BIOINFORMATICS

Vol. 30 ISMB 2014, pages i228–i236
doi:10.1093/bioinformatics/btu278

DrugComboRanker: drug combination discovery based on target network analysis

Lei Huang[1,2,†], Fuhai Li[2,†], Jianting Sheng[2], Xiaofeng Xia[2], Jinwen Ma[1], Ming Zhan[2] and Stephen T.C. Wong[2,*]

[1]Department of Information Science, School of Mathematical Sciences and LMAM, Peking University, Beijing, 100871, China and [2]NCI Center for Modeling Cancer Development, Department of Systems Medicine and Bioengineering, Houston Methodist Hospital Research Institute; Weill Cornell Medical College of Cornell University, Houston, TX 77030, USA

ABSTRACT

Motivation: [illegible]

Results: [illegible] **DrugComboRanker** [illegible]

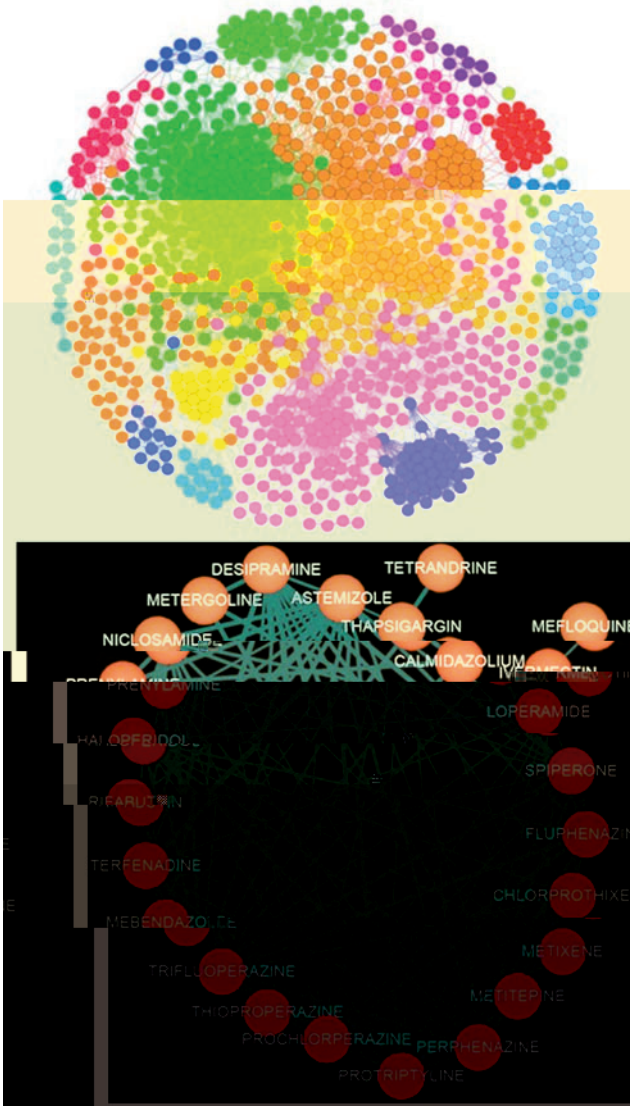

Fig. 2. Snapshots of the drug functional network (top) and an example of the reconstructed drug network community (bottom)

is, each row of $\boldsymbol{W}_N$ is subject to

$$\sum_{j=1}^{K} wn_{ij} = 1 \qquad (2)$$

where wn_{ij} is the element of $\boldsymbol{W}_N$ and it quantifies the membership of node i with respect to community j. To infer the appropriate model order K, we use a Bayesian paradigm for non-negative matrix factorization by placing automatic relevance determination priors with scale hyper-parameters $\boldsymbol{\lambda} = \{\lambda_k\}_{k=1}^{K}$ on the variables w_{ik} and h_{kj}. In this model, the distribution of λ_k is parameterized by fixed parameters a and b, and the fixed parameter β decides the distance measure between the observed interactions $\mathbf{V}$ and the expected interactions $\hat{\mathbf{V}}$. Under these assumptions, the posterior density function can be obtained as

$$p(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{\lambda} | \boldsymbol{V}) = \frac{p(\boldsymbol{V} | \boldsymbol{W}, \boldsymbol{H}) p(\boldsymbol{W} | \boldsymbol{\lambda}) p(\boldsymbol{H} | \boldsymbol{\lambda}) p(\boldsymbol{\lambda})}{p(\boldsymbol{V})} \qquad (3)$$

