

Spatially-constrained sample kernel for sequence classification

Pavel P. Kuksa, Pai-Hsi Huang, and Vladimir Pavlovic
Department of Computer Science, Rutgers University
{pkuksa, paihuang, vladimir}@cs.rutgers.edu

Objective

Fast and accurate biosequence classification methods for structural and functional annotation based on *primary sequences only*

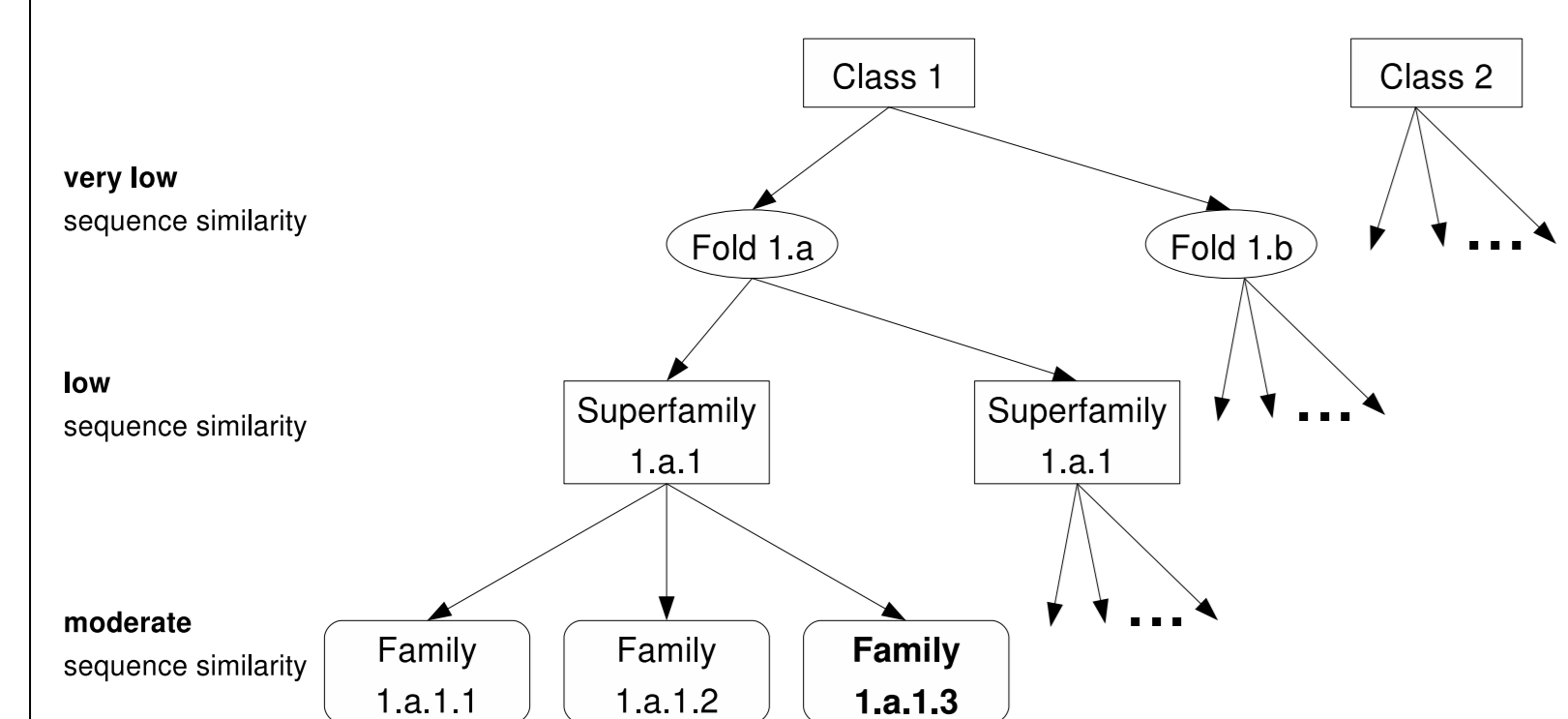
- Remote homology (superfamily) prediction (distant evolutionary relationship)

