KERNEL-BASED DATA FUSION AND ITS APPLICATION TO PROTEIN FUNCTION PREDICTION IN YEAST: SUPPLEMENTARY DATA

GERT R. G. LANCKRIET
Division of Electrical Engineering, University of California, Berkeley

MINGHUA DENG
Department of Biological Sciences, University of Southern California

NELLO CRISTIANINI
Department of Statistics, University of California, Davis

MICHAEL I. JORDAN
Division of Computer Science, Department of Statistics, University of California, Berkeley

WILLIAM STAFFORD NOBLE
Department of Genome Sciences, University of Washington

Table 3: **Classification performance using single kernels.** Each entry in the table is the mean ROC score of a 1-norm SVM with C=1 for a given functional classification, averaged across 15 trials and using as input a single kernel matrix. Kernels are described in the text; 'K'_{Pfam} and K'_{Exp} refer to the enriched versions of the corresponding kernels. The 13 yeast functional classifications are listed in Table 1.

Class	K_{SW}	K_{Pfam}	$K_{Pfam'}$	K_{Exp}
1	.8589 ± .0033	.8373 ± .0037	.8376 ± .0037	.5424 ± .0028
2	.8464 ± .0072	.8107 ± .0113	.8196 ± .0103	.5424 ± .0056
3	.8013 ± .0060	.7547 ± .0062	.7669 ± .0050	.5421 ± .0055
4	.8540 ± .0052	.8085 ± .0048	.8201 ± .0045	.5509 ± .0028
5	.8961 ± .0057	.8349 ± .0058	.8885 ± .0056	.6136 ± .0056
6	.8157 ± .0040	.8069 ± .0046	.7993 ± .0051	.5166 ± .0026
7	.8267 ± .0077	.8010 ± .0074	.7861 ± .0061	.5319 ± .0019
8	.7868 ± .0082	.7023 ± .0089	.7437 ± .0128	.5487 ± .0063
9	.7512 ± .0131	.7309 ± .0113	.7302 ± .0078	.5343 ± .0042
10	.7652 ± .0081	.6906 ± .0079	.6746 ± .0059	.5219 ± .0037
11	.6390 ± .0071	.5952 ± .0148	.5910 ± .0104	.4970 ± .0061
12	.9180 ± .0079	.9331 ± .0038	.9461 ± .0034	.5457 ± .0025
13	.7813 ± .0188	.6678 ± .0227	.7379 ± .0225	.5061 ± .0081

Class	$K_{Exp'}$	K_{TAP}	K_{Phys}	K_{Gen}
1	.7169 ± .0049	.6171 ± .0043	.6450 ± .0045	.6133 ± .0026
2	.7490 ± .0076	.5782 ± .0082	.6349 ± .0077	.6061 ± .0057
3	.6859 ± .0057	.5583 ± .0074	.6651 ± .0056	.6802 ± .0068
4	.7225 ± .0057	.5877 ± .0127	.7033 ± .0075	.6335 ± .0056
5	.8750 ± .0045	.5394 ± .0085	.6671 ± .0072	.6257 ± .0048
6	.6305 ± .0062	.5615 ± .0059	.6429 ± .0062	.6343 ± .0045
7	.5987 ± .0097	.6374 ± .0042	.7422 ± .0068	.6705 ± .0068
8	.6245 ± .0102	.5741 ± .0069	.5984 ± .0057	.5876 ± .0066
9	.6205 ± .0137	.6324 ± .0079	.6713 ± .0119	.5858 ± .0075
10	.6204 ± .0053	.5363 ± .0074	.6581 ± .0076	.6236 ± .0063
11	.4987 ± .0120	.5072 ± .0099	.5383 ± .0122	.5079 ± .0087
12	.7126 ± .0057	.6499 ± .0048	.6655 ± .0043	.6093 ± .0045
13	.5468 ± .0211	.5586 ± .0101	.5911 ± .0163	.5573 ± .0193

Table 4: **Performance using binary data.** Each row corresponds to one yeast functional classification, as given in Table 1. Kernels are computed using binary versions of each data set, as described in the text. The first five columns list the weights assigned to each kernel via SDP. The weights are normalized to sum of the number of kernels. The final three columns list the mean ROC score for the SDP/SVM method, the MRF method and the SVM method using an unweighted sum of kernels. The SVM uses a 1-norm soft margin with $C = 1$.

