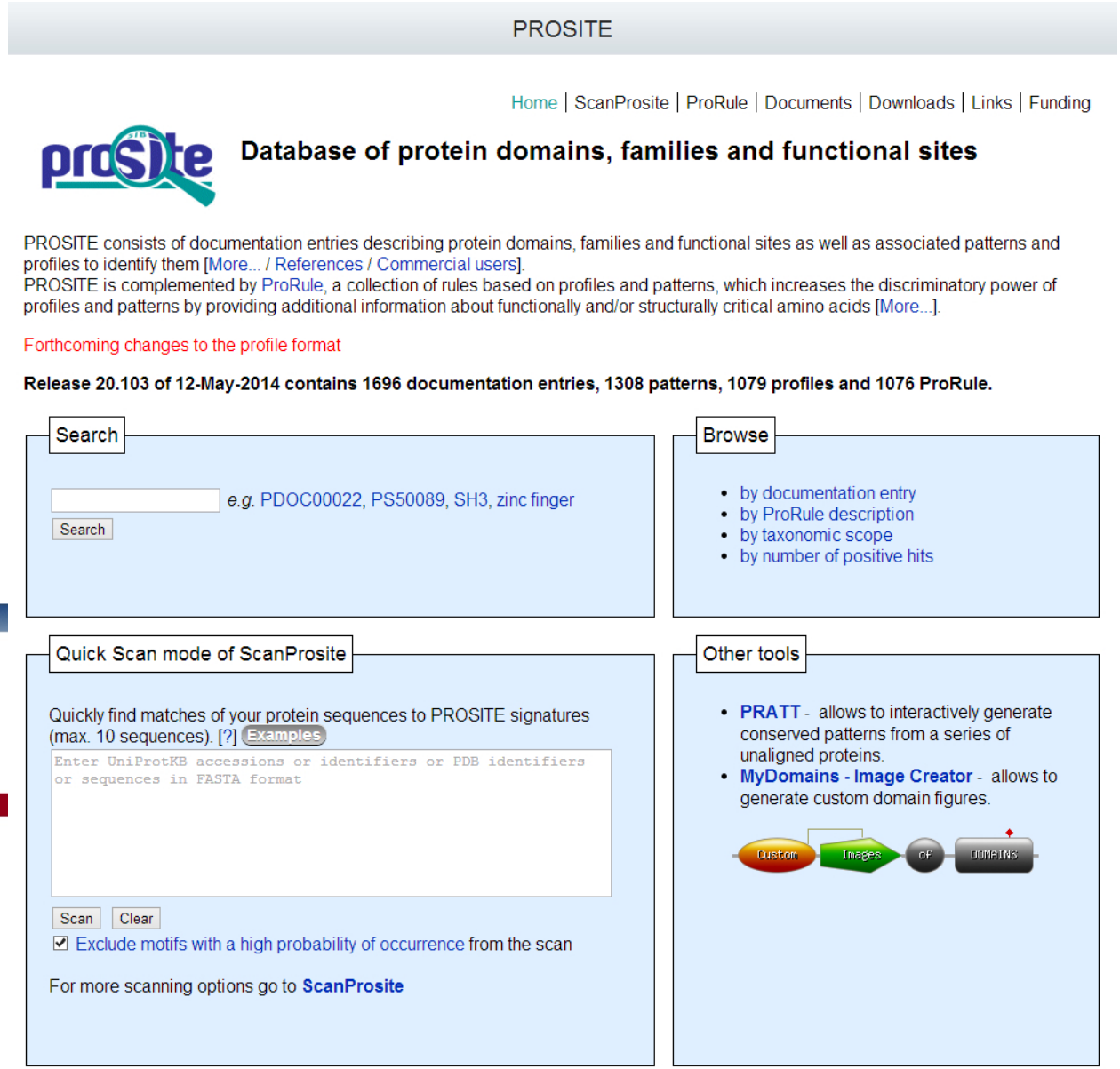
"1 vs Rest"

(In fact a Multi-class SVM)

Method Comparison (cont.)