Challenges

- Highly dissimilar sequences
- Limited labeled examples
- Modeling complex evolutionary transformations

Protein remote homology and fold prediction

- Address the remote homology and fold detection problems in the context of SCOP (Structural Classification of Proteins) [5]



State-of-the-art

- Many methods:
 - generative: profile HMMs [2]
 - model-based discriminative: SVMFisher [3],
 - string-based discriminative approaches: Mismatch [4]
 - others
- Performance still sub-optimal
- Semi-supervised learning
 - significant improvement in accuracy
 - high computational complexity

SCOP data set

- 23 superfamilies, 54 binary prediction problems
- 2862 labeled and 4467 unlabeled sequences

New Sequenced-induced Multi-Resolution Kernels: Fast, Accurate and Sensitive

- Scan sequences at multiple resolutions
- Consider tuples of sequence elements
- Incorporate spatial configuration of sequences
- Efficiently model complex sequence transformations (e.g. mutations, insertions, deletions, etc.)

Defined by three parameters:

$$K^{(t,k,d)}(X, Y) = \sum_{a_i \in \Sigma^k, 0 \leq d_i < d} C(a_1, d_1, \dots, a_{t-1}, d_{t-1}, a_t | X) \cdot C(a_1, d_1, \dots, a_{t-1}, d_{t-1}, a_t | Y)$$

\times

		A	R	N	D	C	Q												
α																			

 \times

		A	R			N	D			C	Q								
$a_1 \leftarrow d_1 \rightarrow a_2 \leftarrow d_2 \rightarrow a_3$																			

Contiguous k-mer feature α of a traditional spectrum/mismatch kernel (top) contrasted with the sparse spatial samples of the proposed kernel (bottom).

Algorithm 1. Spatial Sample Kernel

Input: set of strings S , parameters t, d
Repeat for every $d_1, \dots, d_t \in \{1, \dots, d\}$
 1: Extract t -tuples of features from each $s \in S$
 2: Sort t -tuples with t counting sort rounds
 3: Scan the sorted list and for each distinct feature f update the kernel matrix K for all strings containing f
Output: kernel matrix K

Leveraging Unlabeled Data Using Semi-supervised Learning

$$K^{new}(X, Y) = \frac{\sum_{X' \in N(X), Y' \in N(Y)} K(X', Y')}{|N(X)||N(Y)|}$$

Algorithm 2. Semi-Supervised Spatial Sample Kernel

Input: set of string sets
 $N(S) = \{N(s_1), \dots, N(s_N)\}$, where $N(s_i)$ is a set of neighbors for s_i , parameters t, d
Repeat for every $d_1, \dots, d_t \in \{1, \dots, d\}$
 1: Extract t -tuples of features from each input set $N(s) \in N(S)$
 2: Sort t -tuples with t counting sort rounds
 3: Scan the sorted list and for each distinct feature f update the kernel matrix K for all string groups containing f
Output: kernel matrix K

Complexity Analysis

Method	Time complexity
Triple kernel	$O(d^2 n N + d^2 \Sigma ^3 N^2)$
Double kernel	$O(d n N + d \Sigma ^2 N^2)$
Mismatch	$O(k^{m+1} \Sigma ^m n N + \Sigma ^k N^2)$
Profile kernel	$O(k M_\sigma n N + \Sigma ^k N^2)$
Neighborhood kernels	
Triple kernel	$O(d^2 H n N + d^2 \Sigma ^3 N^2)$
Double kernel	$O(d H n N + d \Sigma ^2 N^2)$
Mismatch	$O(k^{m+1} \Sigma ^m H n N + \Sigma ^k N^2)$

