BIOINFORMATICS

Vol. 00 no. 00 2008
Pages 1–2

Supplementary Material for: A Novel Signaling Pathway Impact Analysis (SPIA)

Adi Laurentiu Tarca[1,2], Sorin Draghici[1,3], Purvesh Khatri[2], Sonia S. Hassan[2], Pooja Mittal[2], Jung-sun Kim[2], Chong Jai Kim[2], Juan Pedro Kusanovic[2] & Roberto Romero[2]

[1]Dept. of Computer Science, Wayne State University, 431 State Hall, Detroit, MI 48202
[2]Perinatology Research Branch-NIH/NICHD, 4 Brush, 3990 John R, Detroit, MI 48201
[3]Corresponding author

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

1 COMPUTING PERTURBATION FACTORS

Let us consider the normalized weighted directed adjacency matrix of the graph describing the gene signaling network:

$$B = \begin{pmatrix} \frac{\beta_{11}}{N_{ds(g_1)}} & \frac{\beta_{12}}{N_{ds(g_2)}} & \cdots & \frac{\beta_{1n}}{N_{ds(g_n)}} \\ \frac{\beta_{21}}{N_{ds(g_1)}} & \frac{\beta_{22}}{N_{ds(g_2)}} & \cdots & \frac{\beta_{2n}}{N_{ds(g_n)}} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\beta_{n1}}{N_{ds(g_1)}} & \frac{\beta_{n2}}{N_{ds(g_2)}} & \cdots & \frac{\beta_{nn}}{N_{ds(g_n)}} \end{pmatrix} \tag{1}$$

In this matrix, β_{ij} is the efficiency with which a unit perturbation of gene j is propagated to gene i, and $N_{ds}(g_i)$ is the number of genes downstream of gene g_i. node) would sum up to 1 if taken in absolute values.

Let the vector of measured log fold-changes be:

$$\Delta E = \begin{pmatrix} \Delta E(g_1) \\ \Delta E(g_2) \\ \cdots \\ \Delta E(g_n) \end{pmatrix} \tag{2}$$

If a gene is not differentially expressed, its log fold-change is assigned the value 0. The vector of gene perturbation factors is:

$$PF = \begin{pmatrix} PF(g_1) \\ PF(g_2) \\ \cdots \\ PF(g_n) \end{pmatrix} \tag{3}$$

Then, the equations defining the perturbations after reaching a stable state:

$$PF(g_i) = \Delta E(g_i) + \sum_{j=1}^{n} \beta_{ij} \cdot \frac{PF(g_j)}{N_{ds}(g_j)} \tag{4}$$

can be re-written as:

$$PF = \Delta E + B \cdot PF \tag{5}$$

while the net accumulations of the perturbations:

$$Acc(g_i) = PF(g_i) - \Delta E(g_i) \tag{6}$$

can also be re-written as:

$$Acc = PF - \Delta E = B \cdot PF \tag{7}$$

From Eq. 5 and 7, and assuming that the matrix $I - \beta$ is non-singular, we can calculate:

$$Acc = B \cdot (I - B)^{-1} \cdot \Delta E \tag{8}$$

2 BOOTSTRAP PROCEDURE FOR COMPUTING A P-VALUE FROM PATHWAY PERTURBATIONS.

The computation of P_{PERT} for a given pathway is based on a bootstrap procedure in which we want to test if the observed global activation or inhibition of the pathway computed with the real data, t_A is unusual compared to a multitude of random scenarios. The step by step procedure we used is:

1. An iteration counter k is initialized ($k = 1$).
2. A set of $N_{de}(P_i)$ gene IDs is selected at random from the pathway P_i where the $N_{de}(P_i)$ is the number of DE genes observed on the pathway with the real data. The log fold-changes for these random gene IDs are assigned by drawing a random sample with replacement from the distribution of all DE genes to be analyzed. item Eq. 8 is used to compute the perturbation accumulations Acc, for each gene in P_i. The net total accumulation is computed as the sum of all perturbation accumulations across each pathway: $T_A(k) = \sum_i Acc(g_{ik})$.
3. Steps 2 and 3 above are repeated a large number of times ($N_{ite} = 2000$).
4. The median of T_A is computed and subtracted from $T_A(k)$ values centering their distribution around 0. The resulting corrected values are denoted with $T_{A,c}(k)$. The observed net total accumulation is also corrected for the shift in the null distribution median to give, $t_{A,c}$.
5. If $t_{A,c}$ is positive then we conclude that the pathway is activated (or positively perturbed). If $t_{A,c}$ is negative then we assume that the pathway is inhibited (or negatively perturbed).

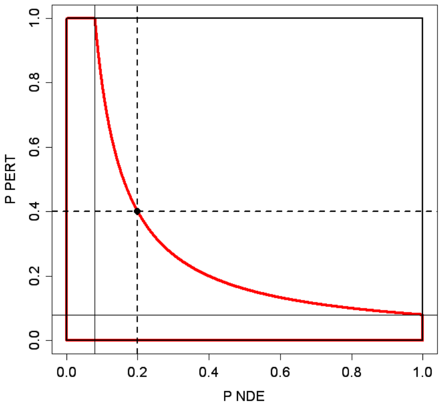

Fig. 1. Combining P_{NDE} and P_{PERT} into a single probability value, P_G. The black rectangle [0,1]x[0,1] contains all possible values that P_{NDE} and P_{PERT} can take. The curve shown is the locus of all combinations of 2 p-values that have the same product $P_{NDE} \cdot P_{PERT} = c$ (which for this example is: $c = 0.2 \cdot 0.4 = 0.08$). The points under and to the left of this curve represent all combinations that would yield a product less than 0.08. The red contour designates the surface whose area is P_G for the chosen example of the pair ($P_{NDE} = 0.2$ and $P_{PERT} = 0.4$) (black dot), under the null hypothesis. The P_G is the probability to have such a combination which can be quantified as the ratio of the area under the curve divided by the entire area of the square (which is 1). In this case, $P_G = 0.282$.

6. The probability to observe such total net inhibition or activation just by chance, P_{PERT}, is computed as:

$$P_{PERT} = \begin{cases} 2 \cdot \frac{\sum_k I(T_{A,c}(k) \geq t_{A,c})}{N_{ite}} & \text{if } t_{A,c} \geq 0 \\ 2 \cdot \frac{\sum_k I(T_{A,c}(k) \leq t_{A,c})}{N_{ite}} & \text{otherwise} \end{cases}$$

   where the identity function $I(x)$ returns 1 if x is true and 0 otherwise.T he multiplication by 2 accounts for a two-tailed test, since we do not have a particular expectation regarding the pathway status (inhibited or activated).

3 COMBINING P_{NDE} AND P_{PERT} AND INTO A GLOBAL PATHWAY SIGNIFICANCE MEASURE.

After computing a p-value for both types of evidence, P_{NDE} and P_{PERT}, we need to combine these two probabilities into one global probability value, P_G, that will be used to rank the pathways and test the research hypothesis, that the pathway is significantly impacted in the condition studied. The probability that a pair of p-values, (P_{NDE}, P_{PERT}), is observed when the null hypothesis is true, can be computed based on the fact that, under the null hypothesis, a p-value is a uniformly distributed random variable on the interval $(0, 1)$. The surface of all theoretically possible values that the variables P_{NDE} and P_{PERT} can take is a square with unity area. The two probability values obtained for a given pathway P_i can be represented as a point within this square ($P_{NDE}(i)$, $P_{PERT}(i)$), as shown in Fig. 1.

