Supplementary Material
Proximity Measures for Clustering Gene Expression Microarray Data: a Validation Methodology and a Comparative Analysis

Pablo A. Jaskowiak, Ricardo J. G. B. Campello, Ivan G. Costa
pablo@icmc.usp.br, campello@icmc.usp.br, igcf@cin.ufpe.br

In this supplement we provide additional results that due to space constraints were not included in our manuscript. In the first part of the supplement, Section S1, we provide complete tabular results regarding ISA and IBSA. In this first part, concerning IBSA, only results for original data are presented, i.e., results for datasets without normalization. In the second part of the supplement, Section S2, we provide complete results concerning the application of IBSA for the *normalized* version of the time-course datasets. We also provide detail regarding the normalization procedure of such data.

S1 - Complete Results - ISA and IBSA

In the following we provide complete results (in tabular format) regarding *Intrinsic Separation Ability* (ISA) and *Intrinsic Biological Separation Ability* (IBSA). Please, note that such results are presented as boxplots in our manuscript (Figures 1 and 3 of our manuscript). Table 1 of this supplement provides results regarding Intrinsic Separation Ability (ISA) analysis, which was performed on cancer datasets (these are the results presented in the boxplots of Figure 1 from our manuscript). In Table 2 we provide results regarding statistical tests for both cDNA and Affymetrix data, concerning ISA on cancer datasets. Tables 3 and 4 provide complete results regarding Intrinsic Biological Separation Ability Analysis (IBSA), which was performed on time-course datasets (their original version, without normalization). Please, note that such results are also presented in our paper in the form of boxplots (Figure 3 of our manuscript).

Table 1: Intrinsic Separation Ability (ISA) - cDNA and Affymetrix datasets.

		PE	GK	KE	SP	RM	WGK	COS	EUC	MAN	SUP
cDNA	alizadeh-2000-v1	0,7	0,7	0,7	0,7	0.71	0,7	0,69	0,6	0,6	0.53
	alizadeh-2000-v2	0.86	0.86	0.86	0.86	0.86	0.86	0.86	0.86	0.86	0.63
	alizadeh-2000-v3	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0,81	0,81	0.63
	bittner-2000	0.71	0,65	0,65	0,65	0,68	0,63	0,65	0,58	0,56	0.54
	bredel-2005	0.75	0,74	0,74	0,74	0,74	0,74	0.75	0,65	0,65	0.61
	chen-2002	0.76	0,75	0,75	0,75	0.76	0.76	0.76	0,58	0,58	0.57
	garber-2001	0.69	0,66	0,66	0,66	0,68	0,66	0.69	0,6	0,6	0.60
	khan-2001	0.90	0,67	0,68	0,67	0,8	0,7	0.90	0,69	0,72	0.60
	lapointe-2004-v1	0,58	0.60	0.60	0.60	0,59	0,58	0,58	0,55	0.60	0.53
	lapointe-2004-v2	0,6	0.61	0.61	0.61	0.61	0,6	0,59	0,55	0,59	0.53
	liang-2005	0,57	0,56	0,56	0,56	0,56	0,56	0,57	0,62	0,59	0.66
	risinger-2003	0.72	0,68	0,68	0,68	0,7	0,68	0,61	0,55	0,59	0.53
	tomlins-2006	0.77	0,76	0,76	0,76	0.77	0,76	0.77	0,69	0,68	0.66
	tomlins-2006-v2	0.73	0,72	0,72	0,72	0.73	0,72	0.73	0,64	0,64	0.60
Affymetrix	armstrong-2002-v1	0,7	0.79	0.79	0.79	0,78	0.79	0,69	0,64	0,67	0.56
	armstrong-2002-v2	0,87	0.90	0,89	0,89	0,89	0.90	0,87	0,75	0,77	0.63
	bhattacharjee-2001	0,77	0,79	0.82	0,81	0,81	0,77	0,77	0,76	0,81	0.72
	chowdary-2006	0,94	0.95	0.95	0.95	0.95	0,94	0,92	0,71	0,78	0.66
	dyrskjot-2003	0,73	0,76	0,76	0,76	0,7	0.82	0,77	0.82	0,77	0.80
	golub-1999-v1	0.79	0,73	0,73	0,73	0,75	0,72	0.79	0,77	0,76	0.64
	golub-1999-v2	0,89	0,76	0,75	0,75	0,86	0,73	0,89	0,77	0,74	0.73
	gordon-2002	0,92	0,94	0,94	0,94	0.95	0,93	0,92	0,68	0,8	0.64
	laiho-2007	0,44	0,58	0,58	0,57	0,45	0,57	0,45	0,65	0.72	0.52
	nutt-2003-v1	0,57	0,59	0,59	0,59	0,59	0,58	0,56	0.71	0.71	0.65
	nutt-2003-v2	0,51	0,51	0,51	0,51	0,52	0,51	0,51	0.66	0,64	0.66
	nutt-2003-v3	0,5	0,54	0,52	0,52	0,52	0,55	0,48	0,89	0.91	0.70
	pomeroy-2002-v1	0.62	0,59	0,58	0,58	0.62	0,59	0,6	0,49	0,49	0.50
	pomeroy-2002-v2	0.87	0,85	0,85	0,85	0,85	0,77	0,86	0,76	0,75	0.73
	ramaswamy-2001	0,9	0.92	0,91	0.92	0.92	0,68	0,9	0,58	0,58	0.59
	shipp-2002-v1	0,62	0,63	0,63	0,63	0,67	0,57	0,64	0,48	0,46	0.49
	singh-2002	0,58	0,58	0,58	0,58	0.59	0,57	0,57	0,56	0,58	0.53
	su-2001	0,91	0,92	0,92	0,92	0.94	0,92	0,91	0,82	0,87	0.76
	west-2001	0,77	0,67	0,67	0,67	0,75	0,67	0,77	0,63	0,65	0.59
	yeoh-2002-v1	0,68	0,6	0,58	0,59	0,62	0,63	0,74	0.75	0,61	0.70
	yeoh-2002-v2	0,77	0,62	0,62	0,62	0,71	0,54	0.78	0,59	0,53	0.60
		PE	GK	KE	SP	RM	WGK	COS	EUC	MAN	SUP

