Electronic Supplementary Material (ESI) for Molecular BioSystems.

Supplementary Data

GTA: a Game Theoretic Approach to Identifying Cancer Subnetwork Markers

This file contains:

- **Supplementary Information 1-2**
- **Supplementary Figures 1-5**

Supplementary Information 1

A simple example of Nash equilibria

As a simple example of Nash equilibria, suppose there are two players in a game. Player 2 has strategy set $\Sigma_1 = \{U,D\}$ and player 1 has strategy set $\Sigma_2 = \{L,M,R\}$ and the payoff matrix is:

$P2$

$P1$		L	M	R
	U	8,0	5,-1	4,-2
	D	10,1	4,0	1,-1

Assuming players rationality, if player 2 chooses the strategy U then the strategy L is the best choice for player 1, also if player 2 chooses the strategy D then L is the best one for player 1. If player 1 chooses strategies L, $M$and R, then, D, U and U respectively provide the best payoffs for player 2. Finally, the pure Nash equilibria is (D, L).

Supplementary Information 2

GTA Scoring Algorithm

In our game theory approach, a scoring scheme is proposed based on a payoff function as a combination of gain function and loss function which are described in the following. It should be noted that joining or leaving a subnetwork are the main strategies that can be chosen by each player.

Gain function. Suppose in a subnetwork G_s, there are $|V_s| = n$ proteins corresponding to n unique genes ($Genes = \{g_1, g_2, \ldots, g_n\}$). These genes have expression values for m different samples in the microarray datasets. The expression vector $x_i = (x_i^1, x_i^2, \ldots, x_i^m)$ contains expression values of gene i, in which $i = 1,2,\ldots,n$ and x_i^j is the expression level of gene i in sample j. For considering phenotypes of the samples in their expressions, we compute the log-likelihood ratio (LLR) of each gene that indicates which phenotype is more likely based on a given expression of that gene. The LLR for gene g_i, for two different phenotypes is defined by

$$LLR_i(x_j^i) = \log\left[f_i^1(x_j^i) \Big/ f_i^2(x_j^i) \right] \qquad (1)$$

Where $f_i^1(x_i^j)$ and $f_i^2(x_i^j)$ are the conditional probability density function (PDF) of the expression level of gene g_i under phenotype 1 and phenotype 2 respectively.

A local scoring (LS) function is also defined for each gene g_i. By this scoring, we try to find the role of each protein in the subnetwork in connecting DEGs. The LS function for gene i with joining strategy is defined as equation (3), in which k is the number of neighbor genes of the gene i in the subnetwork G_s.

$$LS_i = \sum_{j=1}^{k} t - score(LLR_{i_j}) \qquad (3)$$

Where $g_{i_1}, g_{i_2}, \ldots, g_{i_k}$ are k neighbors of gene g_i in the subnetwork.

Furthermore, to score the connectivity of each subnetwork, a density value is assigned. For a subnetwork $G_s = (V_s, E_s)$, the density value is defined by:

$$DE(G_s) = \frac{\sum_{e \in E_s} w(e)}{\binom{|V_s|}{2}} \quad (4)$$

Where $w(e)$is the weight of edge e based on Lage's method.

Finally, the gain function (GF) is determinedas equation (5) for gene i in subnetwork G_s, in which α, β and γ are constants.

$$GF(i,G_s) = \alpha.t - score(LLR_i) + \beta.LS_i + \gamma.DE(G_s) \quad (5)$$

In above equation, α, β and γ are weighting parameters to imply each function's importance and $t - score(LLR_i)$ is the t-test statistics score of the LLR_i.

Loss function. The loss function (LF) for gene i with joining strategy is defined in equation (6), where c is a constant.

$$LF(i,G_s) = c.(|V_s| - 1) \quad (6)$$

Payoff function. Eventually, the payoff function (PF) for a given agent i and the subnetwork G_s is calculated as follows:

$$PF(i,G_s) = GF(i,G_s) - LF(i,G_s) \quad (7)$$

By examining different values for constants in payoff function, using numerical method, the most powerful discriminatory subnetworks were achieved by setting $\alpha = 1.24,\ \beta = 1,\ \gamma = 1$ and $= 2$.