- **TargetLoc**
  - MotifSearch [1]
    - Homology info based on **motifs**
      (In fact motif is a composition of amino acids)
    - PROSITE & NLSdb



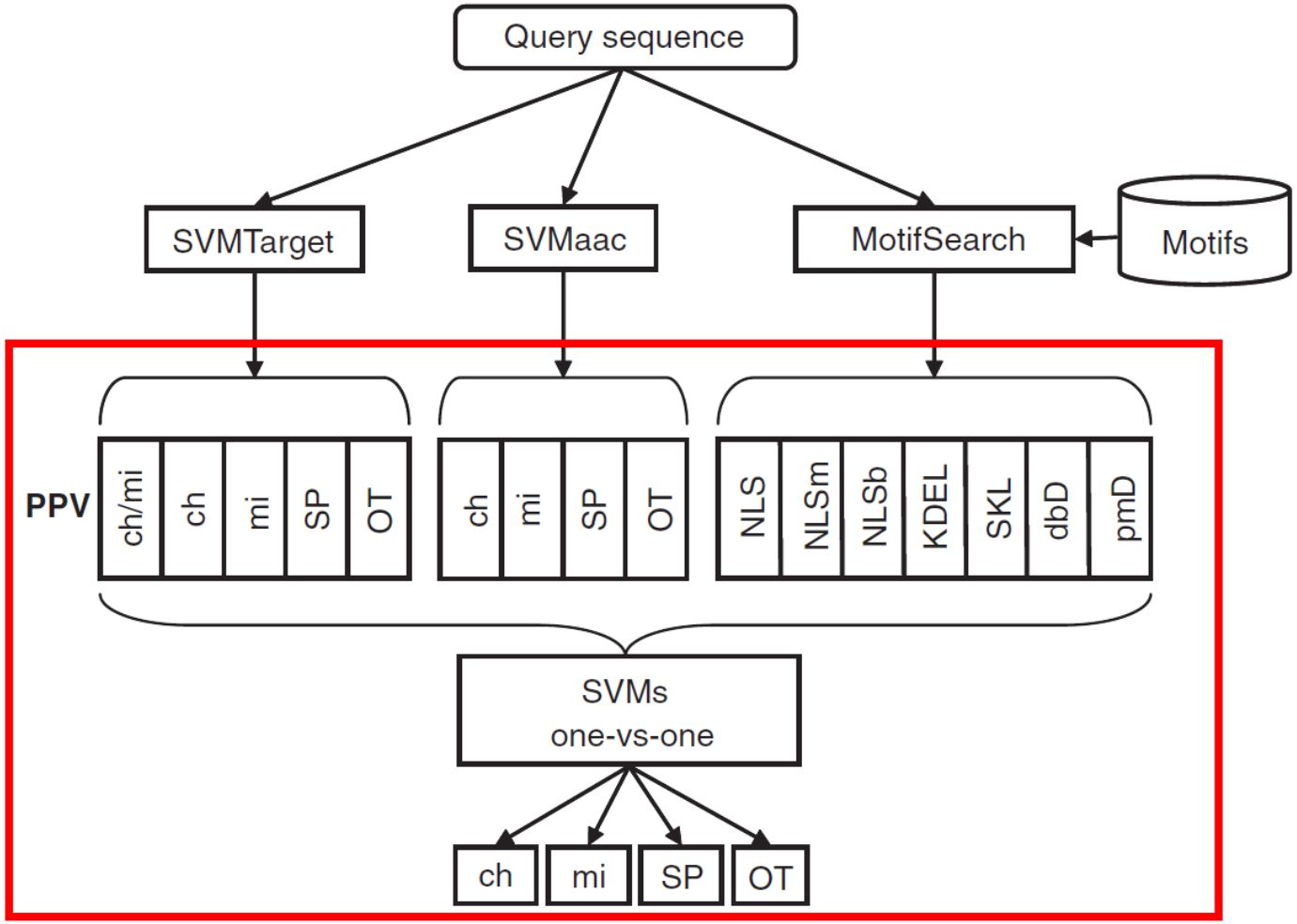


Method Comparison (cont.)

- **TargetLoc**

➢ Second Layer SVM [1]

- All the first layer results form PPV vectors
- **"1 vs 1"** (Comparing each other)
  (In fact a Multi-class SVM)



Method Comparison (cont.)

- **m-SVM** [2]

➢ Where does the similarity come from?

