Noise tolerance of Multiple Classifier Systems in data integration-based gene function prediction: Supplementary Information

Matteo Rè and Giorgio Valentini

Dipartimento di Scienze dell'Informazione, Università degli Studi di Milano
v. Comelico 39 Milano, Italy, `http://www.dsi.unimi.it`

The supplementary material provides additional information about the data sets used in the experiments and detailed gene function prediction results relative to the 56 GO terms and 15 FunCat classes considered in two sets of experiments.

1 Data sets

In the second set of experiments, we predicted the top-level 15 functional classes of the FunCat taxonomy [4] of the model organism *S. cerevisiae*, using 6 different sources of data (Tab. 1). We considered yeast genes common to all data sets (about 1900) and with at least 1 FunCat annotation. We also removed from the list of the target functional classes all those represented by less than 20 genes. This corresponds to restrict our classifications to only 15 FunCat classes (Tab. 2) In other words, we selected the roots of the trees of the FunCat forest (that is the most general and wide functional classes of the overall FunCat taxonomy). Table 2 provides a brief description of the FunCat classes predicted in the experiments. The first column corresponds to the FunCat Identifier of the functional class.

2 Supplementary results

Performance relative to the first set of experiments in presence of artificial noise are reported in Tab. 3. The first three columns are dedicated to the Gene Ontology term ID, the ontology type of the GO term and the number of examples associated to the functional term respectively. Columns 4, 6 and 8 report the averaged AUCs obtained in the 15 folds by integrating the structured, mismatch kernels and a single noisy kernel using the average kernel fusion, the weighted average and decision templates integration methods respectively. Columns 5, 7 and 9 contains the results obtained in the same test repeated by including the second noisy kernel during data integration. Standard deviation results for each classification task are reported in parenthesis.

Table 4 shows the detailed results relative to the 15 FunCat classes considered in the second set of experiments. Results for both single SVMs trained on single sources of data, and kernel fusion (KF), weighted linear and decision template ensembles, with and without noisy data are provided. Note that SVMs trained with a single source of data achieve AUC scores only slightly larger than random guessing. Data integration methods always outperform the best single source classifiers (best results for each FunCat class are highlighted in boldface). Weighted linear and decision templates do not significantly worsen their performances with noisy data, while kernel fusion data integration undergoes a certain decrement of AUC scores.

Table 1: Datasets

Code	Dataset	n.genes	n.features	description
K_{string}	PPI - STRING	2338	2559	protein-protein interaction data from [6]
K_{BG}	PPI - BioGRID	4531	5367	protein-protein interaction data from the *BioGRID* database [5]
K_{pfam1}	Protein domain log-E	3529	5724	Pfam protein domains with log E-values computed by the *HMMER* software toolkit [1]
K_{pfam2}	Protein domain binary	3529	4950	protein domains obtained from *Pfam* database [2]
K_{expr}	Gene expression	4532	250	merged data of Spellman and Gasch experiments [3]
K_{seq}	Pairwise similarity	3527	6349	Smith and Waterman log-E values between all pairs of yeast sequences

Table 2: FunCat classes

Code	**Description**
01	Metabolism
02	Energy
10	Cell cycle and DNA processing
11	Transcription
12	Protein synthesis
14	Protein fate
16	Protein with binding function or cofactor requirement
18	Regulation of metabolism and protein function
20	Cellular transport and transport routes
30	Cellular communication/Signal transduction mechanism
32	Cell rescue, defense and virulence
34	Interaction with the environment
40	Cell fate
42	Biogenesis of cellular components
43	Cell type differentiation

Table 3: Performances of data integration methods in presence of one and two synthetic noisy kernels (AUC averaged across the 15 folds). KF-noisy1 and KF-noisy2 stands for kernel fusion methods with 1 and 2 noisy kernels; Wens-noisy1 and Wens-noisy2 stands for weighted average ensembles with 1 and 2 noisy kernels; DT-noisy1 and DT-noisy2 for decision templates with 1 and 2 noisy kernels.