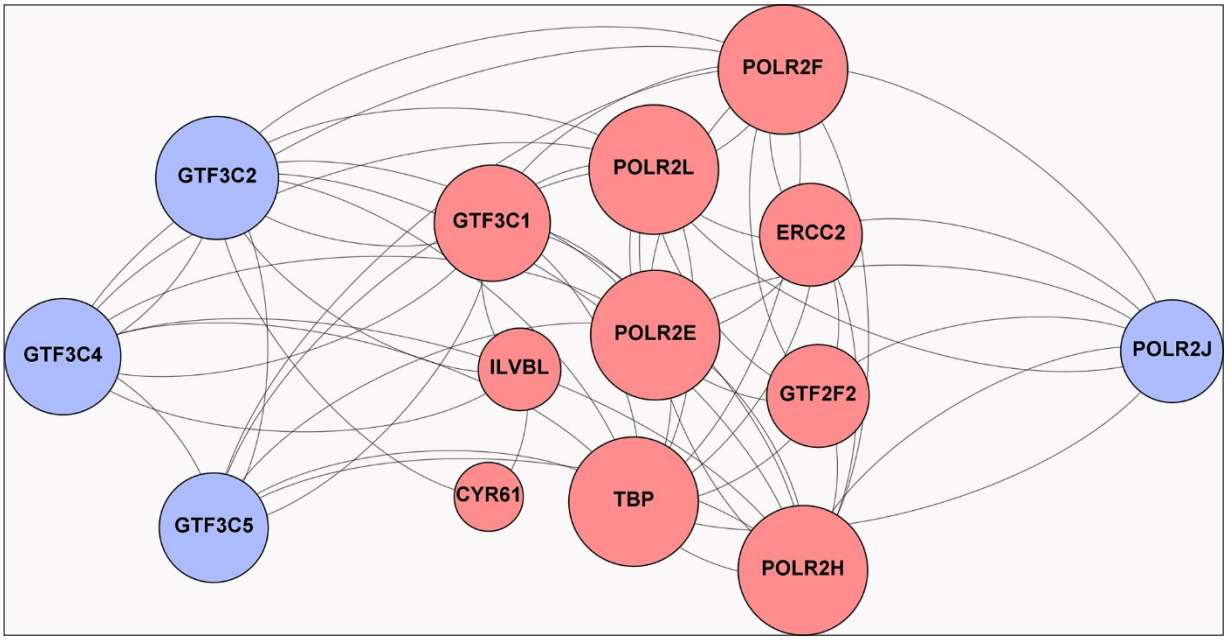
Supplementary Figure 1. The pseudo-code of the GTA algorithm

Input: Weighted PPI Network G=(V,E,w), Absolute t-scores of Genes (t-scores), Number of Subnetwork Markers (N)