$P_{PERT}(i)$). Since under the null hypothesis $P_{NDE}(i)$ and $P_{PERT}(i)$ are independent probabilities, they can be multiplied to give the joint probability of obtaining the observed number of DE genes *and* the observed perturbation at the same time. The geometrical locus of the points with the same joint probability is the hyperbola $P_{NDE}(i) \cdot P_{PERT}(i) = c$. The probability to obtain a set of p-values as extreme or more extreme than ($P_{NDE}(i)$, $P_{PERT}(i)$), is the area under and to the left of this hyperbola. The sought global probability P_G is the probability to have such a combination with a product less than or equal to that observed. Hence, P_G can be quantified as the ratio of the area under the curve divided by the entire area of the square (which is 1):

$$P_G = \int_0^c 1 \cdot dx + \int_c^1 \frac{1}{x} \cdot dx = c + c \cdot \ln x|_c^1 = c - c \cdot \ln c \quad (9)$$

In the example shown in Fig. 1, $P_{NDE}(i) = 0.2$ and $P_{PERT}(i) = 0.4$ which yields $P_G(i) = 0.282$. Eq. 9 can be used to calculate the constant c for any desired significance threshold α. For instance, for the the customary $\alpha = 0.05$, the product of the two individual probabilities can be calculated as $c = 0.0087$, a value which has been independently obtained by others (Loughin, 2004).

Since several pathways are tested simultaneously, we also need to consider adjusting the nominal $P_G(i)$ values for multiple comparisons. For the convenience of the user, the package implementing SPIA provides both Bonferroni- and FDR-corrected p-values.

4 SUPPLEMENTARY TABLES 1-10

REFERENCES

Loughin, T. (2004) A systematic comparison of methods for combining p-values from independent tests. *Computational Statistics and Data Analysis,* **47** (3), 467 – 485.

Table 1. GSEA results on the Colorectal cancer dataset. Enrichement in cancer group. Output from R GSEA V 1.0.

	NOM p-val	FDR q-val	FWER p-val	FDR (median)	glob.p.val
Parkinsons..5020	0.008048	0.34709	0.192	0	0.155
Wnt signal..4310	0.01359	0.38628	0.354	0	0.14
Complement..4610	0.01961	0.31692	0.404	0	0.087
MAPK signa..4010	0.02115	0.17632	0.554	0.11785	0.011
Gap juncti..4540	0.02745	0.19836	0.657	0.14463	0.006
Axon guida..4360	0.02994	0.20308	0.484	0	0.03
Basal cell..5217	0.03571	0.23785	0.479	0	0.044
Colorectal..5210	0.03868	0.18203	0.528	0.12153	0.017
mTOR signa..4150	0.05253	0.16795	0.571	0.10938	0.004
Focal adhe..4510	0.06114	0.28562	0.466	0	0.07
ECM-recept..4512	0.06759	0.19884	0.513	0.13158	0.025
Regulation..4810	0.07968	0.23706	0.732	0.17157	0.01
Renal cell..5211	0.1053	0.34187	0.903	0.28226	0.015
Type II di..4930	0.1076	0.33158	0.848	0.26055	0.029
Melanogene..4916	0.1804	0.3619	0.942	0.3114	0.017

Table 2. GSEA results on the Colorectal cancer dataset. Enrichement in normal group. Output from R GSEA V 1.0.

	NOM p-val	FDR q-val	FWER p-val	FDR (median)	glob.p.val
Huntington..5040	0.1094	1	0.768	1	0.535
Dentatorub..5050	0.2189	1	0.896	1	0.592
SNARE inte..4130	0.2633	1	0.936	1	0.541
PPAR signa..3320	0.3247	1	0.96	1	0.525
Olfactory ..4740	0.3648	1	0.985	1	0.56
GnRH signa..4912	0.3908	0.93841	0.985	0.92687	0.446
Cell cycle..4110	0.4065	1	0.981	1	0.603
ErbB signa..4012	0.4345	0.83531	0.986	0.82639	0.351
Insulin si..4910	0.519	0.94025	0.996	0.96997	0.466
Thyroid ca..5216	0.5369	0.79603	0.997	0.81818	0.249
Ubiquitin ..4120	0.574	0.76453	0.999	0.77493	0.165
Tight junc..4530	0.5866	0.85592	0.996	0.87931	0.342
Phosphatid..4070	0.6321	0.73111	0.999	0.7517	0.091
Maturity o..4950	0.6604	0.72686	1	0.74598	0.058
Adipocytok..4920	0.6654	0.8032	0.999	0.82018	0.233