GO_{term}	Ont	#	KF-noisy1	KF-noisy2	Wens-noisy1	Wens-noisy2	DT-noisy1	DT-noisy2
GO:0008168	MF	108	0.922(0.035)	0.908(0.040)	0.935(0.025)	0.935(0.025)	0.935(0.026)	0.934(0.026)
GO:0005506	MF	129	0.901(0.040)	0.873(0.045)	0.941(0.031)	0.941(0.031)	0.940(0.030)	0.940(0.029)
GO:0006260	BP	109	0.819(0.060)	0.749(0.050)	0.870(0.066)	0.869(0.065)	0.871(0.067)	0.869(0.069)
GO:0048037	MF	118	0.892(0.037)	0.860(0.038)	0.901(0.042)	0.899(0.042)	0.900(0.041)	0.899(0.042)
GO:0046483	BP	128	0.927(0.025)	0.917(0.027)	0.954(0.016)	0.953(0.016)	0.953(0.016)	0953(0.016)
GO:0044255	BP	101	0.818(0.051)	0.757(0.049)	0.882(0.059)	0.881(0.059)	0.881(0.058)	0.882(0.058)
GO:0016853	MF	124	0.773(0.073)	0.749(0.069)	0.842(0.072)	0.840(0.072)	0.835(0.073)	0.837(0.073)
GO:0044262	BP	209	0.883(0.032)	0.856(0.029)	0.900(0.031)	0.899(0.031)	0.899(0.030)	0.900(0.030)
GO:0009117	BP	124	0.825(0.058)	0.783(0.070)	0.877(0.045)	0.877(0.045)	0.877(0.044)	0.877(0.046)
GO:0016829	MF	201	0.901(0.022)	0.879(0.022)	0.931(0.021)	0.931(0.021)	0.932(0.021)	0.932(0.021)
GO:0016779	MF	142	0.830(0.077)	0.800(0.075)	0.863(0.075)	0.863(0.075)	0.862(0.074)	0.862(0.073)
GO:0016043	BP	106	0.779(0.082)	0.723(0.083)	0.874(0.052)	0.872(0.055)	0.876(0.050)	0.878(0.049)
GO:0008270	MF	234	0.862(0.028)	0.833(0.035)	0.889(0.026)	0.889(0.026)	0.888(0.026)	0.888(0.026)
GO:0006066	BP	111	0.917(0.034)	0.887(0.027)	0.933(0.033)	0.933(0.033)	0.932(0.034)	0.930(0.033)
GO:0003723	MF	212	0.847(0.036)	0.797(0.047)	0.895(0.037)	0.894(0.038)	0.893(0.038)	0.893(0.039)
GO:0004518	MF	125	0.786(0.073)	0.761(0.076)	0.813(0.058)	0.816(0.058)	0.810(0.056)	0.813(0.057)
GO:0006811	BP	117	0.715(0.060)	0.722(0.078)	0.762(0.087)	0.762(0.081)	0.763(0.085)	0.764(0.081)
GO:0006725	BP	164	0.858(0.040)	0.835(0.037)	0.894(0.050)	0.894(0.050)	0.887(0.051)	0.887(0.050)
GO:0016491	MF	516	0.938(0.013)	0.925(0.013)	0.946(0.009)	0.946(0.009)	0.946(0.008)	0.946(0.008)
GO:0009405	BP	118	0.925(0.044)	0.909(0.046)	0.942(0.040)	0.942(0.040)	0.943(0.040)	0.943(0.040)
GO:0005524	MF	485	0.875(0.032)	0.856(0.032)	0.892(0.026)	0.893(0.025)	0.891(0.027)	0.891(0.027)
GO:0030246	MF	102	0.904(0.032)	0.887(0.034)	0.919(0.032)	0.918(0.031)	0.918(0.033)	0.918(0.032)
GO:0006508	BP	330	0.924(0.019)	0.904(0.023)	0.934(0.015)	0.934(0.015)	0.933(0.015)	0.933(0.015)
GO:0008652	BP	121	0.912(0.034)	0.905(0.034)	0.922(0.035)	0.921(0.034)	0.920(0.034)	0.920(0.034)
GO:0045184	BP	108	0.789(0.063)	0.765(0.059)	0.849(0.041)	0.850(0.038)	0.849(0.041)	0.849(0.042)
GO:0020037	MF	104	0.987(0.016)	0.986(0.013)	0.993(0.013)	0.994(0.012)	0.993(0.014)	0.993(0.013)
GO:0003700	MF	214	0.907(0.020)	0.886(0.021)	0.931(0.022)	0.931(0.022)	0.931(0.022)	0.931(0.021)
GO:0016070	BP	140	0.868(0.058)	0.839(0.059)	0.914(0.039)	0.913(0.039)	0.915(0.038)	0.915(0.039)