Table 2: Statistical Test Summary - cDNA and Affymetrix - Cancer Datasets.

	PE	GK	KE	SP	RM	WGK	COS	EUC	MAN	SUP
PE	—									
GK		—								
KE			—							
SP				—						
RM					—					
WGK						—				
COS							—			
EUC	□				□			—		
MAN	□				□				—	
SUP	□	∗			⊠					—

Symbols in each cell denote that the measure in the column outperformed the one in the row regarding: ∗ Affymetrix, □ cDNA, ⊠ both.

Table 3: Intrinsic Biological Separation Ability (IBSA) — Molecular Function Ontology (MF) — Original Datasets.

	PE	GK	KE	SP	RM	WGK	COS	EUC	MAN	SUP	JK	STS	LSS	YR1	YS1	YRM1
alpha factor	0.69	0.74	0.73	0.71	0.70	0.69	0.69	0.66	0.63	0.66	0.69	0.66	0.69	0.70	0.72	0.70
cdc 15	0.70	0.71	0.72	0.71	0.70	0.70	0.70	0.67	0.66	0.67	0.70	0.68	0.68	0.70	0.70	0.70
cdc 28	0.70	0.70	0.73	0.69	0.68	0.69	0.70	0.67	0.64	0.66	0.69	0.67	0.68	0.67	0.68	0.67
elutriation	0.70	0.70	0.72	0.70	0.70	0.70	0.70	0.66	0.64	0.67	0.70	0.67	0.67	0.70	0.72	0.70
1mM menadione	0.68	0.74	0.71	0.70	0.65	0.70	0.69	0.59	0.58	0.65	0.69	0.63	0.64	0.69	0.74	0.67
1M sorbitol	0.70	0.70	0.69	0.69	0.64	0.70	0.70	0.64	0.61	0.66	0.68	0.69	0.60	0.70	0.70	0.68
1.5mM diamide	0.68	0.69	0.68	0.69	0.64	0.68	0.67	0.65	0.62	0.69	0.70	0.67	0.62	0.68	0.70	0.65
2.5mM DTT	0.71	0.70	0.69	0.70	0.66	0.71	0.71	0.69	0.65	0.67	0.71	0.67	0.65	0.70	0.73	0.70
constant 32nM H2O2	0.70	0.70	0.71	0.69	0.70	0.70	0.70	0.66	0.62	0.67	0.70	0.65	0.67	0.70	0.74	0.68
diauxic shift	0.69	0.66	0.66	0.67	0.65	0.69	0.69	0.66	0.66	0.69	0.71	0.71	0.64	0.69	0.70	0.70
complete DTT	0.71	0.69	0.68	0.70	0.65	0.71	0.71	0.69	0.68	0.70	0.71	0.69	0.64	0.71	0.72	0.71
heat shock 1	0.71	0.68	0.67	0.67	0.64	0.71	0.71	0.68	0.62	0.66	0.71	0.68	0.64	0.71	0.72	0.64
heat shock 2	0.71	0.69	0.68	0.69	0.64	0.71	0.71	0.69	0.66	0.70	0.71	0.69	0.63	0.71	0.72	0.70
nitrogen depletion	0.71	0.72	0.71	0.72	0.66	0.71	0.71	0.70	0.67	0.70	0.70	0.68	0.68	0.71	0.75	0.69
YPD 1	0.71	0.71	0.68	0.70	0.66	0.71	0.71	0.68	0.65	0.70	0.71	0.71	0.71	0.71	0.72	0.71
YPD 2	0.71	0.72	0.71	0.72	0.66	0.71	0.71	0.70	0.67	0.73	0.71	0.71	0.69	0.71	0.73	0.69
yeast sporulation	0.69	0.68	0.68	0.66	0.61	0.70	0.66	0.69	0.67	0.71	0.63	0.69	0.58	0.70	0.74	0.70
	PE	GK	KE	SP	RM	WGK	COS	EUC	MAN	SUP	JK	STS	LSS	YR1	YS1	YRM1

