BIOINFORMATICS

Vol. 00 no. 00 2009
Pages 1–8

A hypergraph-based learning algorithm for classifying gene expression and arrayCGH data with prior knowledge (supplementary document)

Ze Tian[*], TaeHyun Hwang[*], and Rui Kuang[†]

Department of Computer Science and Engineering, University of Minnesota Twin Cities,
Minneapolis, Minnesota, USA

1 ALGORITHM AND FRAMEWORK

1.1 Description of the algorithm

The *HyperPrior* algorithm first initializes w with a uniform weighting 1 over the hyperedges. Note that $w = 1$ is a solution to the linear system $Hw = diag(D_v)$ by definition of D_v and thus, a valid solution to $\underset{f,w}{\text{minimize}}\ \Phi(f, w)$. In the first step in each iteration, *HyperPrior* fixes w and optimizes $\Phi(f, w = w_t)$ with respect to f in the following optimization problem,

$$\underset{f}{\text{minimize}} \quad \Omega(f, w = w_t) + \mu||f - y||^2 \tag{1}$$

The cost term $\Psi(w = w_t)$ is removed from $\Phi(f, w = w_t)$ since it is a constant in the above optimization problem. Let $L = I - D_v^{-1/2} H W D_e^{-1} H^T D_v^{-1/2}$. In the cost term, we can prove $\Omega(f, w = w_t) = f^T L f$ (see next section). L is positive semi-definite given $\Omega(f, w = w_t) \geq 0$ for any f, which also implies that $\Omega(f, w = w_t)$ is convex in f. Therefore, we can simply take derivative with respect to f to get the optimal solution $f^* = (1 - \alpha)((1 - \alpha)I + \alpha L)^{-1} y$, where $\alpha = \frac{\mu}{1+\mu}$ (Zhou *et al.*, 2006). This is equivalent to solving the linear system $(1 - \alpha)((1 - \alpha)I + \alpha L)f = y$.

In the second step in each iteration, the *HyperPrior* algorithm fixes $f = f_t$ learned in the previous step to learn the optimal weighting of hyperedges w by solving the quadratic programming problem:

$$\underset{w}{\text{minimize}} \quad \Omega(f = f_t, w) + \rho \Psi(w) \tag{2}$$

subject to

$$\begin{array}{ll} w(e) \geq 0 & \text{for } \forall e \in E \\ \sum_{e \in E} h(v, e) w(e) = d(v) & \text{for } \forall v \in V. \end{array}$$

The cost $\mu||f - y||^2$ is removed from $\Phi(f, w = w_t)$ since it is a constant in the above optimization problem, and $\Omega(f = f_t, w)$ is a linear function of w. Since $\Psi(w) = w^T (I - D_\sigma^{-1/2} \Delta D_\sigma^{-1/2}) w \geq 0$ for any w, $I - D_\sigma^{-1/2} \Delta D_\sigma^{-1/2}$ is positive semi-definite, which implies that $\Phi(f = f_t, w)$ is convex in w. In both steps, the total cost $\Phi(f, w)$ is guaranteed to be reduced until there is only very small change. Thus, our algorithm will finally stop at a small total cost. We implemented the *HyperPrior* algorithm in MATLAB and use ILOG/CPLEX package (version 11.1) for quadratic programming.

[*] equal contribution

[†] to whom correspondence should be addressed

```
HyperPrior$(y, H, \Delta, \mu, \rho)$

$t = 0, w_0 = 1, f_0 = y, c_0 = +\infty$
do
  $t = t + 1$
  Use network propagation to find optimal $f_t$
    $f_t = (1-\alpha)(I - \alpha D_v^{-1/2} H W_{t-1} D_e^{-1} H^T D_v^{-1/2})^{-1} y$
  Use quadratic programming to find optimal $w_t$
    $w_t = \text{argmin}_w\ \Omega(f = f_{t-1}, w) + \rho\Psi(w)$
  subject to $Hw = diag(D_v)$ and $diag(W) \succeq 0$
  $c_t = \Omega(f_t, w_t) + \mu||f_t - y||^2 + \rho\Psi(w_t)$
while $(c_{t-1} - c_t > \pi)$
return $(f_t, w_t)$
```

Fig. 1. The *HyperPrior* algorithm.