- Motif – a sequence of amino acid pattern, e.g. "ABC"
- Motif Composition – a permutation of amino acids and gaps, e.g. "■□□■■"
- Motif + Motif Composition = "A□□BC"
- We need to compare 3 things (every latter one depends on the former one):
  1. Amino Acid
  2. Motif
  3. Motif Composition

➢ How to compare similarity?

- **Kernel ≈ Similarity!**

Method Comparison (cont.)

- **m-SVM** [2]

➢ <u>Amino Acid Kernel</u>

- Recall what a substitution matrix is:

  A substitution matrix describes the rate at which one character in a sequence changes to other character states over time. [9]

- BLOSUM is a good option:

$$S_{ij} = \left(\frac{1}{\lambda}\right) \log\left(\frac{p_{ij}}{q_i \cdot q_j}\right) \text{ [10]}$$

- **Amino Acid Kernel**:

  $K_1^{AA}(a,b) = \sum_c p_{ac} - p_{ab}$,

  which is the graph Laplacian.

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	-1	5																		
Asn	-2	0	6																	
Asp	-2	-2	1	6																
Cys	0	-3	-3	-3	9															
Gln	-1	1	0	0	-3	5														
Glu	-1	0	0	2	-4	2	5													
Gly	0	-2	0	-1	-3	-2	-2	6												
His	-2	0	1	-1	-3	0	0	-2	8											
Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

Method Comparison (cont.)

- **m-SVM** [2]

➢ Motif Kernel

- Merely extend the length of objects
- **Motif Kernel:**

  $K_r^{AA}(s,t) = \sum_{i=1}^{r} K_1^{AA}(s_i, t_i)$, where r is the motif length.

➢ Motif Composition Kernel

- For any given pattern, compute the empirical distribution of corresponding motifs from a given amino-acid sequence, which is a histogram of occurrences of each possible r-mer sequence.
- **Motif Composition Kernel:**

  $$K_r^{JS}(p,q) = \sum_{s\in\mathcal{A}^r} \sum_{t\in\mathcal{A}^r} K_r^{AA}(s,t) \cdot \left(p(s) \cdot \log\frac{p(s)}{p(s)+q(t)} + q(s) \cdot \log\frac{q(s)}{p(s)+q(t)}\right),$$

  where Jensen-Shannon divergence is used to compare the similarity of two distribution.

  (NOT arbitrary vector but carrying a special structure, so NO RBF.)

Method Comparison (cont.)