Maximizing the posterior density is equivalent to minimizing the negative log posterior, which can be regarded as a loss function $C_{MAP}(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{\lambda})$ as

$$C_{MAP}(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{\lambda}) \triangleq -\log p(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{\lambda} | \boldsymbol{V}) \qquad (4)$$

$$= -\log p(\boldsymbol{V} | \boldsymbol{W}, \boldsymbol{H}) - \log p(\boldsymbol{W} | \boldsymbol{\lambda}) - \log(\boldsymbol{H} | \boldsymbol{\lambda}) - \log p(\boldsymbol{\lambda}) \qquad (5)$$

where $-\log p(\boldsymbol{V} | \boldsymbol{W}, \boldsymbol{H})$ is the log-likelihood.

The generalized β-divergence is defined by

$$D_\beta(x|y) \triangleq \begin{cases} \frac{x^\beta}{\beta(\beta-1)} + \frac{x^\beta}{\beta} - \frac{xy^{\beta-1}}{\beta-1} & \beta \in \mathbb{R} \setminus \{0, 1\} \\ x \log \frac{x}{y} - x + y & \beta = 1 \\ \frac{x}{y} - \log \frac{x}{y} - 1 & \beta = 0 \end{cases} \qquad (6)$$

The β-divergence can be regarded as a minus log-likelihood for the Tweedie distribution and its probability density function is given by

$$f(x; \mu, \phi, \beta) = h(x, \phi) \exp\left\{\frac{1}{\phi}\left(\frac{1}{\beta - 1} x^{\beta - 1} - \frac{1}{\beta} \mu^{\beta}\right)\right\} \qquad (7)$$

where $h(x, \phi)$ is the base measure function, μ is the mean, ϕ is the dispersion parameter and β is the shape parameter. Assuming that v_{ij} is generated from the Tweedie distribution, the log-likelihood function can be given by

$$\log p(\boldsymbol{V} | \boldsymbol{W}, \boldsymbol{H}) = -\frac{1}{\phi} D_\beta(\boldsymbol{V} | \boldsymbol{W}\boldsymbol{H}) + C \qquad (8)$$

To insure $\boldsymbol{W}$ and $\boldsymbol{H}$ are non-negative, the Half-Normal priors are assigned on them,

$$p(w_{ik} | \lambda_k) = HN(w_{ik} | \lambda_k) \qquad (9)$$

$$p(h_{kj} | \lambda_k) = HN(h_{kj} | \lambda_k) \qquad (10)$$

where $$HN(x | \lambda) \triangleq \left(\frac{2}{\lambda}\right)^{\frac{1}{2}} \exp\left(-\frac{x^2}{2\lambda}\right), \quad x \geq 0 \qquad (11)$$

and place an inverse Gamma priors on each λ_k,

$$p(\lambda_k; a, b) = \frac{b^a}{\Gamma(a)} \lambda_k^{-(a+1)} \exp\left(-\frac{b}{\lambda_k}\right) \qquad (12)$$

Then, according to Equation (5), the objective function $C_{MAP}(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{\lambda})$ can be given as

$$C_{MAP}(\boldsymbol{W}, \boldsymbol{H}) = \frac{1}{\phi} D_\beta(\boldsymbol{V} | \boldsymbol{W}\boldsymbol{H}) + \sum_{k=1}^{K} \frac{1}{\lambda_k}\left(\frac{1}{2} w_k^2 + \frac{1}{2} h_k^2 + b\right) + (N + a + 1) \log \lambda_k + C \qquad (13)$$

To minimize $C_{MAP}(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{\lambda})$ with respect to $\boldsymbol{W}$, $\boldsymbol{H}$ and $\boldsymbol{\lambda}$, we adopt the strategy in (Tan and Fevotte, 2013) by introducing a local majorization–minimization algorithm with efficient multiplicative updates. Finally, we give the overlapping community detection algorithm as follows:

Step 1: Initialize $\mathbf{w}_k \in \mathbb{R}_+^{N \times K}$ and $\mathbf{h}_k \in \mathbb{R}_+^{K \times N}$ to random non-negative values.