1.2 Proof of convexity

Let $L = I - D_v^{-1/2} H W D_e^{-1} H^T D_v^{-1/2}$, where I is the identity matrix and W is the diagonal matrix with $W_{ii} = w(e_i)$. We can show $\Omega(f, w) = f^T L f$ by

$$\begin{aligned}
\Omega(f,w) &= \sum_{e \in E} \sum_{u,v \in V} \frac{w(e)h(u,e)h(v,e)}{d(e)} \left(\frac{f^2(u)}{d(u)} - \frac{f(u)f(v)}{\sqrt{d(u)d(v)}}\right) \\
&= \sum_{e \in E} \sum_{u \in V} \frac{w(e)h(u,e)f^2(u)}{d(u)} \sum_{v \in V} \frac{h(v,e)}{d(e)} - \sum_{e \in E} \sum_{u,v \in V} \frac{w(e)h(u,e)h(v,e)}{d(e)} \frac{f(u)f(v)}{\sqrt{d(u)d(v)}} \\
&= \sum_{u \in V} f^2(u) \sum_{e \in E} \frac{w(e)h(u,e)}{d(u)} - \sum_{e \in E} \sum_{u,v \in V} \frac{f(u)w(e)h(u,e)h(v,e)f(v)}{\sqrt{d(u)d(v)}d(e)} \\
&= \sum_{u \in V} f^2(u) - \sum_{e \in E} \sum_{u,v \in V} \frac{f(u)w(e)h(u,e)h(v,e)f(v)}{\sqrt{d(u)d(v)}d(e)}.
\end{aligned}$$

Step three in the above derivation shows that $\Omega(f, w) = f^T L f$ if and only if $\sum_{e \in E} \frac{w(e)h(u,e)}{d(u)} = 1$. The constraints $\sum_{e \in E} h(v, e)w(e) = d(v)$ for $\forall v \in V$ keep D_v unchanged during the optimization and thus make L always positive semi-definite.

1.3 Convergency of the algorithm

To check the convergence of the *HyperPrior* algorithm, we measured the value of the cost function in each iteration on the real microarray gene expression datasets with selected 1,464 genes. The change of the cost function for different α and ρ parameters is shown in Fig. 2. It is clear that the *HyperPrior* algorithm converges very fast. We also found that the value of f and w variables stay unchanged after the first 2 to 3 iterations.

In *HyperPrior*, we use $w = 1$ as the starting point. However, we also tried random starting points and they all converged to the same solution as long as the initial constraints on w are satisfied. So empirically, the solution of *HyperPrior* is not affected by the starting point.

2 CLASSIFICATION RESULTS

2.1 Parameter tuning for ArrayCGH data

We tested *HyperPrior* on two arrayCGH datasets used by Rapaport *et al.* (2008). By following the same procedure from that paper, we made three classification problems and performed a cross-validation by a leave-one-out (LOO) procedure for them. For the SVMs with linear and RBF kernels, combinations of parameters $C = \{10^{-5}, 10^{-4}, \ldots, 10^4, 10^5\}$ and $\sigma = \{10^{-5}, 10^{-4}, \ldots, 10^4, 10^5\}$ were tested. For the hypergraph-based algorithm and *HyperPrior*, parameters $\alpha = \{0.01, 0.1, 0.3, 0.5, 0.7, 0.9, 0.99\}$, and $\rho = \{10^{-3}, 10^{-2}, \ldots, 10^2, 10^3\}$ (for *HyperPrior* only) were tested. The L_1-SVM and fused SVM were implemented using the source code provided by Rapaport *et al.* (2008). For L_1-SVM and fused SVM, combinations of parameters $\lambda = \{2^0, 2^1, \ldots, 2^9, 2^{10}\}$, and $\mu = \{2^{-10}, 2^{-9}, \ldots, 2^9, 2^{10}\}$ (for fused SVM only) were tested.

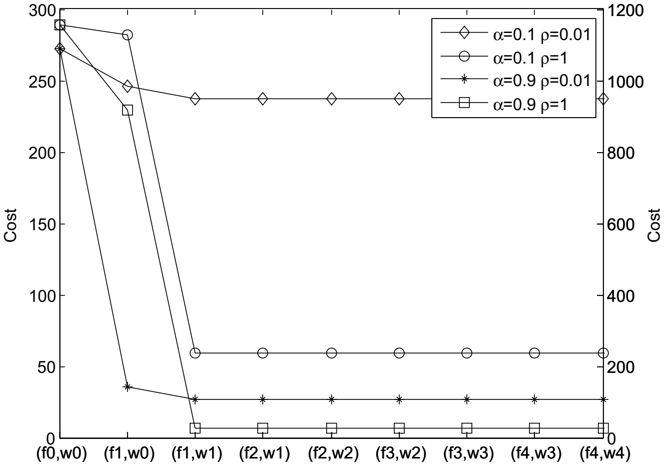


Fig. 2. Convergence of *HyperPrior*. This plot shows the decrease of the cost function after each iteration of *HyperPrior*.