	Kernel weights					ROC	
Class	K_{Pfam}	K_{Exp}	K_{TAP}	K_{Phys}	K_{Gen}	SDP/SVM	MRF
1	1.41	1.88	0.16	0.53	1.01	$.8825 \pm .0042$	$.7532 \pm .0042$
2	1.16	1.94	0.04	0.70	1.17	$.8563 \pm .0100$	$.7173 \pm .0102$
3	0.65	2.05	0.08	0.70	1.52	$.8464 \pm .0053$	$.6990 \pm .0045$
4	1.09	1.55	0.17	0.98	1.21	$.9024 \pm .0028$	$.7409 \pm .0049$
5	1.26	1.58	0.05	0.79	1.32	$.9094 \pm .0050$	$.7375 \pm .0102$
6	1.30	1.54	0.12	0.76	1.28	$.8742 \pm .0049$	$.7183 \pm .0059$
7	0.91	1.53	0.47	1.15	0.93	$.9149 \pm .0040$	$.7534 \pm .0085$
8	0.82	2.37	0.07	0.56	1.19	$.8023 \pm .0094$	$.7285 \pm .0120$
9	0.95	2.13	0.51	0.72	0.69	$.8623 \pm .0091$	$.6849 \pm .0107$
10	0.57	2.31	0.20	0.94	0.98	$.8120 \pm .0078$	$.6954 \pm .0060$
11	0.53	2.58	0.02	0.64	1.23	$.6575 \pm .0093$	$.5691 \pm .0092$
12	2.21	0.93	0.74	0.94	0.18	$.9674 \pm .0023$	$.8575 \pm .0076$
13	0.60	2.90	0.07	0.68	0.75	$.8083 \pm .0091$	$.6612 \pm .0169$

Table 5: **Performance using the enriched expression kernel.** The table contains experimental results using the enriched expression kernel, as described in the text. See the caption to Table 4 for an explanation of individual columns. For comparison, mean ROC scores using the binary data are given in the final column of the table.

	Kernel weights					SDP/SVM ROC	
Class	K_{Pfam}	$K_{Exp'}$	K_{TAP}	K_{Phys}	K_{Gen}	Enriched	Binary
1	1.43	0.38	0.62	0.87	1.70	$.8892 \pm .0045$	$.8825 \pm .0042$
2	1.22	0.48	0.12	1.04	2.14	$.8783 \pm .0064$	$.8563 \pm .0100$
3	1.00	0.42	0.65	0.89	2.04	$.8490 \pm .0060$	$.8464 \pm .0053$
4	1.21	0.39	0.40	1.24	1.75	$.9134 \pm .0031$	$.9024 \pm .0028$
5	1.17	1.08	0.07	1.08	1.61	$.9385 \pm .0047$	$.9094 \pm .0050$
6	1.48	0.27	0.50	1.02	1.74	$.8806 \pm .0054$	$.8742 \pm .0049$
7	1.15	0.24	0.69	1.51	1.41	$.9162 \pm .0049$	$.9149 \pm .0040$
8	0.99	0.26	0.90	1.09	1.76	$.8235 \pm .0104$	$.8023 \pm .0094$
9	1.15	0.31	1.02	1.14	1.39	$.8568 \pm .0097$	$.8623 \pm .0091$
10	0.87	0.63	0.84	1.25	1.41	$.8238 \pm .0068$	$.8120 \pm .0078$
11	1.01	0.50	0.80	1.02	1.67	$.6532 \pm .0097$	$.6575 \pm .0093$
12	2.24	0.19	0.71	1.21	0.64	$.9695 \pm .0026$	$.9674 \pm .0023$
13	1.01	0.73	0.77	1.18	1.30	$.7988 \pm .0180$	$.8083 \pm .0091$