Step 2: Update $\boldsymbol{W}$, $\boldsymbol{H}$ and λ_k by

$$\boldsymbol{H} = \boldsymbol{H} \left(\frac{\boldsymbol{W}^T((\boldsymbol{W}\boldsymbol{H})^{\beta-2} \boldsymbol{V})}{\boldsymbol{W}^T(\boldsymbol{W}\boldsymbol{H})^{\beta-1} + \text{repmat}(\boldsymbol{\lambda}, 1, N)}\right)^{\gamma(\beta)}$$

$$\boldsymbol{W} = \boldsymbol{W} \left(\frac{((\boldsymbol{W}\boldsymbol{H})^{\beta-2} \boldsymbol{V})\boldsymbol{H}^T}{(\boldsymbol{W}\boldsymbol{H})^{\beta-1}\boldsymbol{H}^T + \text{repmat}(\boldsymbol{\lambda}, 1, N)}\right)^{\gamma(\beta)}$$

$$\lambda_k = \left(\frac{1}{2}\sum_i w_{ik}^2 + \frac{1}{2}\sum_j h_{kj}^2 + b\right) \quad (N + a + 1)$$

$$\gamma(\beta) = \begin{cases} 1 & (3 \quad \beta) \quad \beta \quad 2 \\ 1 & (\beta \quad 1) \quad \beta > 2 \end{cases}$$

Step 3: Repeat Step 2 until $\max_{k=1\,2\,\ldots\,K}(\lambda_k^{new} \quad \lambda_k^{old}) \; \lambda_k^{old}$

Step 4: Normalize $\boldsymbol{W}$ to $\boldsymbol{W}_N$, then the number of non-zero columns k of $\boldsymbol{W}_N$ is the number of clusters. Assign each node to the k clusters according to $\boldsymbol{W}_N$. In the above algorithm, $\boldsymbol{X}$ $\boldsymbol{Y}$ denotes element-by-element multiplication of $\boldsymbol{X}$ and $\boldsymbol{Y}$; $\frac{\boldsymbol{X}}{\boldsymbol{Y}}$ denotes element-by-element division of $\boldsymbol{X}$ and $\boldsymbol{Y}$; and $\boldsymbol{X}^{\gamma}$ denotes raising each element of $\boldsymbol{X}$ to the γ^{th} power. In addition, repmat($\boldsymbol{\lambda}$ 1 N) denotes the K N matrix with each column being the vector$\boldsymbol{\lambda}$. Using the BNMFβD approach, we partition the drug functional network into a set of connected network modules (Fig. 2), within which drugs share common targets or related signaling mechanisms.

2.4 Drug combination discovery based on target network analysis

The novel drug combination approach consists of the following three major components.

(a) Disease specific signaling network reconstruction

Several approaches (Barrenäs et al., 2012; Chuang et al., 2007; Ideker et al., 2002) have been proposed to reconstruct signaling networks of diseases based on transcritpome and interactome data. In this study, the approach proposed in (Barrenäs et al., 2012) is used. The integrated protein–protein interactions from BioGRID (Stark et al., 2006) and the manually curated human cancer signaling from (Awan et al., 2007; Cui et al., 2007; Li et al., 2012; Newman et al., 2013) (available at http://www.bri.nrc.ca/wang/) are clustered into functional protein–protein modules, and each module is tested for enrichment (Fisher's exact test, $P < 0.01$) of the differentially expressed genes of the gene expression profile of a disease. The enriched modules are then considered as disease susceptibility modules, from which the highly interconnected genes are identified as the disease-specific signaling network. Figure 3 shows the reconstructed signaling network of lung adenocarcinoma.

(b) Functional drug target prediction using network-based recommendation

Drugs often have multiple targets and affect distinct signaling modules, but only parts of them are known for given drugs. The drug communities embed targeting signaling modules (functional targets instead of physical targets) of drugs. To uncover the targeting signaling modules of drugs, we propose a network-based recommendation approach as follows. Let $D = \{d_1\ d_2\ \ldots\ d_m\}$ denote drugs in a given community, and $T = \{t_1\ t_2\ \ldots\ t_n\}$ be the known drug targets. The drug–target interaction network can be described as a bipartite graph G(D T E). The E indicates the known drug and target associations. This drug–target network can be represented by an adjacent matrix $A = \{a_{ji}\}_{n\ m}$, where a_{ji} is the weight that quantifies the association between d_i and t_j. Then the novel network-based algorithm is designed based on a bipartite network projection technique (Zhou et al., 2010) as follows.

$$R = F \quad A \qquad (14)$$

where $R = \{r_{ij}\}_{n\ m}$ is the recommendation score (the functional association possibility between drug i and target j). The $F = \{f_{ij}\}_{n\ m}$ indicates the transition matrix from drug i to drug j and is defined as:

$$f_{ij} = \frac{1}{\Gamma(i\ j)} \sum_{l=1}^{m} \frac{a_{il} a_{jl}}{k(x_l)} \qquad (15)$$