	σ/C	0.0001	0.001	0.01	0.1	1	10	100	1000	10000
SVM (linear)		0.218	**0.205**	0.244	0.244	0.231	0.244	0.244	0.244	0.244
SVM (RBF)	10	0.218	0.218	0.218	0.218	0.218	0.218	0.231	0.231	0.231
	100	0.218	0.218	0.218	0.218	0.218	**0.205**	0.244	0.244	0.231
	1000	0.218	0.218	0.218	0.218	0.218	0.218	0.218	**0.205**	0.244
	10000	0.218	0.218	0.218	0.218	0.218	0.218	0.218	0.218	0.218

Table 1. SVMs on 78 training samples from van 't Veer *et al.* dataset with 231 genes

C/percentage	0.2	0.4	0.6	0.8	1
0.0001	0.231	0.218	0.231	0.231	0.231
0.001	0.244	0.231	0.269	0.231	0.218
0.01	0.218	0.269	0.282	0.269	0.256
0.1	0.231	0.269	0.218	0.218	0.244
1	0.231	0.256	0.218	**0.192**	0.231
10	0.244	0.269	0.218	**0.192**	0.231
100	0.256	0.269	0.218	**0.192**	0.231
1000	0.256	0.269	0.218	**0.192**	0.231
10000	0.256	0.269	0.218	**0.192**	0.231

Table 2. Rapaport *et al.*'s method on 78 training samples from van 't Veer *et al.* dataset with 231 genes

2.2 Gene expression data

We evaluated *HyperPrior* on two breast cancer gene expression datasets, the van 't Veer *et al.* dataset with 97 samples (van 't Veer *et al.*, 2002) and the van de Vijver *et al.* dataset with 295 samples (van de Vijver *et al.*, 2002), using as a prior a large curated human protein-protein interaction network with 57,235 interactions, which is integrated from yeast two-hybrid experiments, predicted interactions from orthology and co-citatioin, and other literature reviews (Chuang *et al.*, 2007). The classification task is to classify patients who developed metastasis or were free of metastasis in five years after prognosis.

2.2.1 Parameter tuning for van 't Veer et al. dataset As suggested by van 't Veer *et al.* (2002), 231 genes are selected on a training set of 78 patients and the remaining 19 patients are held out as the test set in the van 't Veer *et al.* dataset. To select the parameters used on test set, we performed a leave-one-out cross-validation on 78 training samples and report the training error rate for each algorithm in Table 1,2,3 and 4.

2.2.2 5-fold cross-validation for van de Vijver et al. dataset In the experiments on van de Vijver *et al.* dataset, we used for classification two subsets of hypothetical cancer susceptibility genes: 326 genes from *Ingenuity* and 1,464 genes from Cancer Genomics tool (`http://cbio.mskcc.org/CancerGenes/Select.action`). We randomly run 5-fold cross-validation multiple times on van de Vijver *et al.* dataset and measure the average AUC. Note that within each experiment of a 5-fold cross-validation, another 4-fold cross-validation is

λ_1/λ_2	0.0001	0.001	0.01	0.1	1	10	100	1000	10000
0.0001	0.256	0.256	0.269	0.269	0.269	0.269	0.269	0.269	0.269
0.001	0.256	0.256	0.256	0.269	0.269	0.269	0.269	0.269	0.269
0.01	0.269	0.256	0.282	0.269	0.269	0.269	0.269	0.244	0.295
0.1	**0.231**	**0.231**	0.244	0.282	0.308	0.321	0.295	0.282	0.269
1	0.346	0.346	0.346	0.372	0.359	0.333	0.308	0.256	**0.231**
10	0.295	0.295	0.295	0.295	0.295	0.282	0.282	0.282	0.282
100	0.564	0.564	0.564	0.564	0.564	0.564	0.564	0.564	0.564
1000	0.564	0.564	0.564	0.564	0.564	0.564	0.564	0.564	0.564
10000	0.564	0.564	0.564	0.564	0.564	0.564	0.564	0.564	0.564

Table 3. Li *et al.*'s method on 78 training samples from van 't Veer *et al.* dataset with 231 genes

	ρ/α	0.01	0.1	0.3	0.5	0.7	0.9	0.09
Hypergraph		**0.218**	0.231	0.231	0.231	0.244	0.269	0.231
HyperPrior (LP)	10	**0.244**	0.256	0.256	0.256	0.256	0.256	0.256
	1	**0.244**	0.256	0.256	0.256	0.256	0.256	0.256
	0.1	0.256	0.256	0.269	0.269	0.256	0.256	0.256
	0.01	0.308	0.321	0.295	0.295	0.269	0.256	0.321
	0.001	0.346	0.333	0.333	0.308	0.321	0.282	0.333
	0.0001	0.462	0.474	0.487	0.397	0.333	0.321	0.462
HyperPrior (NB)	10	**0.244**	**0.244**	**0.244**	**0.244**	0.256	0.269	**0.244**
	1	**0.244**	**0.244**	**0.244**	**0.244**	0.256	0.269	**0.244**
	0.1	**0.244**	**0.244**	**0.244**	**0.244**	0.256	0.269	**0.244**
	0.01	0.308	0.308	0.282	0.282	0.282	0.269	0.308
	0.001	0.346	0.333	0.333	0.308	0.321	0.269	0.333
	0.0001	0.462	0.462	0.487	0.385	0.333	0.321	0.462