N - number of sequences, n - sequence length,
 H is the sequence neighborhood size, $|\Sigma|$ is the alphabet size
 k, m are the mismatch kernel parameters ($k = 5, 6$ and $m = 1, 2$ in most cases)
 M_σ is the profile neighborhood size, $k^m |\Sigma|^m \leq M_\sigma \leq |\Sigma|^k$

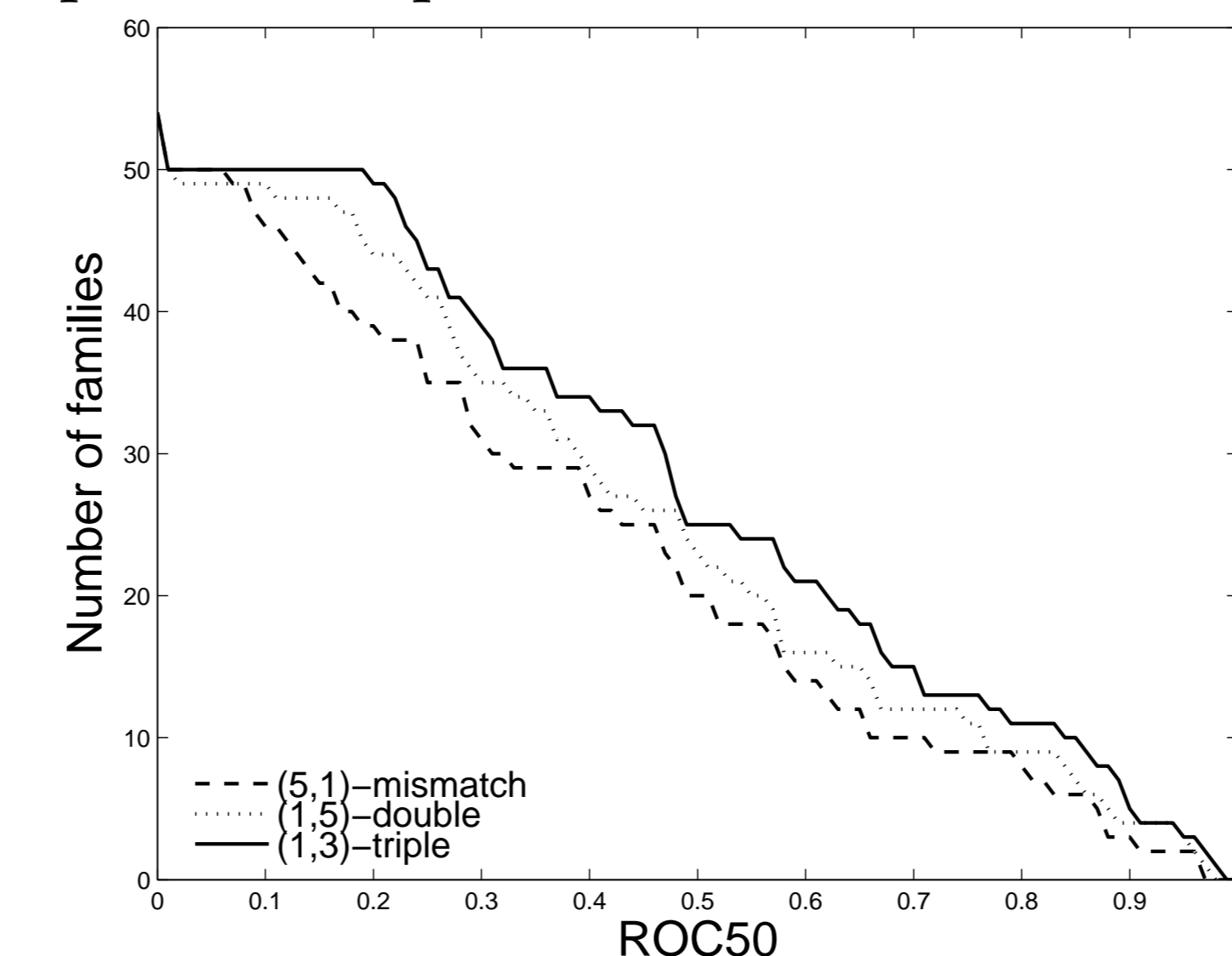
Experiments: Multi-Class Fold Recognition