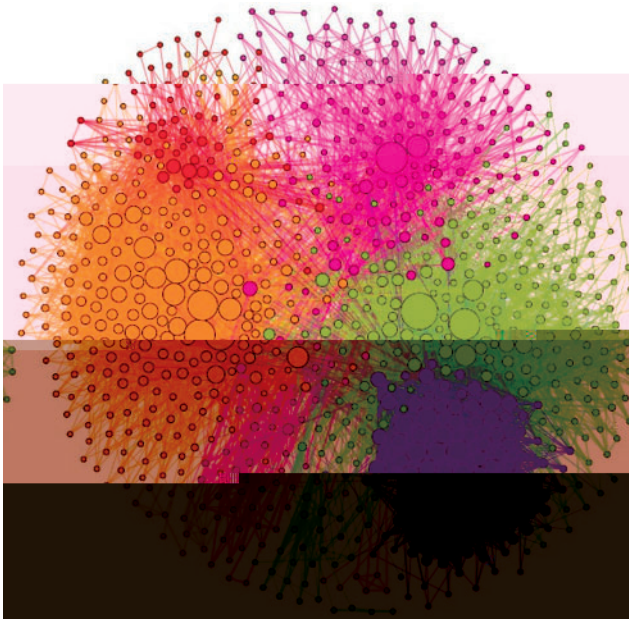

Fig. 3. The reconstructed signaling network of lung adenocarcinoma. Different node colors indicate different modules, and the node size indicates the degree of nodes

where $\Gamma(i\ j) = k(t_i)^1\ {}^{\lambda}k(t_j)^{\lambda}$ and k(x) is the degree of the x node in the bipartite network. Targets with recommendation scores greater than given threshold, 0.1, are kept as the active functional targets of given drugs.

(c) Disease-specific drug combination discovery

With the constructed disease signaling network and the predicted drug targets, drug combinations are then prioritized by combining the following synergistic scores. Given two candidate drugs, d_i and d_j, from different clusters, suppose d_i 2 C_k and d_j 2 C_h, and $T_k = \{t_{k1}\ t_{k2}\ \ldots\ t_{km}\}$ denote the targets of d_i in C_k, and $T_h = \{t_{h1}\ t_{h2}\ \ldots\ t_{hn}\}$ denote the targets of d_j in C_h. The first synergistic score is defined as follows.

$$S_1(i\ j) = \frac{\sum_i CS(t_{ki}) \exp(\ D\{t_{ki}\ T_h\}\ n^2)}{\sum_i CS(t_{ki})} + \frac{\sum_j CS(t_{hj}) \exp(\ D\{t_{hj}\ T_k\}\ m^2)}{\sum_j CS(t_{hj})} \qquad (16)$$

where $CS(t_{ki})$ is the centrality score of target t_{ki} in the reconstructed disease signaling network, and it is an additive of betweenness (Bn), closeness (Cn) (Brandes and Fleischer, 2005) and PageRank (Pr) score (Page et al., 1999) of protein t_{ki}, that is

$$CS(t_{ki}) = Bn(t_{ki}) + Cn(t_{ki}) + Pr(t_{ki}) \qquad (17)$$

These are three different but correlated centrality measurements, and the reason of combing them is to get a robust centrality score. The min D $\{t_{ki}\ T_h\}$ is the minimum shortest path from t_{ki} to T_h. The first synergistic score, $S_1(i\ j)$, prefers drug combinations, whose targets are in the center (hubs) of disease signaling network and closely connected.

The second synergistic score is defined as

$$S_2(i\ j) = \frac{\sum_{i\ j} \mathrm{Sim}(t_{ki}\ t_{hj})}{(m + n)(m + n \quad 1)} \qquad (18)$$

where $\mathrm{Sim}(t_{ki}\ t_{hj})$ is the semantic similarity of gene ontology (GO) annotations of t_{ki} and t_{hj} (Couto et al., 2007), which is computed based on the overlap of the GO terms that are associated with t_{ki} and t_{hj}, and is defined as

$$\mathrm{Sim}(t_{ki}\ t_{hj}) = \frac{2\log_2 \max\{p(A)\}}{(\log_2 p(GO_{ki}) + \log_2 p(GO_{hj}))} \quad (19)$$

Where GO_{ki} is the GO term that associated with t_{ki}, and A is a GO term that is an ancestor of both GO_{ki} and GO_{hj}, and

$$p(GO_{ki}) = \frac{\mathrm{Freq}(GO_{ki})}{\mathrm{MaxFreq}} \quad (20)$$

$\mathrm{Freq}(GO_{ki})$ is the frequency of GO term GO_{ki} occurring in GO annotations, which are taken from GO database. MaxFreq is the maximum occurrences frequency of GO terms that are associated with all the targets and the predicted drug targets in the GO annotations. The second synergistic score, $S_2(i\ j)$, prefers drug combinations that block genes with similar functions, e.g. cell proliferation.