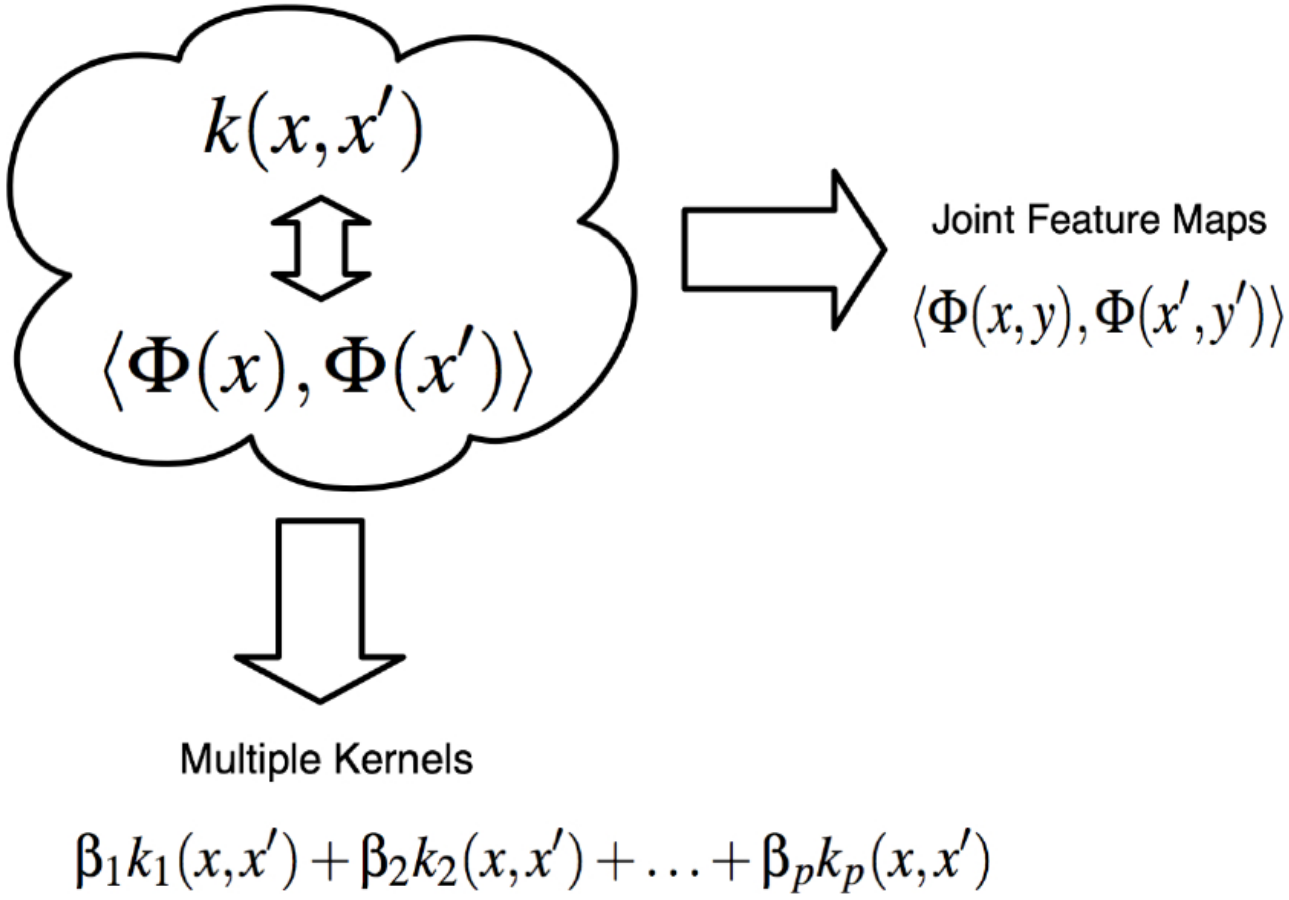
- **m-SVM** [2][3]

➢ <u>Final Kernel</u>

- Joint Feature Space:
  Describe an object from various angles (features) at the same time.

- Weighted finite mixture model

- **Final Kernel:**
  $K(\boldsymbol{x}, \boldsymbol{x}') = \sum_i \beta_i \cdot K_i(\boldsymbol{x}, \boldsymbol{x}')$



Method Comparison (cont.)

- **m-SVM** [2][11]