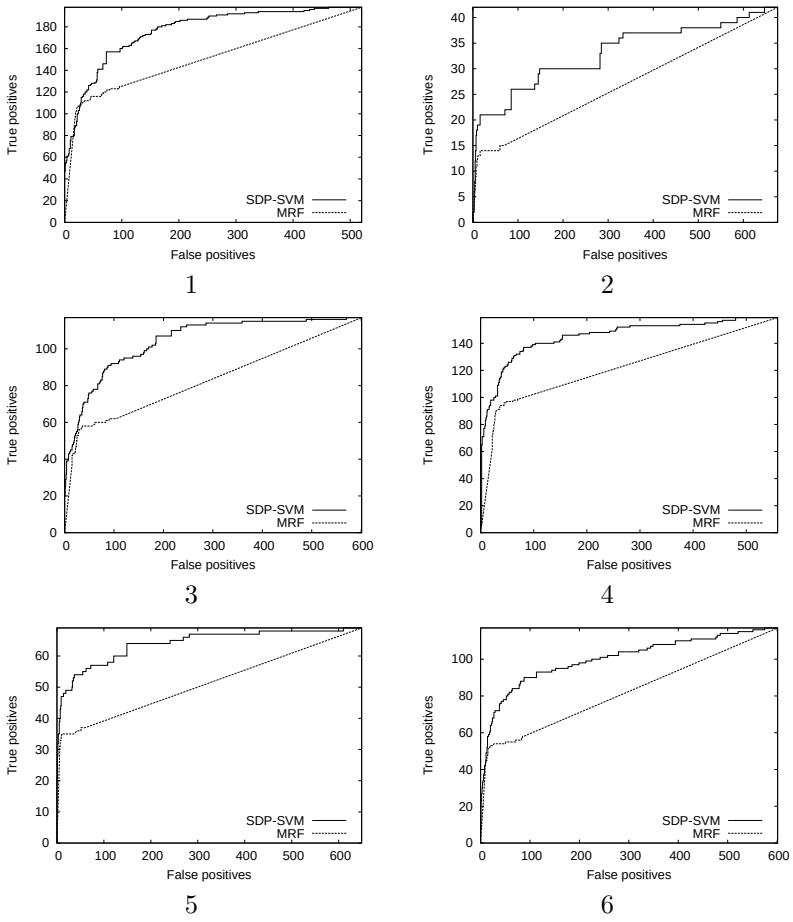

Figure 1: **ROC curves for each class.** Each plot includes the (unnormalized) ROC curve for the MRF and SDP/SVM method on one class. Each method is run using all six input data types in binary format. Diagonal segments of the MRF curves correspond to runs of genes that are all assigned the same probability.