Table 4. Hypergraph and *HyperPrior* on 78 training samples from van 't Veer *et al.* dataset with 231 genes

AUC	mean	std	vs. SVM (linear)	vs. SVM (RBF)	vs. Rapaport *et al.*	vs. Li and Li	vs. Hypergraph	vs. HyperPrior-LP	vs. HyperPrior-NB
SVM (linear)	0.676	0.061	1.000	0.403	0.297	0.001	0.037	1.031E-04	1.086E-04
SVM(RBF)	0.681	0.063	0.403	1.000	0.792	0.015	0.225	0.003	0.003
Rapaport *et al.*	0.682	0.072	0.297	0.792	1.000	0.041	0.392	0.011	0.012
Li and Li	0.695	0.071	0.001	0.015	0.041	1.000	0.170	0.737	0.749
Hypergraph	0.687	0.060	0.037	0.225	0.392	0.170	1.000	0.062	0.065
HyperPrior-LP	**0.697**	0.061	1.031E-04	0.003	0.011	0.737	0.062	1.000	0.985
HyperPrior-NB	**0.697**	0.060	1.086E-04	0.003	0.012	0.749	0.065	0.985	1.000

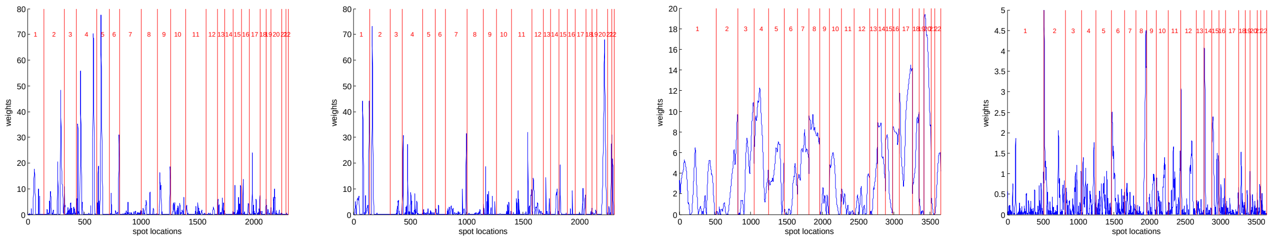
Table 5. All algorithms on van de Vijver *et al.* dataset with 326 genes

AUC	mean	std	vs. SVM (linear)	vs. SVM (RBF)	vs. Rapaport *et al.*	vs. Li and Li	vs. Hypergraph	vs. HyperPrior-LP	vs. HyperPrior-NB
SVM (linear)	0.671	0.066	1.000	0.425	0.296	0.018	0.019	2.960E-04	3.232E-04
SVM(RBF)	0.667	0.060	0.425	1.000	0.763	0.093	0.001	4.282E-06	4.766E-06
Rapaport *et al.*	0.665	0.067	0.296	0.763	1.000	0.189	0.001	3.497E-06	3.876E-06
Li and Li	0.657	0.068	0.018	0.093	0.189	1.000	2.926E-06	3.326E-09	3.745E-09
Hypergraph	0.685	0.063	0.019	0.001	0.001	2.926E-06	1.000	0.187	0.196
HyperPrior-LP	**0.692**	0.062	2.960E-04	4.282E-06	3.497E-06	3.326E-09	0.187	1.000	0.978
HyperPrior-NB	**0.692**	0.062	3.232E-04	4.766E-06	3.876E-06	3.745E-09	0.196	0.978	1.000

Table 6. All algorithms on van de Vijver *et al.* dataset with 1,464 genes

used on the training set to determine the best parameters for *HyperPrior* and the baseline algorithms to test the held-out set. Table 5 and 6 list the cross-validation results of all algorithms on van de Vijver *et al.* dataset with 326 and 1,464 cancer genes. p-values from two-sample t-test are also listed.