Table 4: Intrinsic Biological Separation Ability (IBSA) — Biological Process Ontology (BP) — Original Datasets.

	PE	GK	KE	SP	RM	WGK	COS	EUC	MAN	SUP	JK	STS	LSS	YR1	YS1	YRM1
alpha factor	0.66	0.68	0.64	0.63	0.65	0.64	0.66	0.60	0.58	0.58	0.66	0.58	0.61	0.61	0.61	0.60
cdc 15	0.58	0.65	0.65	0.63	0.61	0.61	0.58	0.65	0.63	0.58	0.61	0.60	0.61	0.64	0.65	0.64
cdc 28	0.54	0.6	0.63	0.57	0.54	0.54	0.54	0.57	0.58	0.58	0.54	0.53	0.61	0.56	0.57	0.56
elutriation	0.62	0.62	0.63	0.62	0.65	0.66	0.61	0.60	0.59	0.59	0.61	0.61	0.59	0.62	0.66	0.65
1mM menadione	0.60	0.68	0.66	0.65	0.60	0.63	0.66	0.58	0.55	0.59	0.62	0.61	0.54	0.63	0.64	0.61
1M sorbitol	0.60	0.63	0.62	0.63	0.60	0.60	0.63	0.60	0.56	0.60	0.61	0.62	0.57	0.60	0.65	0.64
1.5mM diamide	0.60	0.64	0.63	0.63	0.61	0.61	0.60	0.62	0.58	0.61	0.6	0.62	0.52	0.61	0.64	0.59
2.5mM DTT	0.64	0.65	0.64	0.62	0.60	0.64	0.69	0.67	0.63	0.64	0.64	0.60	0.58	0.64	0.64	0.64
constant 32nM H2O2	0.46	0.45	0.45	0.46	0.43	0.46	0.47	0.47	0.47	0.43	0.46	0.42	0.46	0.46	0.45	0.46
diauxic shift	0.66	0.64	0.63	0.63	0.61	0.63	0.66	0.62	0.63	0.65	0.66	0.62	0.62	0.66	0.66	0.64
complete DTT	0.64	0.63	0.62	0.62	0.61	0.64	0.63	0.65	0.63	0.63	0.65	0.67	0.53	0.64	0.65	0.65
heat shock 1	0.66	0.66	0.65	0.66	0.66	0.66	0.66	0.65	0.60	0.62	0.66	0.57	0.64	0.66	0.70	0.66
heat shock 2	0.63	0.64	0.64	0.65	0.62	0.63	0.66	0.64	0.62	0.64	0.64	0.61	0.59	0.63	0.65	0.65
nitrogen depletion	0.60	0.64	0.63	0.61	0.58	0.60	0.64	0.66	0.63	0.66	0.59	0.51	0.53	0.65	0.70	0.67
YPD 1	0.65	0.65	0.61	0.65	0.63	0.65	0.65	0.66	0.63	0.63	0.65	0.61	0.64	0.65	0.64	0.65
YPD 2	0.67	0.66	0.65	0.68	0.64	0.67	0.67	0.66	0.63	0.67	0.67	0.62	0.65	0.67	0.70	0.67
yeast sporulation	0.64	0.66	0.66	0.66	0.63	0.67	0.66	0.64	0.62	0.65	0.64	0.66	0.57	0.67	0.70	0.67
	PE	GK	KE	SP	RM	WGK	COS	EUC	MAN	SUP	JK	STS	LSS	YR1	YS1	YRM1