➢ Classifier

- **Multiple Kernel SVM**

- **Confidence Function**

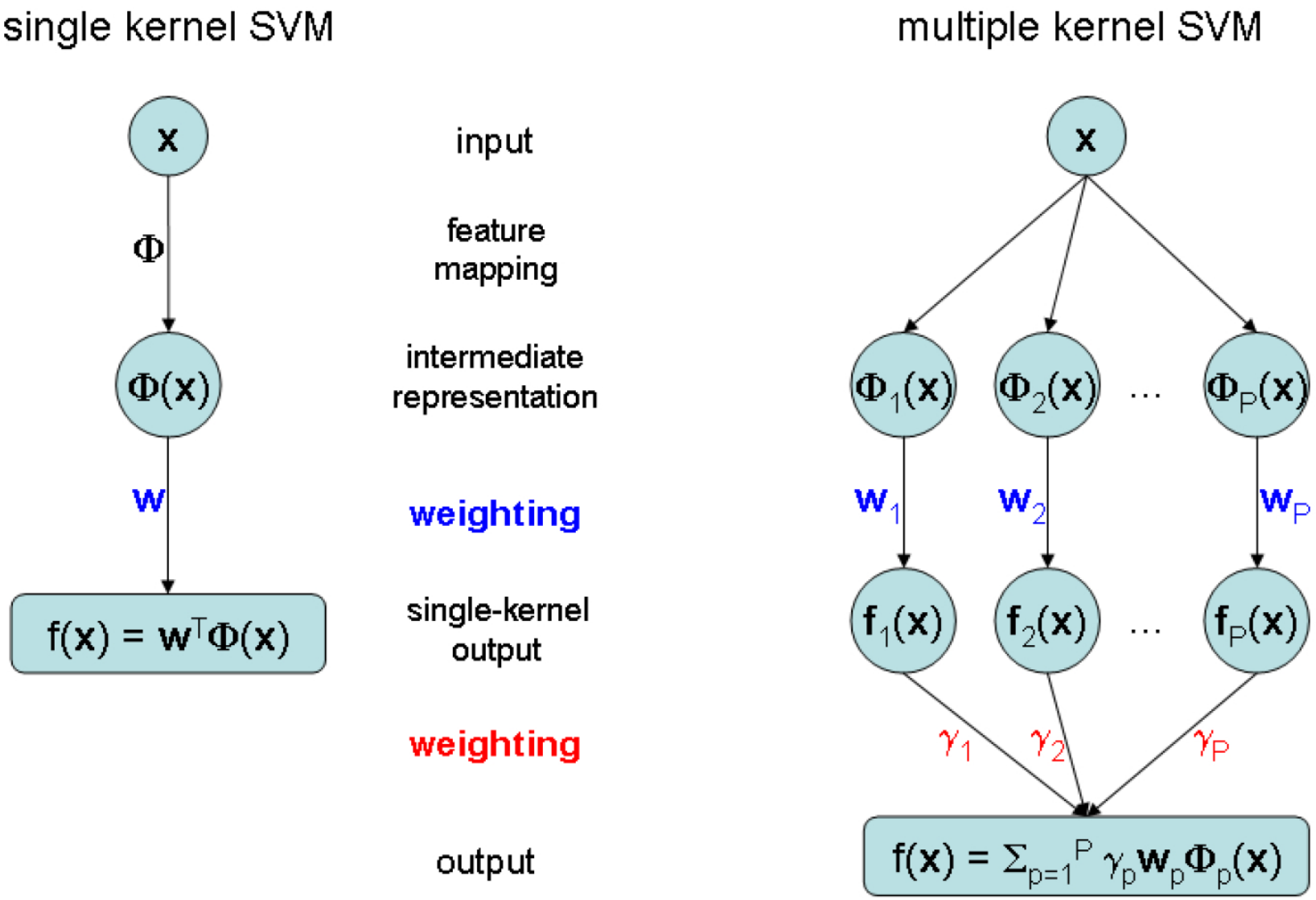
For each class u,

$$f_u(\boldsymbol{x}) = \left\langle \boldsymbol{w}_u, \sum_i \beta_i \Phi_i(\boldsymbol{x}) \right\rangle$$

$$= \sum_j \alpha_{ju} \sum_i \beta_i K_i(\boldsymbol{x}, \boldsymbol{x}_j)$$

- **Classification**

$$\hat{y}(\boldsymbol{x}) = argmax_u f_u(\boldsymbol{x})$$



Experimental Results

- **m-SVM Experimental Settings** [1][11]
  - Motif kernels up to length 5.
  - Compute the motif kernels on different sections of the protein sequences, namely the first 15 and 60 amino acids from the N-terminus and the 15 amino acids from the C-terminus ($4 \times 2^{5-1} = 64$ motif kernels).
  - 3 BLAST similarity kernel
    - Linear kernel on E-Values
    - Gaussian kernel on E-Values, width 1000
    - Gaussian kernel on log E-Values, width 1e5
  - 2 phylogenetic kernels [12]
    - Linear kernel
    - Gaussian kernel, width 300

Experimental Results (cont.)

Table 2. Performance comparison of TargetLoc against the TargetP and iPSORT methods, using the TargetP size non-equalized dataset (940 plant and 2738 non-plant proteins)

Version	Method	Category	*SE*	*SP*	*MCC*	correct[%]
Plant	TargetLoc	ch	0.88	0.76	0.78	89.7 (±1.6)
		mi	0.87	0.94	0.84	
		SP	0.93	0.97	0.93	
		OT	0.92	0.84	0.86	
	TargetP	ch	0.85	0.69	0.72	85.3 (±3.5)
		mi	0.82	0.90	0.77	
		SP	0.91	0.95	0.90	
		OT	0.85	0.78	0.77	
	iPSORT	ch	0.68	0.71	0.64	83.4
		mi	0.84	0.86	0.75	
		SP	0.91	0.98	0.92	
		OT	0.83	0.70	0.71	
Non-plant	TargetLoc	mi	0.91	0.77	0.81	92.5 (±1.2)
		SP	0.95	0.92	0.91	
		OT	0.91	0.97	0.86	
	TargetP	mi	0.89	0.67	0.73	90.0 (±0.7)
		SP	0.96	0.92	0.92	
		OT	0.88	0.97	0.82	
	iPSORT	mi	0.74	0.68	0.67	88.5
		SP	0.92	0.92	0.90	
		OT	0.90	0.92	0.78	