Method	Error	Top 5 Error	Balanced Error	Top 5 Balanced Error
Supervised				
SVM(D&D) [1]	-	-	56.5	-
Mismatch(5,1)	51.17	22.72	53.22	28.86
Double(1,5)	44.13	23.50	46.19	23.92
Triple(1,3)	41.51	18.54	44.99	21.09
Semi-supervised (Non-redundant data set)				
Profile(5,7,5)	31.85	15.14	32.17	16.73
Double(1,5)	28.72	14.99	24.74	11.6
Triple(1,3)	24.28	12.79	22.38	11.79

27 folds, less than 35% sequence identities

Experiments: Remote Homology Detection

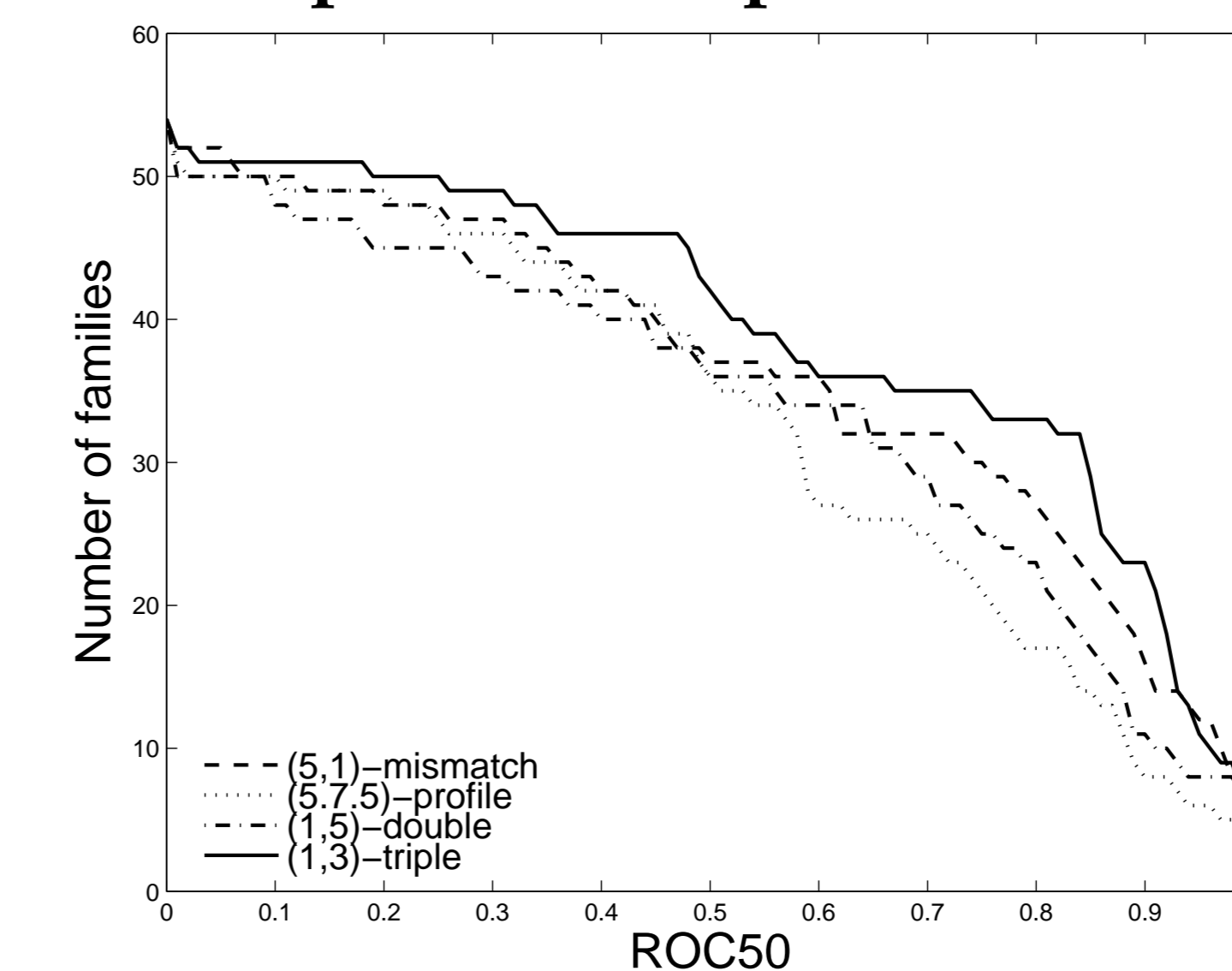
Supervised Experiments



Method	ROC	ROC50	# dim.	Time (s)
(5, 1)-mismatch	0.8749	0.4167	3200000	938
SVM-pairwise	0.8930	0.4340	-	-
Fisher	0.7730	0.2500	-	-
(1,5) double	0.8901	0.4629	2000	54
(1,3) triple	0.9148	0.5118	72000	112

- Order-of-magnitude running time improvement
- Significant improvement in accuracy

Semi-Supervised Experiments



Semi-supervised Methods	ROC	ROC50