Output: List of Ranked Subnetwork Markers

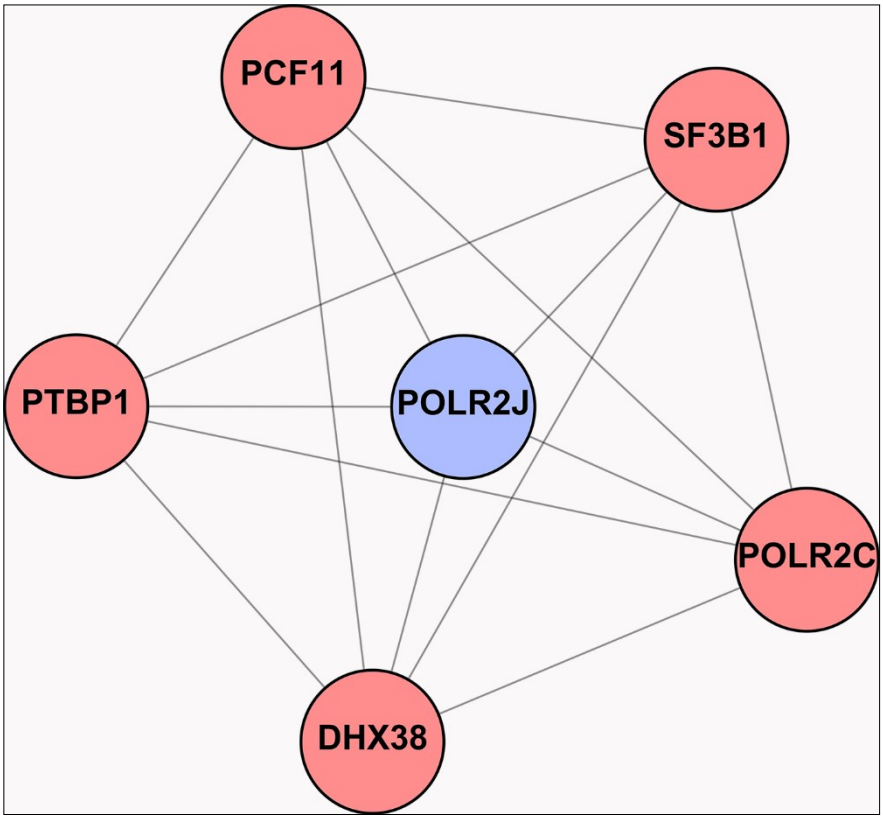
1. Sort Gene List by Their Absolute t-scores in Decreasing Order ;
2. For i=1: N
   - Deg= Degree of Gene i ;
   - If (Deg >= Average Degree of PPIN nodes)
     - ✓ Seed=Gene i ;
     - ✓ Candidate_Subnetwork=BFS (Seed, 2) ; //Using Breadth First Search and starting from the seed gene, nodes with at most two interactions away from the seed are returned
     - ✓ Subgames= Divide candidate subnetwork into several subgames;
     - ✓ For each subgame do
       - Payoffs=Calculate the payoff value for each player;
       - Equilibriums=Calculate Nash equilibria;
       - Selected= Choose the best of the Nash equilibria // Based on the average payoff values of associated genes
     - ✓ Optimized_Subnetwork=Merge all selected equilibria of subgames;
     - ✓ End
   - Subnetwork_Markers(i)=Optimized subnetwork;
   - End
3. Do Post-proccessing on each optimized subnetwork markers; // Based on K-means

Supplementary Figure 2. POLR2J-based subnetwork in the Netherland dataset.



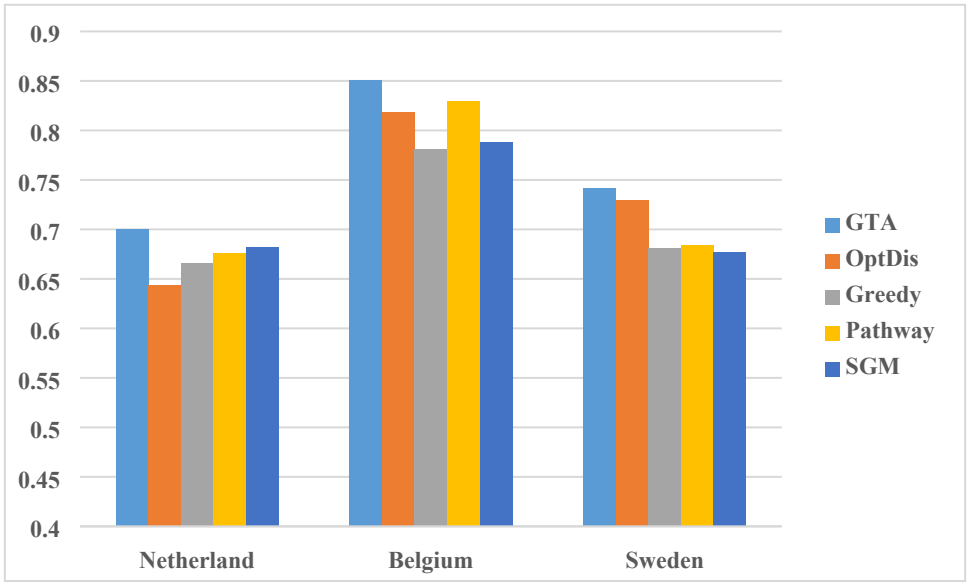
Node colour represents changes in level of expression where red and blue node are DEGs and non-DEGs respectively. Node degree is proportional to the diameter of each node. All of the edges have a confidence-weight of 1.0, indicating high confidence of interactions in the subnetwork.