Table 3. SPIA results on the Vessels dataset

KEGG Pathway	P_{NDE}	P_{PERT}	P_G	$P_{G,FDR}$	$P_{G,FWER}$	Status
Antigen pr..4612	0.0067	0.0004	0.0000	0.0016	0.0016	Activated
Axon guida..4360	0.0002	0.0908	0.0002	0.0045	0.0090	Inhibited
Neuroactiv..4080	0.0006	0.1992	0.0012	0.0170	0.0514	Inhibited
Focal adhe..4510	0.0003	0.5364	0.0016	0.0170	0.0681	Inhibited
Wnt signal..4310	0.0008	0.4244	0.0032	0.0251	0.1356	Activated
Regulation..4810	0.0042	0.0948	0.0035	0.0251	0.1508	Activated
Type I dia..4940	0.0011	1.0000	0.0083	0.0469	0.3556	Inhibited
Complement..4610	0.0023	0.4812	0.0087	0.0469	0.3750	Activated
Notch sign..4330	0.0392	0.0468	0.0134	0.0579	0.5756	Activated
ECM-recept..4512	0.0024	0.7560	0.0135	0.0579	0.5789	Inhibited
Cytokine-c..4060	0.0453	0.2172	0.0553	0.2161	1.0000	Inhibited
Gap juncti..4540	0.0970	0.1236	0.0650	0.2331	1.0000	Inhibited
TGF-beta s..4350	0.0262	0.5224	0.0724	0.2396	1.0000	Inhibited
Tight junc..4530	0.2171	0.0700	0.0788	0.2421	1.0000	Inhibited
Adherens j..4520	0.0598	0.3112	0.0927	0.2659	1.0000	Activated

Table 4. ORA results on the Vessels dataset

KEGG Pathway	P_{NDE}	$P_{NDE,FDR}$	$P_{NDE,FWER}$
Axon guida..4360	0.0002	0.0065	0.0083
Focal adhe..4510	0.0003	0.0065	0.0131
Neuroactiv..4080	0.0006	0.0086	0.0257
Wnt signal..4310	0.0008	0.0089	0.0357
Type I dia..4940	0.0011	0.0091	0.0453
Complement..4610	0.0023	0.0150	0.1000
ECM-recept..4512	0.0024	0.0150	0.1050
Regulation..4810	0.0042	0.0225	0.1801
Antigen pr..4612	0.0067	0.0321	0.2886
TGF-beta s..4350	0.0262	0.1127	1.0000
Notch sign..4330	0.0392	0.1390	1.0000
Renal cell..5211	0.0405	0.1390	1.0000
MAPK signa..4010	0.0442	0.1390	1.0000
Cytokine-c..4060	0.0453	0.1390	1.0000
GnRH signa..4912	0.0502	0.1438	1.0000

Table 5. GSEA results on the Vessels dataset, enrichement in UA group. Output from R GSEA V 1.0.