S2 - Complete Results - IBSA - Normalized Datasets

In this section we provide results concerning the comparison of proximity measures in the normalized version of the time-course datasets. Before presenting the results, we give details on the source of the original data and how we proceeded to normalize the datasets. In brief, all the datasets employed in our study are publicly available at the *Gene Expression Omnibus* (GEO) website in raw format [1], except from the dataset *cdc 28*, for which we were unable to obtain raw data. Since we were unable to obtain original data for the *cdc 28* dataset, our experiments on normalized data were performed on 16 datasets. Such datasets may be easily obtained from GEO by performing a search by the name of the authors of the study that generated the dataset in question, or even the title of their original paper. We also note, that the preprocessed

version of such datasets, which was employed in the experiments described in this section are publicly available as a benchmark set at our website: `http://www.icmc.usp.br/~campello/Sub_Pages/IEEEACM_TCBB_arquivos/`.

After obtaining the raw versions of the datasets which provide, among other information, the values of green and red intensities, background intensities, spot names and, spot locations, we were able to began with the normalization procedure of each dataset. As kindly suggested by one of the reviewers of our paper, in order to remove the problem of dynamic range of data represented by their log ratio values, we employed the procedure referred to as *Multiple-Slide Normalization*, as used in [2]. Normalization was carried out with the help of the `marray` package [3] from Bioconductor [4]. Please, note that the main purpose of our paper is to compare different proximity measures to both the clustering of cancer samples and the clustering of time-course data. The experiments considering the normalized version of the datasets were performed in order to verify if our results remain valid in this case. Having made such considerations, for further information on the Multiple-Slide Normalization procedure, please, refer to [2], or to the documentation of the `marray` package [3]. After performing normalization we removed genes for which 10% or more expression values were missing. After this removal no genes with missing values remained (note that the datasets have a few features, i.e., time points). Finally, for further analysis, we select genes that displayed a difference of at least l-fold in at least c samples from their mean expression level [5]. We consider $c = 1$ and adjust the value of l in order to select about 1000 genes, number previously employed in several works (e.g., [6, 7, 8, 9]). A summary of the 16 datasets, after normalization and filtering, is given in Table 5.

Table 5: Time-course datasets used in the experiments — Version for which we applied Multiple-Slide Normalization.

Name	Source	$\#s$	$\#\hat{g}$	$\#g$
alpha factor	Spellman *et al.* (1998)	16	6178	978
cdc 15		24	6178	1010
elutriation		14	6178	975
1mM menadione	Gasch *et al.* (2000)	9	6152	1033
1M sorbitol		6	6152	1057
1.5mM diamide		8	6152	995
2.5mM DTT		8	6152	1024
constant 32nM H2O2		9	6152	850
diauxic shift		7	6152	1012
complete DTT		7	6152	1023
heat shock 1		8	6152	991
heat shock 2		5	6152	1023
nitrogen depletion		9	6152	998
YPD 1		10	6152	997
YPD 2		9	6152	1001
yeast sporulation	Chu *et al.* (1998)	7	6118	982

Once again, we note that the preprocessed version of such datasets, which comprise normalization and the filtering of genes is publicly available for download at: `http://www.icmc.usp.br/~campello/Sub_Pages/IEEEACM_TCBB_arquivos/`.