3 FUNCTIONAL ANALYSIS OF DISCRIMINATIVE CHROMOSOMAL REGIONS

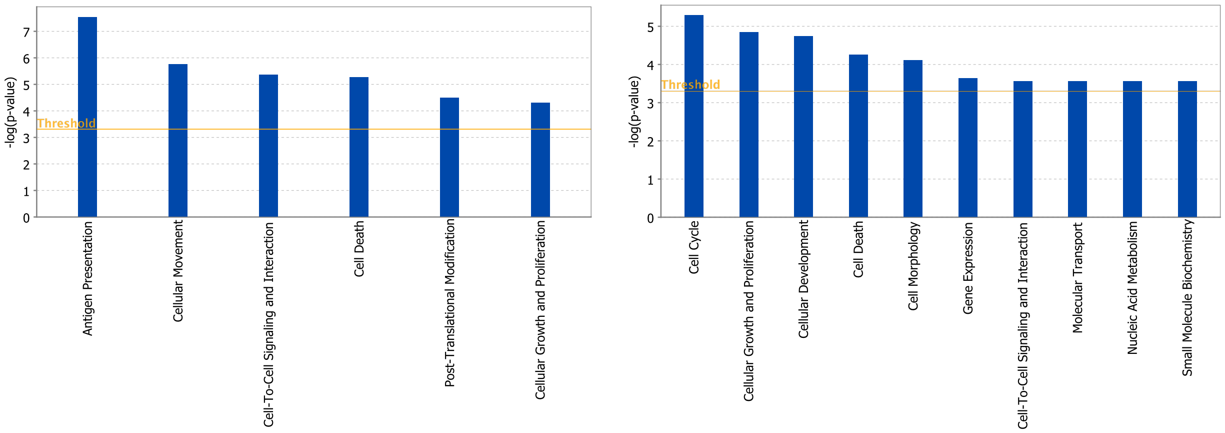


weights of DNA amplifications weights of DNA deletions
(A) Bladder cancer

weights of DNA amplifications weights of DNA deletions
(B) Melanoma cancer

Fig. 3. Discriminative regions of DNA amplification and deletion. The figures plot separately the weights of regions of "amplification state" and "deletion state", assigned by *HyperPrior* with the α and ρ parameters giving the best results in cross-validation for the grade classification on bladder tumor samples and melanoma tumor samples. The spots are ordered by their locations on chromosomes and the corresponding weights are plotted in blue curves. Red lines represent the chromosome separations.

For the two arrayCGH datasets, the weights of spots assigned by *HyperPrior* are plotted in Fig. 3. We analyze with *Ingenuity* (`http://www.ingenuity.com/`) the biological functions of the genes located in the highly weighted chromosome regions to check whether the genes involve over-represented GO categories and biological pathways that are related to bladder cancer and melanoma cancer. We select the chromosome regions associated with the top 20 highly weighted amplification states and the top 20 deletion states on both datasets. Inside these chromosome regions, 130 genes are found in the amplification regions and 255 genes are found in the deletion regions of the bladder cancer dataset, while on the melanoma cancer dataset, 205 genes are found in the amplification regions and 28 genes are found in the deletion regions . Using these genes as input, *Ingenuity* identifies 6 and 10 enriched functions scoring a p-value less than 0.0005 on the bladder and melanoma cancer datasets, respectively. The enriched functions on the bladder cancer dataset include post-translation modification, antigen presentation and cellular movement, which are all consistent with those identified by Saban *et al.* (2007); Konstantinopoulos *et al.* (2007); Smith *et al.* (2009). The enriched functions on the melanoma cancer dataset also include known gene functions related to cancer development such as cell cycle, cellular growth and proliferation, cellular development, and cell morphology (Hanahan and Weinberg, 2000; Onken *et al.*, 2006).



(A) Bladder cancer (B) Melanoma cancer

Fig. 4. Enriched biological functions in discriminative chromosomal regions.

4 CANCER GENE RANKING

We ranked the 1,464 cancer genes on van de Vijver *et al.* dataset and compare the ranking of known breast cancer genes with the ranking by correlation coefficients.

We also introduced some noise to the PPI network to make the degree of each node no less than one half of the maximum degree in the network. The top 100 genes ranked by *HyperPrior* with two groups of parameters and with a PPI to which the noise is introduced are listed in the following table:

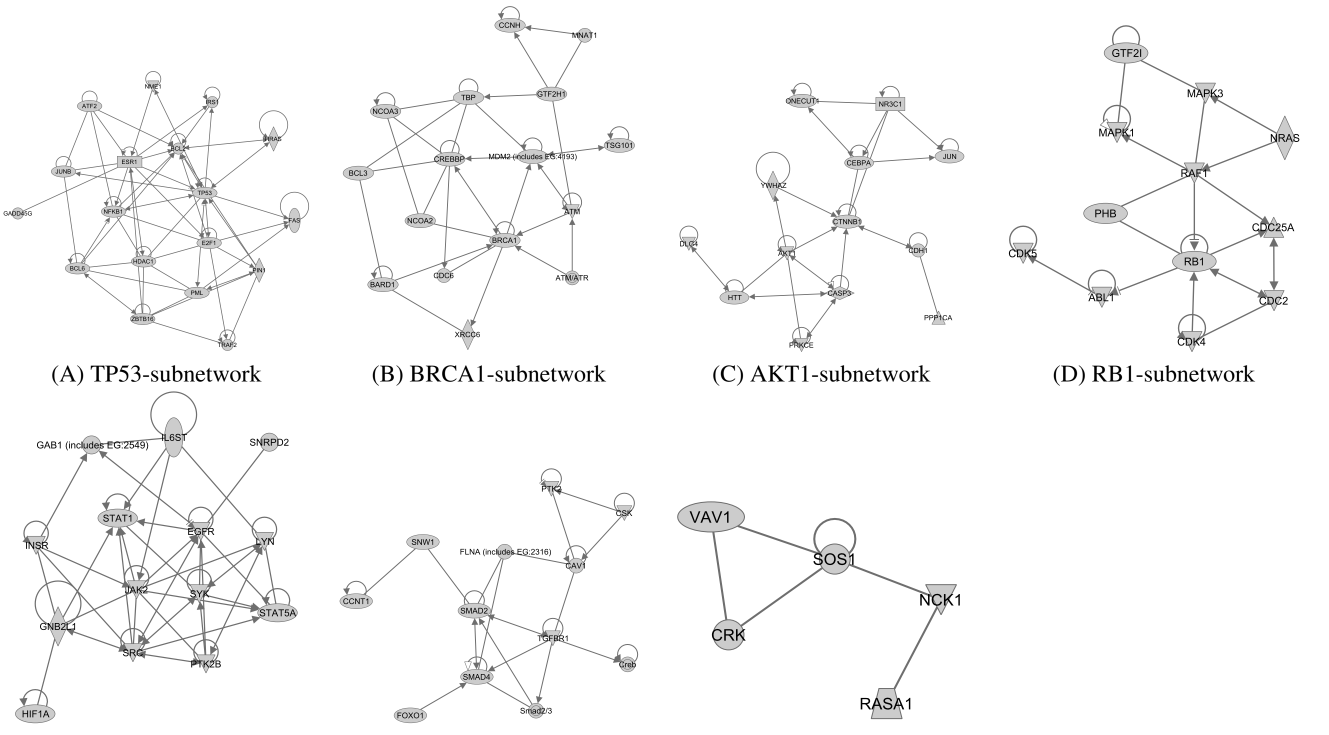
Known	**Gene Ranking**		
Disease	*HyperPrior* (LP)	*HyperPrior* (LP)	CC
Gene	$\alpha = 0.5, \rho = 1$	$\alpha = 0.5, \rho = 0.1$	
TP53	1	1	1166
BRCA1	11	12	1285
KRAS2	15	19	1057
ESR1	17	16	122
HRAS	18	14	73
BARD1	56	62	350
ATM	60	59	1154
AKT1	70	79	737
TGFB1	107	112	628
CASP8	108	120	636
PTEN	129	137	708
SERPINE1	185	136	179
PPM1D	188	116	243
BRCA2	226	258	856
PIK3CA	450	421	127
STK11	588	588	1278

Table 7. The ranking of known breast cancer (OMIM#114480) susceptibility genes. We compared the ranking of the known cancer genes obtained by the *HyperPrior* algorithm with the ranking calculated by Correlation Coefficients (CC). We set $\alpha = 0.5$ and $\rho = 1$ and 0.1 to test *HyperPrior* algorithm.