(5, 1)-mismatch	0.9093	0.6745
(5,7,5)-profile	0.9190	0.6069
(1,5)-double	0.9131	0.6279
(1,3)-triple	0.9382	0.7262

- Significantly improved accuracy

Features Induced by Kernels

$S = \text{HKYNQLIM}$

spectrum-5	mismatch(5,1)	double-(1,5)
HKYNQ	XKYNQ	XYNQL
KYNQL	HKYNQ	KXNQL
YNQLI	HKXNQ	KYXQL
NQLIM	HKYXQ	KYNXL
	HKYXN	YKNQX
	XNQLI	XQLIM
	YXQLI	NXLIM
	YNXLI	NQXIM
	YNQXI	NQLXM
	YNQLX	NQLIX

'X' corresponds to $|\Sigma|$ features

- Spatial kernels: low-dimensional, sparse representation, very few features
- Mismatch kernels: high-dimensional

Modeling Sequence Evolution

- Example: a slightly diverged region

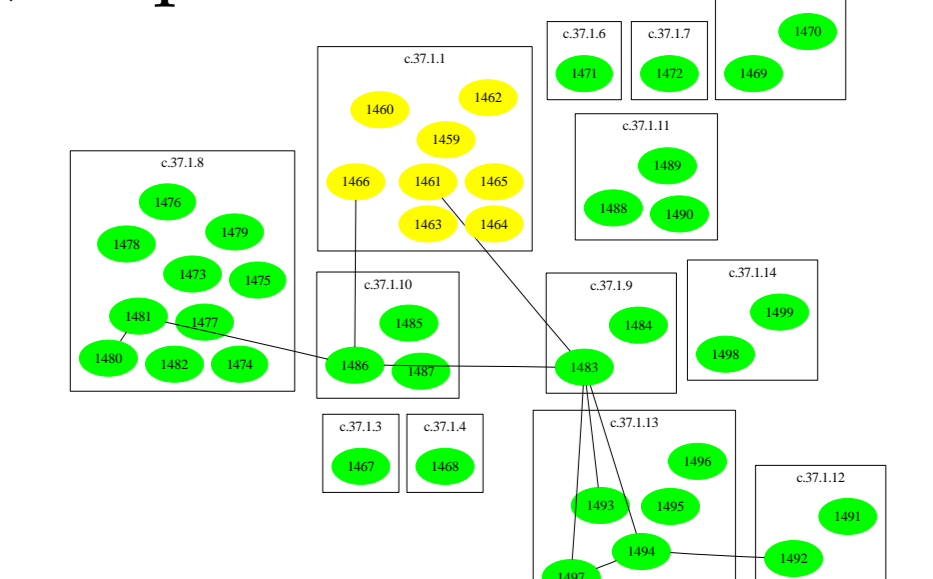
	$S = \text{HKYNQLIM}$	$S' = \text{HKINQIIM}$
mismatch (5,1)	XKYNQ	XKINQ
	HKYNQ	HKINQ
	HKXNQ	HKXNQ
	HKYXQ	HKIXQ
	HKYXN	HKINX
	YNQLI	YNQI
double-(1,5)	YN	YN
	YN	YN
	YN	YN
	YN	YN
	YN	YN
	YN	YN