S2.1 - Results - IBSA - Normalized Datasets

In the following we present the results for the normalized versions of the datasets from Table 5. Such results are not described in our paper due to space constraints. In Figure 1 we depict IBSA values for the distances under evaluation considering the normalized versions of the datasets (presented in Table 5). These results are also presented in tabular form on Tables 6 and 7. In brief, the results for the normalized datasets are in agreement with the ones obtained for the original datasets. YS1 presents the best results among the proximity measure under comparison, whereas the worst results are displayed by LSS. Although the differences among different distances seem smaller in the boxplots from Figure 1, they are still present. As shown by Table 8, there are statistical differences in favor of YS1 in a number of cases for both MS and BP ontologies. Finally, the consistency of the results, i.e., the agreement between the results for the original datasets and the normalized datasets, becomes clearer when the average ranks for each distance, regarding all datasets, are observed. To this extent, we depict in Table 9 the average ranks for each distance, considering the original and normalized version of the datasets. Note that there is moderate to high correlation between the ranks for original datasets and normalized datasets, for both BP and MF, indicating an agreement between such results. Finally, in Figure 2, we depict results regarding noise experiments in the normalized datasets (for methodological details on such experiments, please, refer to Section 4.3 of our paper). For *alpha factor* and *cdc 15* IBSA analysis was performed employing MF ontology, whereas for *1.5mM diamide* and *YPD1*, BP ontology was used. Such results are also in agreement with results for the original datasets.

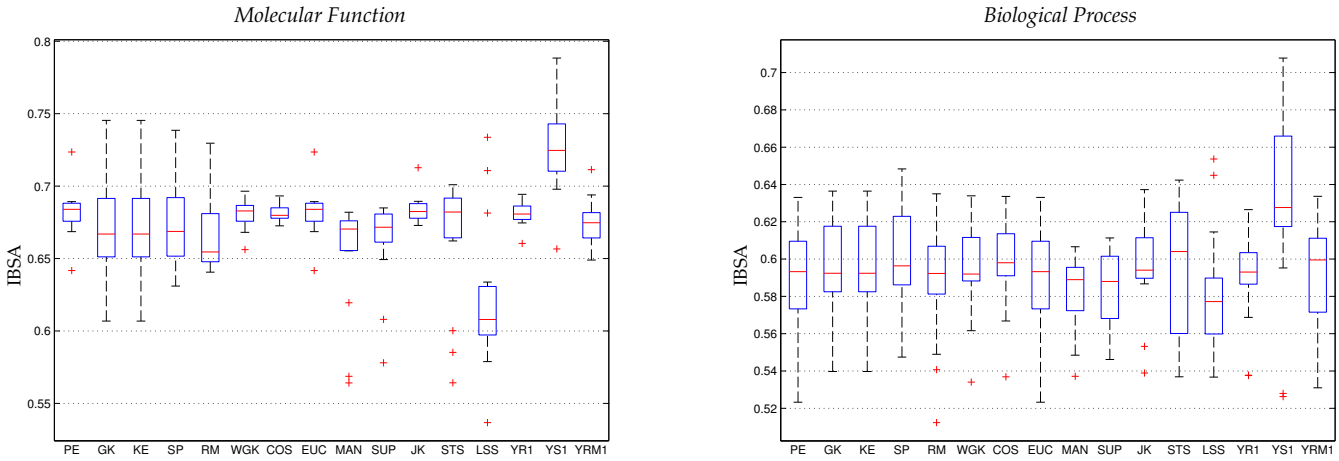

Figure 1: Boxplots depict IBSA values for the distances under evaluation considering the normalized datasets.

Table 6: Intrinsic Biological Separation Ability (IBSA) — Molecular Function Ontology (MF) — Original Datasets.