GO:0005102	MF	120	0.917(0.046)	0.902(0.047)	0.930(0.043)	0.928(0.042)	0.929(0.042)	0.930(0.041)
GO:0006355	BP	340	0.908(0.017)	0.891(0.023)	0.926(0.021)	0.926(0.020)	0.925(0.022)	0.925(0.021)
GO:0016874	MF	161	0.872(0.039)	0.838(0.040)	0.885(0.035)	0.885(0.036)	0.884(0.034)	0.883(0.034)
GO:0006468	BP	160	0.860(0.056)	0.844(0.063)	0.892(0.056)	0.892(0.057)	0.893(0.055)	0.893(0.055)
GO:0016798	MF	227	0.969(0.021)	0.955(0.028)	0.967(0.017)	0.966(0.017)	0.967(0.017)	0.967(0.017)
GO:0006118	BP	392	0.937(0.013)	0.919(0.017)	0.961(0.012)	0.960(0.012)	0.961(0.012)	0.961(0.012)
GO:0004672	MF	164	0.873(0.063)	0.856(0.050)	0.900(0.048)	0.899(0.048)	0.900(0.046)	0.900(0.046)
GO:0004872	MF	138	0.904(0.047)	0.895(0.051)	0.933(0.030)	0.932(0.033)	0.933(0.029)	0.931(0.030)
GO:0015075	MF	110	0.755(0.072)	0.757(0.055)	0.815(0.059)	0.808(0.062)	0.814(0.059)	0.813(0.061)
GO:0005489	MF	196	0.935(0.024)	0.915(0.024)	0.965(0.021)	0.965(0.021)	0.965(0.021)	0.965(0.021)
GO:0005576	CC	352	0.876(0.031)	0.862(0.032)	0.896(0.022)	0.896(0.022)	0.895(0.020)	0.895(0.020)
GO:0019012	CC	101	0.841(0.064)	0.813(0.078)	0.879(0.056)	0.879(0.056)	0.878(0.054)	0.877(0.055)
GO:0030234	MF	132	0.808(0.059)	0.773(0.062)	0.830(0.052)	0.829(0.053)	0.827(0.057)	0.828(0.056)
GO:0016021	CC	136	0.698(0.054)	0.675(0.049)	0.789(0.054)	0.786(0.051)	0.787(0.051)	0.787(0.051)
GO:0006412	BP	170	0.865(0.038)	0.825(0.043)	0.910(0.037)	0.910(0.037)	0.910(0.036)	0.910(0.037)
GO:0005634	CC	347	0.901(0.026)	0.882(0.027)	0.927(0.026)	0.927(0.026)	0.927(0.026)	0.927(0.026)
GO:0017111	MF	154	0.796(0.056)	0.780(0.066)	0.866(0.049)	0.865(0.050)	0.866(0.048)	0.867(0.047)
GO:0005737	CC	490	0.877(0.027)	0.849(0.030)	0.903(0.024)	0.903(0.024)	0.901(0.023)	0.901(0.023)
GO:0051188	BP	118	0.784(0.054)	0.763(0.066)	0.817(0.051)	0.817(0.051)	0.814(0.052)	0.814(0.051)
GO:0043232	CC	153	0.804(0.060)	0.781(0.060)	0.842(0.050)	0.844(0.048)	0.842(0.050)	0.843(0.050)
GO:0043234	CC	414	0.787(0.028)	0.764(0.027)	0.846(0.025)	0.844(0.025)	0.843(0.025)	0.842(0.025)
GO:0005509	MF	173	0.901(0.033)	0.886(0.037)	0.920(0.021)	0.918(0.021)	0.918(0.020)	0.918(0.020)
GO:0050874	BP	144	0.883(0.051)	0.863(0.056)	0.900(0.044)	0.900(0.044)	0.902(0.041)	0.901(0.042)
GO:0006732	BP	119	0.788(0.058)	0.752(0.070)	0.850(0.050)	0.849(0.050)	0.841(0.052)	0.841(0.051)
GO:0007242	BP	140	0.900(0.029)	0.883(0.041)	0.925(0.031)	0.925(0.031)	0.926(0.032)	0.926(0.032)
GO:0005525	MF	104	0.906(0.043)	0.912(0.043)	0.925(0.042)	0.931(0.045)	0.925(0.042)	0.928(0.044)
GO:0004252	MF	140	0.923(0.046)	0.915(0.055)	0.924(0.042)	0.926(0.042)	0.924(0.042)	0.924(0.041)
GO:0005198	MF	179	0.819(0.031)	0.811(0.035)	0.837(0.035)	0.839(0.036)	0.836(0.038)	0.836(0.038)