- Mismatch kernel: low similarity score, very few sequence features retained
- Spatial kernel: high similarity score, still significant overlap in sequence features

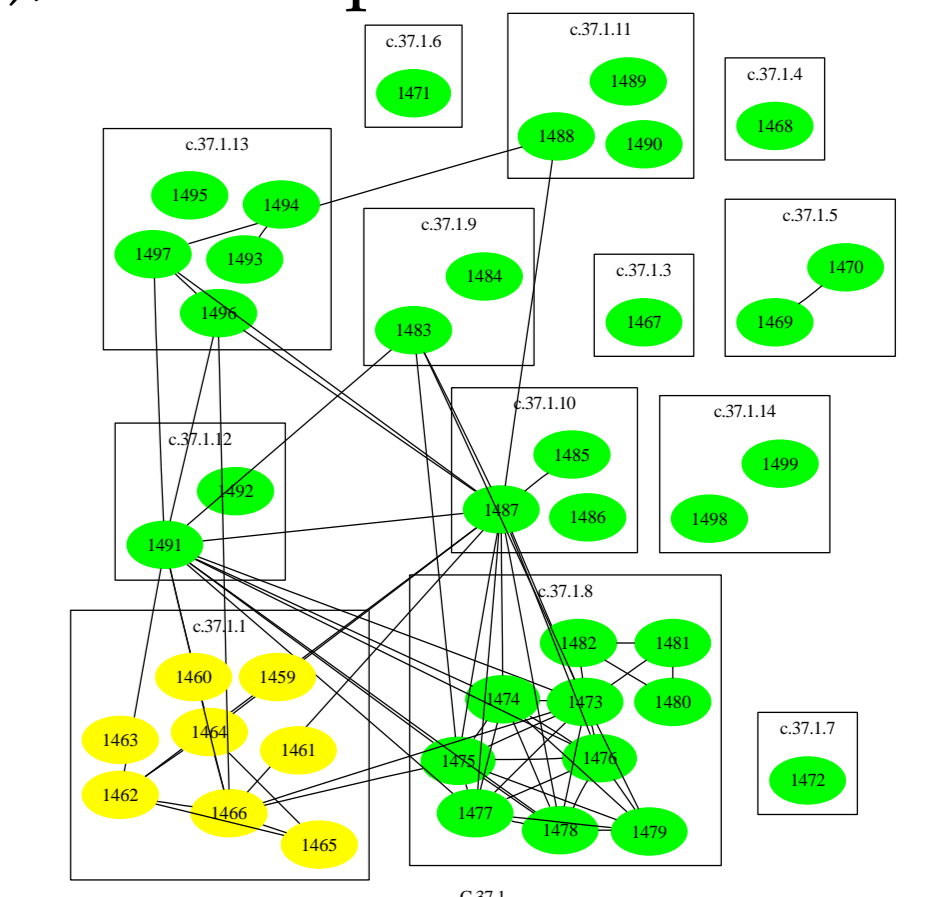
Kernel-induced data manifolds

- Only used unlabeled sequences in SCOP
- Normalized kernel matrices

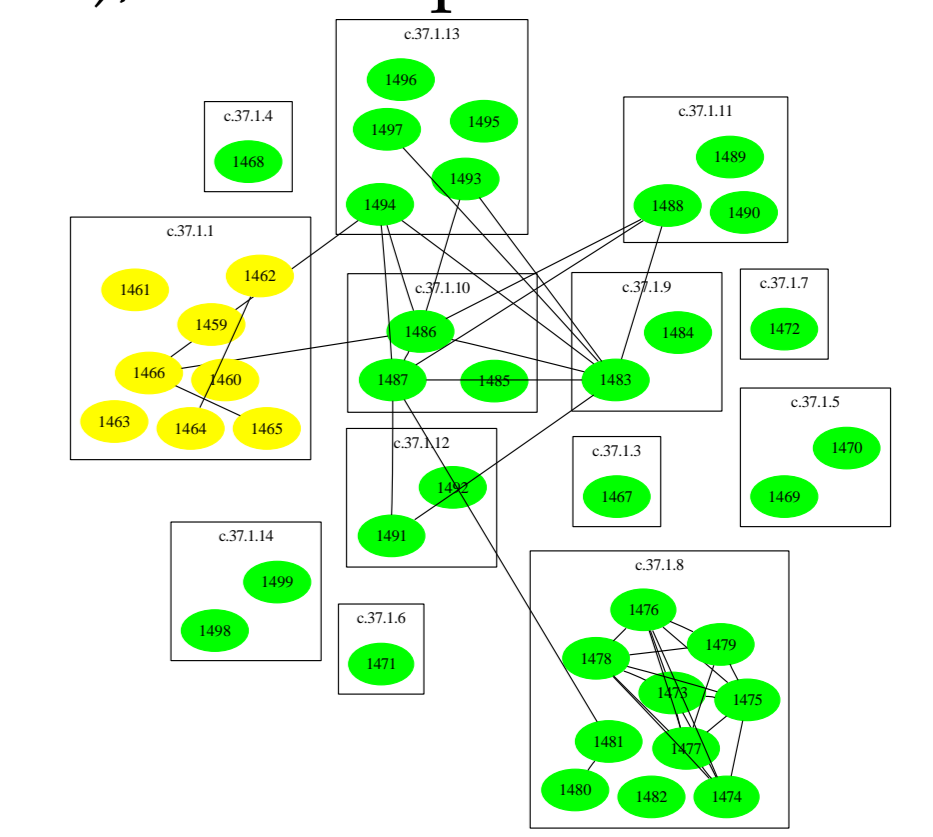
Triple(1,3), supervised:



Triple(1,3), semi-supervised



Profile(5,7,5), semi-supervised:



Thresholded at low (< 10%) FDR
Green: training; Yellow: testing

Conclusions and Future Works

- Significantly improved performance and sensitivity for remote homology prediction
- Highly scalable; work with large unlabeled data set in the future
- Applications in text, music categorization

References

- Chris H.Q. Ding and Inna Dubchak. Multi-class protein fold recognition using support vector machines an d neural networks. *Bioinformatics*, 17(4):349–358, 2001.
- SR Eddy. Profile hidden Markov models. *Bioinformatics*, 14(9):755–763, 1998.
- Tommi Jaakkola, Mark Diekhans, and David Haussler. A discriminative framework for detecting remote protein homologies. In *Journal of Computational Biology*, volume 7, pages 95–114, 2000.
- Christina S. Leslie, Eleazar Eskin, Jason Weston, and William Stafford Noble. Mismatch string kernels for svm protein classification. In *NIPS*, pages 1417–1424, 2002.
- L. Lo Conte, B. Ailey, T.J. Hubbard, S.E. Brenner, A.G. Murzin, and C. Chothia. SCOP: a structural classification of proteins database. *Nucleic Acids Res.*, 28:257–259, 2000.