	PE	GK	KE	SP	RM	WGK	COS	EUC	MAN	SUP	JK	STS	LSS	YR1	YS1	YRM1
alpha factor	0.69	0.73	0.73	0.72	0.68	0.69	0.67	0.69	0.57	0.61	0.69	0.69	0.71	0.68	0.79	0.69
cdc 15	0.72	0.75	0.75	0.74	0.73	0.70	0.69	0.72	0.62	0.67	0.71	0.69	0.73	0.69	0.79	0.71
elutriation	0.69	0.72	0.72	0.70	0.69	0.69	0.69	0.69	0.68	0.68	0.69	0.70	0.68	0.69	0.78	0.69
1mM menadione	0.68	0.70	0.70	0.70	0.65	0.68	0.68	0.68	0.66	0.66	0.68	0.69	0.63	0.68	0.74	0.68
1M sorbitol	0.68	0.64	0.64	0.65	0.65	0.68	0.68	0.68	0.67	0.65	0.68	0.67	0.58	0.68	0.71	0.68
1.5mM diamide	0.68	0.66	0.66	0.66	0.65	0.67	0.68	0.68	0.66	0.67	0.67	0.68	0.60	0.67	0.72	0.66
2.5mM DTT	0.67	0.67	0.67	0.67	0.64	0.67	0.68	0.67	0.66	0.68	0.68	0.70	0.60	0.68	0.74	0.67
constant 32nM H2O2	0.64	0.68	0.68	0.68	0.68	0.66	0.67	0.64	0.56	0.58	0.68	0.59	0.63	0.66	0.73	0.65
diauxic shift	0.68	0.65	0.65	0.66	0.68	0.68	0.68	0.68	0.68	0.68	0.69	0.70	0.61	0.68	0.72	0.67
complete DTT	0.68	0.64	0.64	0.64	0.65	0.68	0.68	0.68	0.68	0.68	0.68	0.69	0.59	0.68	0.70	0.67
heat shock 1	0.69	0.67	0.67	0.67	0.66	0.69	0.69	0.69	0.67	0.67	0.69	0.68	0.61	0.69	0.73	0.68
heat shock 2	0.68	0.61	0.61	0.63	0.65	0.68	0.68	0.68	0.67	0.67	0.68	0.66	0.54	0.68	0.66	0.66
nitrogen depletion	0.67	0.67	0.67	0.67	0.65	0.67	0.67	0.67	0.66	0.66	0.67	0.56	0.63	0.68	0.72	0.67
YPD 1	0.60	0.58	0.58	0.58	0.60	0.6	0.60	0.60	0.59	0.60	0.6	0.54	0.58	0.60	0.63	0.60
YPD 2	0.69	0.68	0.68	0.67	0.64	0.68	0.68	0.69	0.68	0.68	0.68	0.67	0.61	0.69	0.71	0.66
yeast sporulation	0.69	0.65	0.65	0.65	0.67	0.70	0.69	0.69	0.67	0.67	0.69	0.69	0.59	0.69	0.71	0.69
	PE	GK	KE	SP	RM	WGK	COS	EUC	MAN	SUP	JK	STS	LSS	YR1	YS1	YRM1

Table 7: Intrinsic Biological Separation Ability (IBSA) — Biological Process Ontology (BP) — Normalized Datasets

	PE	GK	KE	SP	RM	WGK	COS	EUC	MAN	SUP	JK	STS	LSS	YR1	YS1	YRM1
alpha factor	0.63	0.63	0.63	0.65	0.63	0.63	0.63	0.63	0.55	0.56	0.64	0.64	0.65	0.6	0.71	0.56
cdc 15	0.56	0.60	0.60	0.6	0.57	0.59	0.61	0.56	0.57	0.55	0.59	0.64	0.64	0.57	0.66	0.57
elutriation	0.62	0.61	0.61	0.62	0.62	0.62	0.62	0.62	0.60	0.60	0.62	0.63	0.61	0.62	0.67	0.62
1mM menadione	0.61	0.64	0.64	0.63	0.59	0.62	0.63	0.61	0.60	0.60	0.61	0.64	0.59	0.63	0.67	0.61
1M sorbitol	0.59	0.58	0.58	0.59	0.59	0.59	0.59	0.59	0.57	0.58	0.59	0.58	0.55	0.59	0.62	0.60
1.5mM diamide	0.59	0.59	0.59	0.6	0.59	0.59	0.60	0.59	0.59	0.58	0.59	0.60	0.56	0.59	0.63	0.59
2.5mM DTT	0.61	0.62	0.62	0.62	0.61	0.61	0.61	0.61	0.59	0.61	0.61	0.62	0.59	0.61	0.66	0.63
constant 32nM H2O2	0.52	0.54	0.54	0.55	0.51	0.56	0.54	0.52	0.54	0.55	0.54	0.56	0.56	0.54	0.53	0.53
diauxic shift	0.59	0.59	0.59	0.60	0.59	0.59	0.59	0.59	0.59	0.59	0.59	0.62	0.56	0.59	0.63	0.61
complete DTT	0.60	0.59	0.59	0.60	0.60	0.6	0.60	0.60	0.60	0.60	0.6	0.61	0.57	0.60	0.63	0.61
heat shock 1	0.62	0.62	0.62	0.63	0.62	0.62	0.62	0.62	0.61	0.61	0.62	0.61	0.59	0.62	0.67	0.63
heat shock 2	0.59	0.57	0.57	0.59	0.59	0.59	0.59	0.59	0.59	0.58	0.59	0.59	0.54	0.59	0.60	0.60
nitrogen depletion	0.58	0.61	0.61	0.60	0.59	0.59	0.59	0.58	0.57	0.58	0.59	0.54	0.58	0.59	0.63	0.57
YPD 1	0.68	0.66	0.66	0.65	0.65	0.68	0.68	0.68	0.68	0.68	0.69	0.60	0.63	0.68	0.70	0.66
YPD 2	0.56	0.59	0.59	0.59	0.54	0.59	0.60	0.56	0.59	0.60	0.59	0.55	0.55	0.59	0.61	0.58
yeast sporulation	0.55	0.54	0.54	0.56	0.55	0.53	0.57	0.55	0.59	0.55	0.55	0.56	0.59	0.54	0.53	0.54
	PE	GK	KE	SP	RM	WGK	COS	EUC	MAN	SUP	JK	STS	LSS	YR1	YS1	YRM1