Supplementary Figure 3. POLR2J-based subnetwork in the Sweden dataset.



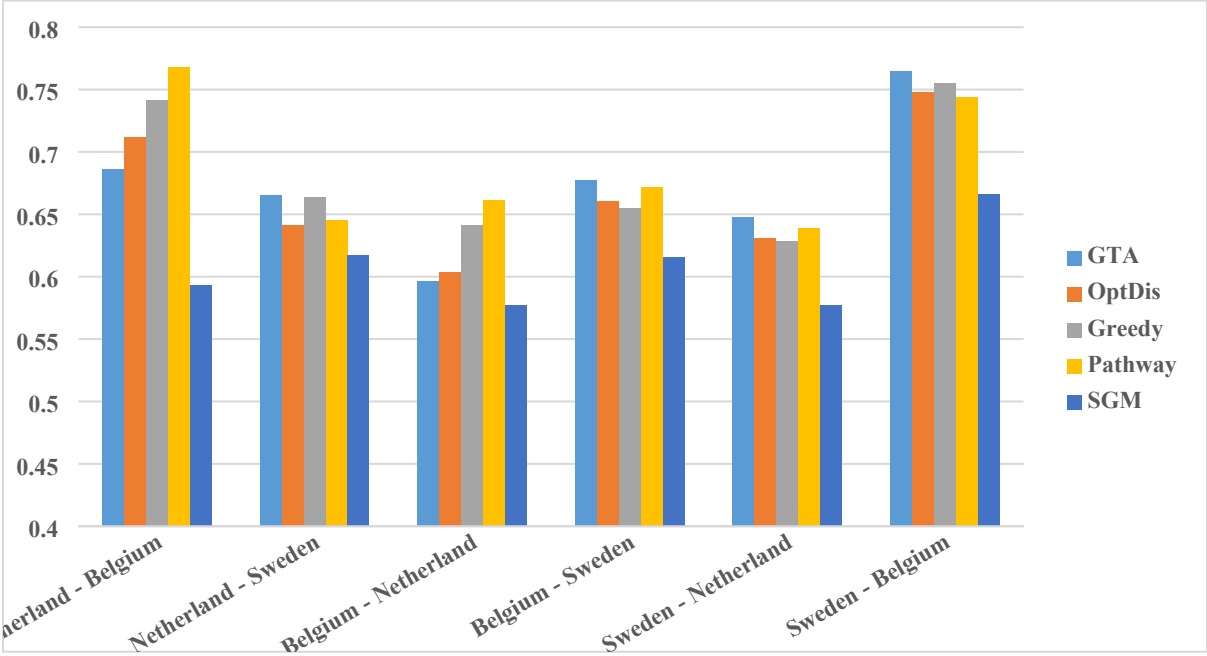
Node colour represents changes in level of expression where red and blue node are DEGs and non-DEGs respectively. Node degree is proportional to the diameter of each node. All of the edges have a confidence-weight of 1.0, indicating high confidence of interactions in the subnetwork.

Supplementary Figure 4. The average accuracy of within-dataset experiments



The bar chart shows the results of the within-dataset experiments based on the Netherland, Belgium and Sweden datasets. It shows the average accuracy of the classifier constructed by markers identified by GTA, OptDis method, the greedy method, pathway- and gene-based methods.

Supplementary Figure 5. The average accuracy of cross-dataset experiments testing reproducibility



The bar chart shows the average accuracy of the SVM classifier that uses subnetwork markers identified by GTA, OptDis method, the greedy method, pathway- and gene-based methods. In order to evaluate the reproducibility of various markers, we used the first dataset to identify markers and the second dataset to train the classifier.