KEGG Pathway	NOM p-val	FDR q-val	FWER p-val	FDR(median)	glob.p.val
Renal cell..5211	0.002953	1	0.5845	0.90625	0.4725
Cell cycle..4110	0.02559	0.71623	0.638	0.53704	0.274
Huntington..5040	0.07707	0.80367	0.787	0.66667	0.334
Thyroid ca..5216	0.1374	0.69938	0.8915	0.64444	0.2555
SNARE inte..4130	0.1621	0.607	0.7885	0.51786	0.217
Gap juncti..4540	0.2102	0.9406	0.941	0.90344	0.432
Axon guida..4360	0.2205	0.75634	0.958	0.74571	0.285
Maturity o..4950	0.224	0.80301	0.8865	0.74839	0.3435
Melanogene..4916	0.2933	0.80577	0.9755	0.79091	0.337
Focal adhe..4510	0.344	0.83775	0.958	0.82857	0.3575
Long-term ..4730	0.3685	0.78456	0.9805	0.77679	0.307
Tight junc..4530	0.3835	0.91601	0.9555	0.88176	0.4235
TGF-beta s..4350	0.4477	0.78065	0.9815	0.78733	0.297
ECM-recept..4512	0.4477	0.90872	0.9905	0.91143	0.418
PPAR signa..3320	0.5045	0.76883	0.994	0.78628	0.256

Table 6. GSEA results on the Vessels dataset, enrichement in UV group. Output from R GSEA V 1.0.

KEGG Pathway	NOM p-val	FDR q-val	FWER p-val	FDR(median)	glob.p.val
Insulin si..4910	0.01515	0.47717	0.7915	0.39286	0.1165
Type II di..4930	0.06825	0.74158	0.7255	0.61111	0.303
Toll-like ..4620	0.07475	1	0.6295	0.81481	0.445
Parkinsons..5020	0.1381	0.60681	0.779	0.52381	0.202
Neuroactiv..4080	0.1388	0.42792	0.8725	0.39286	0.039
ErbB signa..4012	0.1756	0.66679	0.9595	0.64706	0.218
Type I dia..4940	0.2022	0.39259	0.8285	0.36667	0.04
Antigen pr..4612	0.2377	0.46468	0.8265	0.44	0.091
MAPK signa..4010	0.2721	0.74577	0.9815	0.73333	0.2695
Adipocytok..4920	0.2964	0.69667	0.9845	0.69565	0.2325
Natural ki..4650	0.3005	0.67092	0.9695	0.65812	0.2045
Cytokine-c..4060	0.338	0.69065	0.986	0.6875	0.1915
Alzheimers..5010	0.5056	0.89088	0.993	0.91124	0.401
Taste tran..4742	0.5474	0.74295	0.993	0.76389	0.155
Epithelial..5120	0.562	0.76919	0.993	0.78571	0.2225

Table 7. SPIA results on the LaborM dataset

KEGG Pathway	P_{NDE}	P_{PERT}	P_G	$P_{G,FDR}$	$P_{G,FWER}$	Status
Cytokine-c..4060	0.0000	0.0000	0.0000	0.0000	0.0000	Activated
ErbB signa..4012	0.0000	0.0112	0.0000	0.0001	0.0002	Activated
Jak-STAT s..4630	0.0000	0.2140	0.0000	0.0004	0.0011	Activated
Epithelial..5120	0.0044	0.0024	0.0001	0.0015	0.0059	Activated
Complement..4610	0.0003	0.0740	0.0003	0.0023	0.0113	Inhibited
MAPK signa..4010	0.0011	0.4076	0.0038	0.0288	0.1726	Activated
Toll-like ..4620	0.0007	0.9344	0.0058	0.0370	0.2591	Activated
Adipocytok..4920	0.0063	0.4524	0.0195	0.1099	0.8793	Activated
PPAR signa..3320	0.0044	1.0000	0.0284	0.1218	1.0000	Inhibited
TGF-beta s..4350	0.0408	0.1084	0.0284	0.1218	1.0000	Activated
Insulin si..4910	0.0159	0.2944	0.0298	0.1218	1.0000	Inhibited
Type II di..4930	0.0857	0.1668	0.0750	0.2813	1.0000	Inhibited
Thyroid ca..5216	0.1189	0.1528	0.0910	0.2897	1.0000	Inhibited
Wnt signal..4310	0.1143	0.1828	0.1017	0.2897	1.0000	Inhibited
Circadian ..4710	0.0774	0.2704	0.1019	0.2897	1.0000	Inhibited