Table 8: Statistical Test Summary — MF and BP Ontologies — Normalized Datasets.

	PE	GK	KE	SP	RM	WGK	COS	EUC	MAN	SUP	JK	STS	LSS	YR1	YS1	YRM1
PE	—														□	
GK		—													⊛	
KE			—												⊛	
SP				—											*	
RM	*				—	*		*			*				⊛	
WGK						—										
COS							—								*	
EUC								—							□	
MAN				□					—			□			⊛	
SUP										—					⊛	
JK											—					
STS												—			*	
LSS	*			□		*	*	*			*		—	*	⊛	
YR1														—	⊛	
YS1															—	
YRM1															*	—

Symbols in each cell denote that the measure in the column outperformed the one in the row regarding: $*$ MF ontology, □ BP ontology, ⊛ both.

Table 9: Intrinsic Biological Separation Ability (IBSA) average ranks for Biological Process (BP) and Molecular Function (MF) ontologies regarding Original Datasets (OD) and Normalized Datasets (ND). In the last row we depict the Pearson Correlation Coefficient value between IBSA average ranks considering OD and ND for both ontologies.

	BP		MF	
	OD	ND	OD	ND
PE	8.58	9.43	5.52	6.62
GK	5.47	8.75	6.64	8.93
KE	7.82	8.75	7.52	8.93
SP	6.47	5.75	7.11	9.12
RM	12.0	9.37	12.64	13.0
WGK	7.47	8.50	5.82	6.62
COS	6.64	7.18	5.76	7.50
EUC	7.47	9.43	11.94	6.62
MAN	11.88	12.56	15.0	11.18
SUP	10.29	11.0	10.0	10.0
JK	8.17	7.31	5.79	6.43
STS	12.23	6.75	10.76	9.00
LSS	13.23	12.0	13.26	13.5
YR1	7.52	8.93	7.00	7.68
YS1	4.29	2.75	1.94	1.62
YRM1	6.41	7.50	9.23	9.18
Correlation	0.67		0.81	