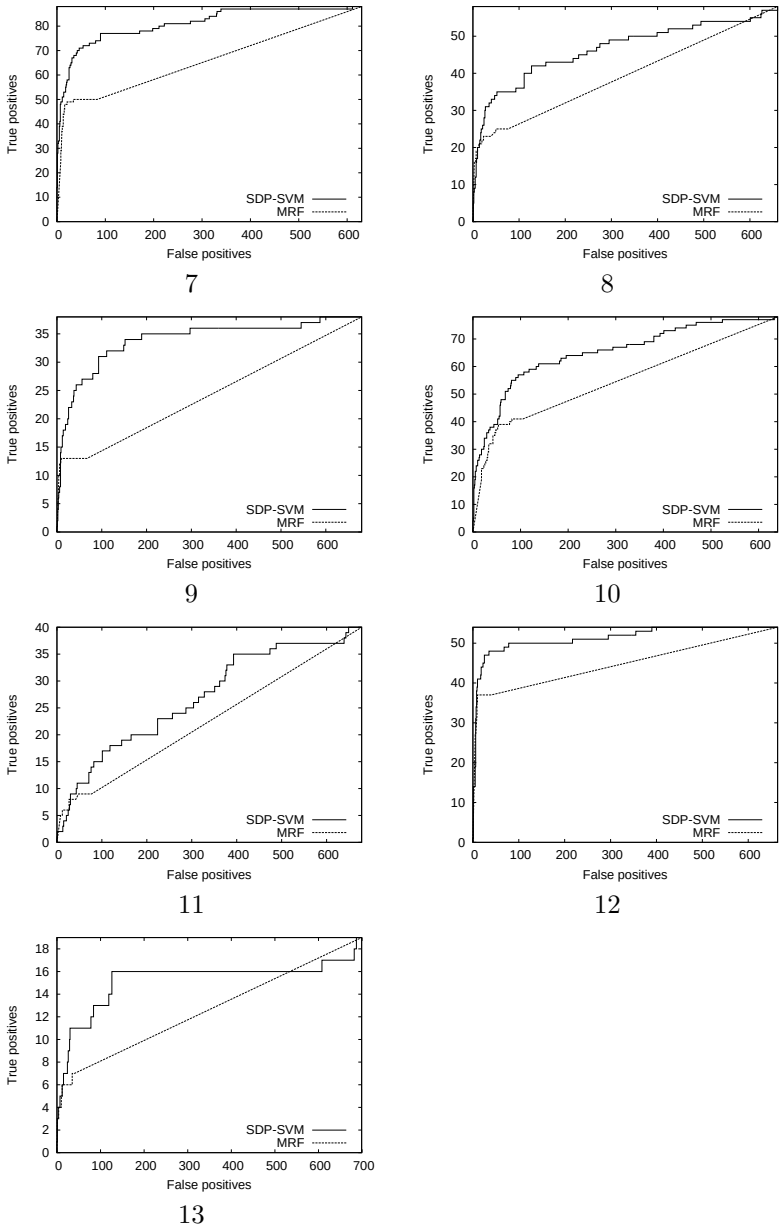

Figure 2: **ROC curves for each class (continued).**

Table 6: **Performance using enriched Pfam and expression kernels.** The table contains experimental results using the enriched Pfam and expression kernels, as described in the text. See the caption to Table 4 for an explanation of individual columns. For comparison, mean ROC scores using the binary data are given in the final column of the table.

	Kernel weights					SDP/SVM ROC	
Class	$K_{Pfam'}$	$K_{Exp'}$	K_{TAP}	K_{Phys}	K_{Gen}	Enriched	Binary
1	1.60	0.31	0.49	0.95	1.65	$.8875 \pm .0037$	$.8825 \pm .0042$
2	1.27	0.47	0.18	1.01	2.08	$.8764 \pm .0073$	$.8563 \pm .0100$
3	1.12	0.42	0.56	0.90	2.01	$.8487 \pm .0049$	$.8464 \pm .0053$
4	1.38	0.34	0.32	1.17	1.79	$.9161 \pm .0032$	$.9024 \pm .0028$
5	1.55	0.94	0.06	1.04	1.41	$.9419 \pm .0046$	$.9094 \pm .0050$
6	1.58	0.25	0.42	1.03	1.72	$.8743 \pm .0051$	$.8742 \pm .0049$
7	1.10	0.20	0.84	1.49	1.36	$.9116 \pm .0051$	$.9149 \pm .0040$
8	1.01	0.26	0.85	1.09	1.79	$.8167 \pm .0117$	$.8023 \pm .0094$
9	1.11	0.19	1.22	1.18	1.30	$.8616 \pm .0087$	$.8623 \pm .0091$
10	1.05	0.56	0.81	1.23	1.35	$.8162 \pm .0058$	$.8120 \pm .0078$
11	1.10	0.53	0.75	1.02	1.60	$.6560 \pm .0086$	$.6575 \pm .0093$
12	2.18	0.17	0.59	1.30	0.76	$.9767 \pm .0022$	$.9674 \pm .0023$
13	1.01	0.82	0.70	1.22	1.25	$.7980 \pm .0157$	$.8083 \pm .0091$