	$\alpha = 0.5, \rho = 1$	$\alpha = 0.5, \rho = 0.1$	$\alpha = 0.5, \rho = 1$ with noise								
1	TP53	TP53	TP53	35	MNAT1	ZNF145	MMP11	68	NCOA3	NCOR1	PPP2R5C
2	RB1	RB1	EGFR	36	JAK2	TNFRSF6	RPS13	69	CEBPA	VAV1	RPL6
3	MADH3	MADH3	RB1	37	SNW1	ONECUT1	BPAG1	70	AKT1	SKP2	RNF6
4	MAPK3	MAPK3	MADH3	38	CDC2	PML	P4HB	71	TRAF6	ABL1	PTHLH
5	EGFR	EGFR	CREBBP	39	NFKB1	INSR	E2F1	72	SYK	NCOA3	TYMS
6	CREBBP	JUN	JUN	40	PML	CDC25A	NESP55	73	BCL2	BCL2	CP
7	JUN	CREBBP	CTNNB1	41	CDK4	IL6ST	PSMD7	74	PIN1	GAB1	HUMGT198A
8	RAF1	CTNNB1	MAP2K4	42	GTF2H1	CDC2	RPL4	75	NCOR1	PPP1CA	FLJ20030
9	MADH2	RAF1	SLC2A5	43	CSK	E2F1	RPL11	76	E2F1	PTK2B	LRP6
10	CTNNB1	STAT1	SIL	44	HIF1A	CSK	PLOD3	77	PHB	TRAF6	NUDT2
11	BRCA1	RASA1	SH3BGRL	45	IL6ST	TRAF2	SKP2	78	BCL3	PIN1	ADRA2B
12	STAT1	BRCA1	SLC16A1	46	SNRPD2	MNAT1	KPNA2	79	ITGA6	AKT1	PSMD1
13	MDM2	MADH2	SELP	47	CASP3	CDK4	FGG	80	TBP	CEBPA	ERBB4
14	MDM2	HRAS	BRCA1	48	FOXO1A	GTF2H1	FANCA	81	DLG4	SYK	GABARAP
15	KRAS2	YWHAZ	SDHD	49	STAT5A	CRK	DKFZP564A063	82	PTK2B	PSMD7	APPL
16	MAPK1	ESR1	SFRS3	50	G22P1	HIF1A	G6PD	83	CDK5	BCR	DGKQ
17	ESR1	PTK2	SDHB	51	SAM68	USP4	DLG2	84	GAB1	TBP	INPPL1
18	HRAS	MDM2	SDHC	52	CRKL	FOXO1A	DUSP9	85	JUNB		FHL2
19	PTK2	KRAS2	LIF	53	FLNA	SNRPD2	GLTSCR2	86	GTF2I	BCL2	RPL22
20	SOS1	MDM2	EPHB2	54	ZNF145	STAT5A	TGFBR3	87	CDC6	CDC6	BMP1
21	YWHAZ	MAPK1	SFRP1	55		CRKL	CPR2	88	BCL2	BCL3	FLT3
22	NRAS	JAK2	SET	56	BARD1	HD	M17S2	89	USP4	LCK	ING3
23	ATF2	EEF1A1	MYBL2	57	TRAF2	CASP3	PP15	90	LCK	CCNA2	RANBP9
24	EGF	EGF	EEF1A1	58		FLNA	CASP2	91	LYN	GTF2I	CDC25C
25	RASA1	NRAS	BIRC5	59	CBL	ATM	MSF	92	NCOA2	JUNB	RASA1
26	HDAC1	SOS1	BUB1B	60	ATM	G22P1	TGFBI	93	GADD45A	PHB	DOC-1R
27	CAV1	ATF2	CREBL2	61	VAV1	SAM68	JAK2	94	PPP1CA	ITGA6	ZNF145
28	TNFRSF6	SNW1	IGBP1	62	CRK	BARD1	CHEK1	95	HDAC2	DLG4	NDUFS8
29	CCNH	CAV1		63	CDH1		PPGB	96	BCR	CDK5	CCNE2
30	CDC25A	NFKB1	CCNA2	64	EEF1A1	GADD45G	CCNB2	97	CCNT1	PSMD1	DDEF1
31	ONECUT1	HDAC1	GNAS1	65	ABL1		GJA1	98	HTATIP	LYN	EXT1
32	INSR	CCNH	PKMYT1	66	HD	CBL	ZNF361	99	MAP3K7	NME1	SERPINE1
33	NR3C1	NR3C1	MCM7	67	GADD45G	CDH1	PTPN13	100	NONO	NCOA2	RPS10
34	CSNK2A1	CSNK2A1	PGR								

Table 8. The top 100 genes ranked by *HyperPrior*.

5 CANCER SUBNETWORKS



(A) TP53-subnetwork (B) BRCA1-subnetwork (C) AKT1-subnetwork (D) RB1-subnetwork

(E) STAT1-subnetwork (F) SMAD2-subnetwork (G) SOS1-subnetwork

Fig. 5. Seven interaction networks of the top 100 marker genes on van de Vijver *et al.* dataset. Known breast cancer causative genes such as TP53, ESR1 and BRCA1 play a central role in the networks. Other known susceptibility genes such as v-akt murine thymoma viral oncogene homolog 1 (AKT1), retinoblastoma 1 (RB1), signal transducer and activator of transcription 1, 91kDa (STAT1), SMAD family member 2 (SMAD2), and son of sevenless homolog 1 (SOS1) also tend to be hubs and interact with many other susceptibility genes in the networks. Note that we remove those marker genes that do not directly interact with other known susceptibility genes.

6 ENRICHED FUNCTIONS

We also analyzed the biological functions of the biomarker genes from van de Vijver *et al.* dataset by Gene Ontology (GO) annotations and pathway analysis with *Ingenuity* (version 5.5). We investigated whether the identified marker genes involve significantly over-represented GO categories and biological pathways that are related with breast cancer. With the top 100 marker genes as input, *Ingenuity* identifies 17 enriched functions scoring a p-value less than $1.0e-9$ on van de Vijver *et al.* dataset. Fig. 6 shows the enriched biological functions from van de Vijver *et al.* dataset. All the 17 enriched functions of top 100 marker genes shows strong consistency with those identified by Hanahan and Weinberg (2000) and Wang *et al.* (2005), indicating that these processes are significantly involved with the progression of cancer. Especially, the most significant functions such as cell cycle (p-value = $4.03e-47$), cell death (p-value = $3.44e-44$) , gene expression (p-value = $2.43e-43$), and cellular growth and proliferation (p-value = $2.7e-36$) are well known to be functionally involved with metastasis and development of breast cancer (Sotiriou *et al.*, 2006; Wang *et al.*, 2005; Chuang *et al.*, 2007; van 't Veer *et al.*, 2002). Note that among the 17 functions, 11 functions are closely or exactly matched with the 21 functions discovered previously in Wang *et al.* (2005).