Table 4: Performances of SVMs trained on single sources of data compared with performances of data integration methods (AUC averaged across the 5 folds) with 6 no-noisy and 6 noisy data. First column: Identifiers starting with K correspond to SVMs trained on single sources of data. The subscripts correspond to those of the data sets of Table 1. KF stands for Kernel Fusion, $Wens$ for weighted linear ensembles and DT for Decision Templates. The ending "n" refers to the SVM or data integration method trained with the corresponding noisy data. The other columns show the AUC results for the 15 FunCat classes. FunCat classes are represented through their two-digits identifiers (see Table 2).

Methods	01	02	10	11	12	14	16	18
K_{BG}	0.7783	0.7432	0.8516	0.8717	0.8969	0.7738	0.7001	0.7739
K_{BGn}	0.5399	0.5546	0.5212	0.5142	0.5412	0.5271	0.5110	0.5791
K_{pfam1}	0.7731	0.7629	0.7430	0.7205	0.8104	0.7056	0.6730	0.7404
K_{pfam1n}	0.5196	0.5505	0.5257	0.5162	0.5283	0.5168	0.5187	0.5517
K_{pfam2}	0.7667	0.6920	0.6393	0.6625	0.7567	0.7295	0.6710	0.6096
K_{pfam2n}	0.5152	0.5125	0.5248	0.5167	0.5206	0.5241	0.5090	0.5174
K_{string}	0.7423	0.7558	0.7855	0.7957	0.8460	0.7403	0.6127	0.6319
$K_{stringn}$	0.5098	0.5291	0.5271	0.5224	0.5323	0.5282	0.5131	0.5540
K_{seq}	0.8105	0.7650	0.7548	0.7773	0.8207	0.7332	0.7192	0.8067
K_{seqn}	0.5184	0.5245	0.5344	0.5350	0.5241	0.5413	0.5210	0.5698
K_{expr}	0.8091	0.7787	0.7594	0.7928	0.8171	0.7327	0.7080	0.8206
K_{exprn}	0.5158	0.5232	0.5414	0.5132	0.5316	0.5228	0.5202	0.5643
KF	**0.8966**	0.8451	**0.8881**	**0.9193**	0.9268	0.8483	**0.8085**	0.8403
KFn	0.8725	0.8207	0.8793	0.9025	0.9103	0.8082	0.7727	0.7767
$Wens$	0.8898	0.8415	0.8689	0.9026	0.9233	0.8426	0.7935	0.8415
$Wensn$	0.8887	**0.8462**	0.8747	0.9040	**0.9280**	**0.8538**	0.7906	**0.8572**
DT	0.8817	0.8364	0.8679	0.8972	0.9209	0.8300	0.7754	0.8409
DTn	0.8810	0.8403	0.8739	0.9012	0.9246	0.8430	0.7735	0.8497