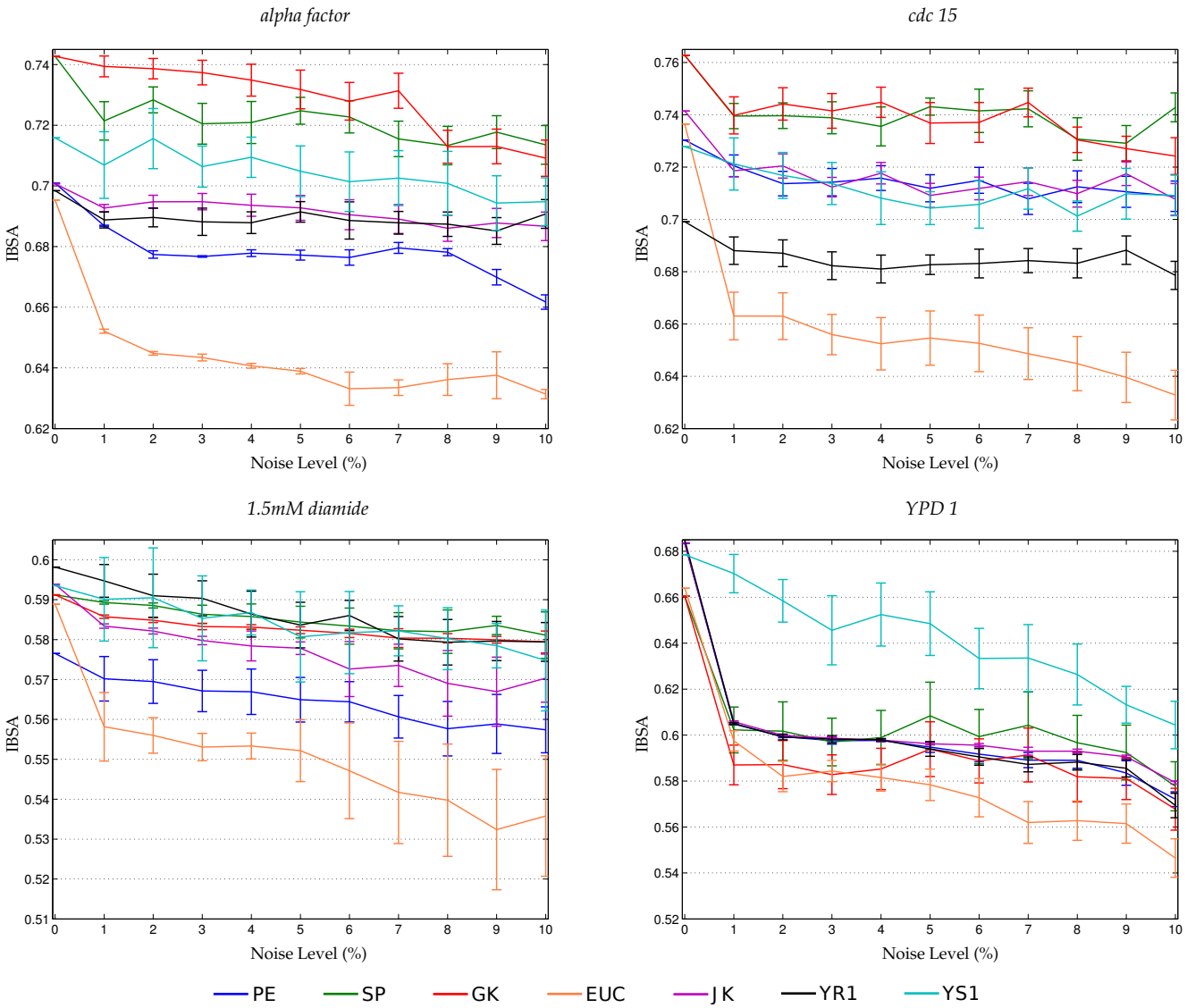

Figure 2: IBSA for different noise levels (%) regarding Pearson (PE), Spearman (SP), Goodman-Kruskal (GK), Euclidean distance (EUC), Jackknife (JK), YR1, and YS1 in four datasets. Lines correspond to mean IBSA values (bars account for standard deviations) for executions performed in 100 noisy datasets for each noise level between 1% and 10%.

Bibliography

[1] R. Edgar, M. Domrachev, and A. E. Lash, "Gene expression omnibus: Ncbi gene expression and hybridization array data repository," *Nucleic Acids Research*, vol. 30, no. 1, pp. 207–210, 2002.

[2] L. P. S. T. Yang YH, Dudoit S, "Normalization for cdna microarray data," in *Microarrays: Optical Technologies and Informatics. SPIE BIOS*, vol. 4266, 2001, pp. 141–152.

[3] Y. H. Y. with contributions from Agnes Paquet and S. Dudoit., *marray: Exploratory analysis for two-color spotted microarray data*, 2009, r package version 1.36.0. [Online]. Available: http://www.maths.usyd.edu.au/u/jeany/

[4] R. Gentleman, V. Carey, D. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Yang, and J. Zhang, "Bioconductor: open software development for computational biology and bioinformatics," *Genome Biology*, vol. 5, no. 10, pp. R80+, 2004.

[5] K. Faceli, A. A. C. d. Carvalho, and W. A. Silva Jr, "Evaluation of gene selection metrics for tumor cell classification," *Genetics and Molecular Biology*, vol. 27, pp. 651 – 657, 00 2004.

[6] P. Tamayo *et al.*, "Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation," *Proc Natl Acad Sci U S A*, vol. 96, no. 6, pp. 2907–2912, 1999.

[7] A. P. Gasch *et al.*, "Genomic expression programs in the response of yeast cells to environmental changes," *Molecular Biology of the Cell*, vol. 11, no. 12, pp. 4241–4257, 2000.

[8] Z. S. Qin, "Clustering microarray gene expression data using weighted chinese restaurant process," *Bioinformatics*, vol. 22, no. 16, pp. 1988–1997, 2006.

[9] M. Souto, I. Costa, D. de Araujo, T. Ludermir, and A. Schliep, "Clustering cancer gene expression data: a comparative study," *BMC Bioinformatics*, vol. 9, no. 1, p. 497, 2008.