Table 8. ORA results on the LaborM dataset

KEGG Pathway	P_{NDE}	$P_{NDE,FDR}$	$P_{NDE,FWER}$
Cytokine-c..4060	0.0000	0.0000	0.0000
Jak-STAT s..4630	0.0000	0.0002	0.0004
ErbB signa..4012	0.0000	0.0005	0.0014
Complement..4610	0.0003	0.0032	0.0130
Toll-like ..4620	0.0007	0.0067	0.0335
MAPK signa..4010	0.0011	0.0081	0.0485
PPAR signa..3320	0.0044	0.0249	0.1989
Epithelial..5120	0.0044	0.0249	0.1989
Adipocytok..4920	0.0063	0.0315	0.2833
Insulin si..4910	0.0159	0.0715	0.7150
Natural ki..4650	0.0283	0.1158	1.0000
TGF-beta s..4350	0.0408	0.1531	1.0000
Renal cell..5211	0.0521	0.1714	1.0000
ECM-recept..4512	0.0533	0.1714	1.0000
Circadian ..4710	0.0774	0.2263	1.0000

Table 9. GSEA results on the LaborM dataset, enrichement in TL group. Output from R GSEA V 1.0.

KEGG Pathway	NOM p-val	FDR q-val	FWER p-val	FDR(median)	glob.p.val
Epithelial..5120	0.00497	0.18574	0.1105	0	0.0965
MAPK signa..4010	0.005066	0.13574	0.3995	0	0.007
Cytokine-c..4060	0.008214	0.22727	0.318	0	0.0585
ErbB signa..4012	0.01091	0.16509	0.3675	0	0.0185
TGF-beta s..4350	0.01094	0.29343	0.2815	0	0.1215
Jak-STAT s..4630	0.01132	0.14696	0.3835	0	0.013
VEGF signa..4370	0.02198	0.18533	0.3405	0	0.0305
Adipocytok..4920	0.02402	0.14489	0.491	0	0.006
Complement..4610	0.04893	0.1504	0.4675	0	0.0085
Toll-like ..4620	0.06592	0.16007	0.5515	0.11176	0.006
Fc epsilon..4664	0.07407	0.16072	0.5785	0.11144	0.0035
Insulin si..4910	0.08918	0.2541	0.828	0.20956	0.004
Type II di..4930	0.09164	0.25983	0.77	0.20276	0.0105
Renal cell..5211	0.09323	0.26766	0.7935	0.20879	0.0105
Natural ki..4650	0.1095	0.16695	0.613	0.11728	0.0035

Table 10. GSEA results on the LaborM dataset, enrichement in TNL group. Output from R GSEA V 1.0.

KEGG Pathway	NOM p-val	FDR q-val	FWER p-val	FDR(median)	glob.p.val
Parkinsons..5020	0.2297	1	0.88	1	0.7075
Melanogene..4916	0.268	1	0.985	1	0.8135
Basal cell..5217	0.3283	1	0.9955	1	0.566
Amyotrophi..5030	0.4055	1	0.9955	1	0.755
Wnt signal..4310	0.4554	1	0.9995	1	0.628
Phosphatid..4070	0.4956	1	0.9995	1	0.502
Long-term ..4730	0.5726	0.97732	1	0.96571	0.46
Thyroid ca..5216	0.6073	0.94022	1	0.93914	0.423
Tight junc..4530	0.6633	0.90019	1	0.90794	0.3625
Gap juncti..4540	0.8018	0.93991	1	0.9602	0.4185
Olfactory ..4740	0.9208	1	1	1	0.9045
Regulation..4140	0.9756	1	1	1	0.964
Taste tran..4742	0.9902	0.98757	1	1	0.759