Methods	20	30	32	34	40	42	43
K_{BG}	0.8278	0.8393	0.7109	0.7981	0.8594	0.7822	0.8264
K_{BGn}	0.5332	0.5577	0.5278	0.5255	0.5299	0.5414	0.5199
K_{pfam1}	0.7431	0.7995	0.6151	0.6818	0.7206	0.6013	0.7669
K_{pfam1n}	0.5242	0.5584	0.5253	0.5383	0.5529	0.5266	0.5500
K_{pfam2}	0.7275	0.7940	0.6844	0.6414	0.6631	0.5899	0.6174
K_{pfam2n}	0.5156	0.5356	0.5186	0.5160	0.5324	0.5172	0.5310
K_{string}	0.7794	0.7118	0.6862	0.7592	0.7697	0.7099	0.7705
$K_{stringn}$	0.5082	0.5596	0.5381	0.5377	0.5566	0.5372	0.5523
K_{seq}	0.7770	0.8476	0.6669	0.7039	0.7574	0.6560	0.7930
K_{seqn}	0.5175	0.5557	0.5217	0.5282	0.5383	0.5135	0.5283
K_{expr}	0.7671	0.8492	0.6469	0.6982	0.7626	0.6521	0.7957
K_{exprn}	0.5317	0.5490	0.5328	0.5312	0.5595	0.5302	0.5398
KF	**0.9037**	0.8914	0.7718	**0.8464**	0.8614	**0.8229**	0.8769
KFn	0.8637	0.8637	0.7442	0.7942	0.8236	0.8024	0.8522
$Wens$	0.8906	**0.9059**	0.7708	0.8363	0.8671	0.7953	0.8792
$Wensn$	0.8847	0.9029	**0.7761**	0.8280	0.8655	0.7941	0.8763
DT	0.8859	0.9053	0.7509	0.8317	**0.8700**	0.7972	**0.8779**
DTn	0.8800	0.9037	0.7628	0.8200	0.8632	0.7956	0.8723

References

[1] SR Eddy. Profile hidden markov models. *Bioinformatics*, 14(9):755–763, 1998.

[2] R.D. Finn, J. Tate, J. Mistry, P.C. Coggill, J.S. Sammut, H.R. Hotz, G. Ceric, K. Forslund, S.R. Eddy, E.L. Sonnhammer, and A. Bateman. The Pfam protein families database. *Nucleic Acids Research*, 36:D281–D288, 2008.

[3] P. Gasch and M. Eisen. Exploring the conditional regulation of yeast gene expression through fuzzy k-means clustering. *Genome Biology*, 3:11, 2002.

[4] A. Ruepp, A. Zollner, D. Maier, K. Albermann, J. Hani, M. Mokrejs, I. Tetko, U. Guldener, G. Mannhaupt, M. Munsterkotter, and H.W. Mewes. The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. *Nucleic Acids Research*, 32(18):5539–5545, 2004.

[5] C. Stark, B. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers. BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.*, 34:D535–D539, 2006.

[6] C. vonMering et al. STRING: a database of predicted functional associations between proteins. *Nucleic Acids Research*, 31:258–261, 2003.