REFERENCES

Chuang, H. Y. *et al.* (2007). Network-based classification of breast cancer metastasis. *Mol Syst Biol*, **3**.

Hanahan, D. and Weinberg, R. (2000). The hallmarks of cancer. *Cell*, **100**, 57–70.

Konstantinopoulos, P. A. *et al.* (2007). Post-translational modifications and regulation of the ras superfamily of gtpases as anticancer targets. *Nat Rev Drug Discov*, **6**(7), 541–555.

Onken, M. D. *et al.* (2006). Functional Gene Expression Analysis Uncovers Phenotypic Switch in Aggressive Uveal Melanomas. *Cancer Res*, **66**(9), 4602–4609.

Rapaport, F. *et al.* (2008). Classification of arrayCGH data using fused SVM. *Bioinformatics*, **24**, i375–i382.

Saban, M. R. *et al.* (2007). Repeated BCG treatment of mouse bladder selectively stimulates small GTPases and HLA antigens and inhibits single-spanning uroplakins. *BMC Cancer*, **7**(1), 204.

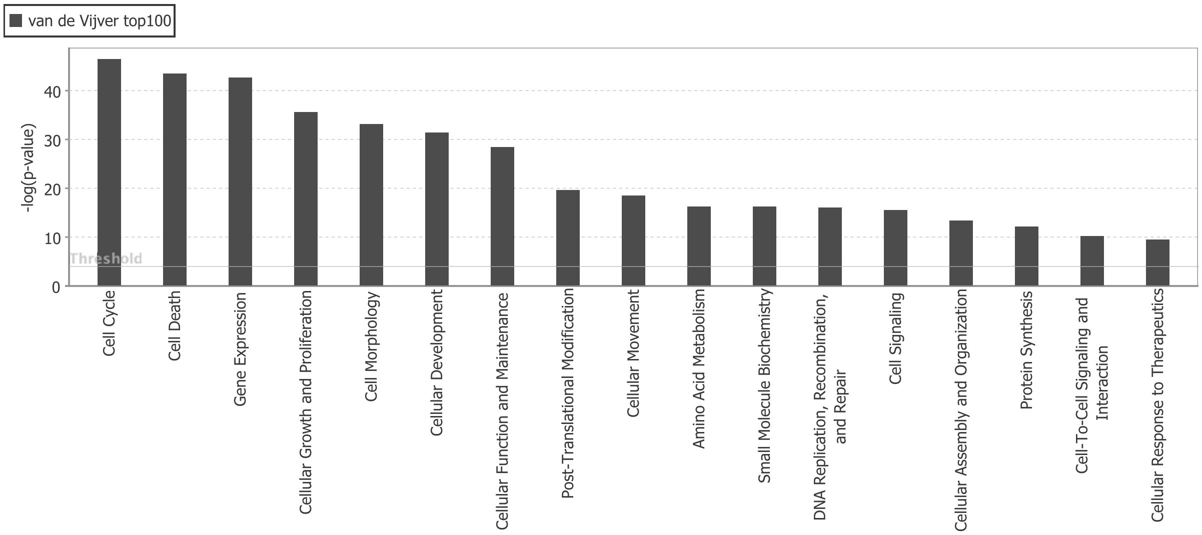


Fig. 6. Enriched biological functions by the top 100 marker genes on the van de Vijver *et al* dataset. The enriched functions are sorted by p-values calculated using the right-tailed Fisher Exact Test. All the enriched functions have p-value less than $1.0e - 9$.

Smith, S. C. *et al.* (2009). Profiling bladder cancer organ site-specific metastasis identifies LAMC2 as a novel biomarker of hematogenous dissemination. *Am J Pathol*, **174**(2), 371–379.

Sotiriou, C. *et al.* (2006). Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *Journal of the National Cancer Institute*, **98**(4), 262–272.

van de Vijver, M. J. *et al.* (2002). A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med*, **347**, 1999–2009.

van 't Veer, L. J. *et al.* (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**, 530–536.

Wang, Y. *et al.* (2005). Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *The Lancet*, **365**, 671–679.

Zhou, D. *et al.* (2006). Learning with hypergraphs: Clustering, classification, and embedding. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1601–1608.