Table 7: **Performance using SW and enriched Pfam kernels.** The table contains experimental results using the enriched Pfam kernel and replacing the expression kernel with the SW kernel. A maximum of five kernel matrices would fit in the available memory, and the expresion kernel was removed because it received the smallest weights in the previous experiment (Table 6). See the caption to Table 4 for an explanation of individual columns. For comparison, mean ROC scores using the binary data are given in the final column of the table.

	Kernel weights					SDP/SVM ROC	
Class	K_{SW}	$K_{Pfam'}$	K_{TAP}	K_{Phys}	K_{Gen}	Enriched	Binary
1	2.33	0.82	0.19	0.61	1.06	$.9022 \pm .0024$	$.8825 \pm .0042$
2	2.65	0.19	0.12	0.72	1.32	$.8665 \pm .0081$	$.8563 \pm .0100$
3	2.58	0.24	0.11	0.62	1.46	$.8572 \pm .0040$	$.8464 \pm .0053$
4	2.15	0.74	0.15	0.85	1.12	$.9203 \pm .0039$	$.9024 \pm .0028$
5	2.22	1.10	0.04	0.70	0.94	$.9347 \pm .0046$	$.9094 \pm .0050$
6	2.40	0.67	0.11	0.68	1.14	$.8895 \pm .0047$	$.8742 \pm .0049$
7	1.80	0.40	0.56	1.22	1.02	$.9193 \pm .0043$	$.9149 \pm .0040$
8	2.64	0.02	0.35	0.71	1.29	$.8419 \pm .0086$	$.8023 \pm .0094$
9	1.81	0.44	0.95	0.92	0.88	$.8629 \pm .0073$	$.8623 \pm .0091$
10	2.81	0.01	0.40	0.93	0.85	$.8289 \pm .0066$	$.8120 \pm .0078$
11	2.81	0.01	0.40	0.79	1.00	$.6953 \pm .0109$	$.6575 \pm .0093$
12	1.72	1.58	0.49	1.01	0.21	$.9733 \pm .0029$	$.9674 \pm .0023$
13	2.93	0.01	0.76	0.64	0.66	$.8176 \pm .0173$	$.8083 \pm .0091$

Table 8: **Percentage true positives at one percent false positives.** The table contains experimental results for SDP/SVM when using only kernels on binary data, as well as for using the enriched expression and Pfam kernel and for using the enriched Pfam kernel and replacing the expression kernel with the SW kernel. Those results are compared with the MRF results. For this performance measure, the SDP/SVM method outperforms the MRF method, except for classes that have very few positive examples (classes 2, 8, 9, 11, 13).

Class	MRF	SDP/SVM		
		binary	enriched	SW-enriched
1	12.93%	28.03%	32.61%	38.53%
2	31.97%	29.25%	37.69%	44.68%
3	12.16%	31.95%	32.88%	35.53%
4	15.68%	40.97%	34.10%	37.52%
5	41.65%	51.93%	70.80%	64.06%
6	16.45%	34.18%	39.17%	42.96%
7	16.99%	47.57%	50.77%	53.08%
8	29.78%	25.68%	26.58%	39.86%
9	31.44%	35.15%	36.84%	40.08%
10	10.50%	25.12%	23.97%	29.70%
11	7.59%	9.45%	9.89%	8.57%
12	65.74%	78.40%	76.44%	80.74%
13	28.38%	18.75%	20.36%	31.43%