Data	Class	Our Method				
		Accuracy	Precision	Recall	F1-Score	MCC
plant	ch	96.7 ± 0.4	95.4	84.4	89.5 ± 1.4	87.8 ± 1.5
	mi	95.3 ± 0.4	92.0	97.3	94.6 ± 0.4	90.5 ± 0.8
	SP	97.4 ± 0.3	96.0	94.5	95.2 ± 0.7	93.5 ± 0.9
	OT	95.6 ± 0.3	87.3	86.7	86.9 ± 1.4	84.3 ± 1.6
	avg	96.2 ± 0.4	92.9	92.7	92.7 ± 0.8	89.9 ± 1.1
nonplant	mi	96.9 ± 0.2	87.8	90.1	88.9 ± 0.9	87.1 ± 1.0
	SP	96.8 ± 0.3	94.4	93.6	94.0 ± 0.6	91.8 ± 0.8
	OT	94.9 ± 0.3	95.9	95.7	95.8 ± 0.3	89.3 ± 0.7
	avg	95.7 ± 0.3	94.4	94.4	94.4 ± 0.4	89.7 ± 0.8

Future Work

- **Everything reasonable could be feature**
  - Recall feature actually is description from a given perspective
  - How about adding Text-Based features? [13][14]
  - Something else known to be useful

- **What is the underlying principle that m-SVM could outperform “1 vs 1” and even “1 vs Rest” scheme?**

- **How could we know which sub-kernel contributes the result more?**

- **Could we use the method used in Boosting to compute the weights?**

Conclusion

- m-SVM could achieve comparable or even better experimental results.
- Multi-layer prediction system using binary SVMs could be converted to a single kernel SVM.
- By applying m-SVM in a joint feature space, it provides a framework to introduce more comprehensive and diverse features to describe an object.
- Kernel for each feature is emphasized more, which means it makes learning with kernels more interesting.

References

[1] Höglund, Annette, et al. "MultiLoc: prediction of protein subcellular localization using N-terminal targeting sequences, sequence motifs and amino acid composition." Bioinformatics 22.10 (2006): 1158-1165.

[2] Ong, Cheng Soon, and Alexander Zien. "An automated combination of kernels for predicting protein subcellular localization." Algorithms in Bioinformatics. Springer Berlin Heidelberg, 2008. 186-197.

[3] Zien, Alexander, and Cheng Soon Ong. "Multiclass multiple kernel learning." Proceedings of the 24th international conference on Machine learning. ACM, 2007.

[4] Park, Keun-Joon, and Minoru Kanehisa. "Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs." Bioinformatics 19.13 (2003): 1656-1663.

[5] Scholkopf, Bernhard, and Alex Smola. "Learning with kernels." (2002).

References (cont.)

[6] Gupta, Shobhit, et al. "Quantifying similarity between motifs." Genome Biol 8.2 (2007): R24.

[7] Nakai, Kenta, and Paul Horton. "PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization." Trends in biochemical sciences 24.1 (1999): 34-35.

[8] Brock, Tom. "Histone Methylation: SET versus Jumonji.“

[9] http://en.wikipedia.org/wiki/Substitution_matrix

[10] http://en.wikipedia.org/wiki/BLOSUM

[11] http://raetschlab.org/lectures/mkl-tutorial.pdf/at_download/file

[12] Pellegrini, Matteo, et al. "Assigning protein functions by comparative genome analysis: protein phylogenetic profiles." Proceedings of the National Academy of Sciences 96.8 (1999): 4285-4288.

References (cont.)

[13] Shatkay, Hagit, et al. "SherLoc: high-accuracy prediction of protein subcellular localization by integrating text and protein sequence data." Bioinformatics 23.11 (2007): 1410-1417.

[14] Brady, Scott, and Hagit Shatkay. "EpiLoc: a (working) text-based system for predicting protein subcellular location." Pacific Symposium on Biocomputing. Vol. 13. 2008.