Finally, the synergistic score of drug d_i and d_j is given by

$$S_y(i\ j) = (S_1(i\ j) + S_2(i\ j))\ d_E(i\ j) \quad (21)$$

where

$$d_E(i\ j) = \max\left\{S_G(i\ j)\ \sqrt{S_G(i\ j)\ S_s(i\ j)}\ S_G(i\ j) + S_s(i\ j)\right\} \quad (22)$$

where $S_G(i\ j)$ reflects the distance of their expression pattern. In summary, drug combinations targeting on the disease-specific signaling network, with similar functions, through alternative targets are prioritized.

3 RESULTS

We have applied the BNMFβD algorithm to cluster the drug–drug network into overlapping drug communities. Table 1 lists the parameters in the analysis. We set a as one of {10,50,100,150,200,250,300,350,400,450}, and b to be equal to a. When a and b are set to be 450, the BNMFβD algorithm

Table 2. Synergistic alternative drugs combining with Gefitinib

Drug combination (Community number)	Synergistic score	Rank	Literature evidence
Gefitinib/Paclitaxel (63/55)	2.905	2	PMID:19596955 PMID:14990633
Gefitinib/Celecoxib (63/48)	2.804	3	PMID:18379355 PMID:16914589
Gefitinib/Genistein (63/102)	2.765	4	PMID:22160570
Gefitinib/Fulvestrant (63/55)	2.529	9	PMID:24268810
Gefitinib/Irinotecan (63/66)	2.468	11	PMID:21915126 PMID:16713012
Gefitinib/Vorinostat (63/102)	2.464	12	PMID:21271222
Gefitinib/Lovastatin (63/34)	2.102	27	PMID:19760159
Gefitinib/Rosiglitazome (63/13)	2.023	32	PMID:168386327
Gefitinib/MS-275(63/102)	2.007	34	PMID:16424029

Table 3. Synergistic alternative drugs combining with LY-294002

Drug combination (Community number)	Synergistic score	Rank	Literature evidence
yiLii			

predicted for ER-positive BRCA based on the reconstructed ER-positive breast cancer signaling network.

To validate the prediction results, we searched the literature evidence of the top 50 ranked combinations. Tables 2–4 show the literature evidence of the effective drug combinations in our top 50 lists for lung adenocarcinoma. Surprisingly, 19 different drug combinations have been reported to be synergistic combining with Gefitinib, Paclitaxel and LY-294002 in non–small-cell lung cancer. Tables 5 and 6 show the literature evidence of effective drug combinations in the top 50 lists for ER-positive breast cancer. Also 14 different drug combinations have been reported to be synergistic combining with Tamoxifen and Letrozole. These results show the strong drug combination prediction capacity of DrugComboRanker.

To make the evaluation fair and sound, we further compared the predicted results with CDA, and a random combination method (RCM), which randomly picks up 50 drugs from the available drug lists to combine with the designated drugs. As for CDA, we picked the top 50 drug combinations with the designated drugs. Figures 5 and 6 show the comparison results on the lung adenocarcinoma and ER-positive breast cancer in terms of literature supports of those top-ranked 50 drug combinations, respectively (The numbers on the bars of RCM are standard deviations). For RCM, we repeated the random selection 100 times; for each simulation, we checked the literature evidence (In total, we checked all the 5000 random combinations). As can be seen, the proposed approach outperforms the CDA and random selection significantly.

The predicted drug targets in the disease-specific network could indicate the molecular mechanism of synergistic drug combinations. Here, we map the responsive genes of Gefitinib, Paclitaxel, Vorinostat, LY-294002 and Quercetin to the lung adenocarcinoma-specific signaling network to capture the distinct synergistic responses induced by three agent combinations, Gefitinib and Paclitaxel, LY-294002 and Quercetin, Gefitinib and Vorinostat. As shown in Figure 7, Gefitinib and Paclitaxel combinations can affect the EGFR signaling pathway (endothelial cell proliferation), TP53 signaling pathways, as well as

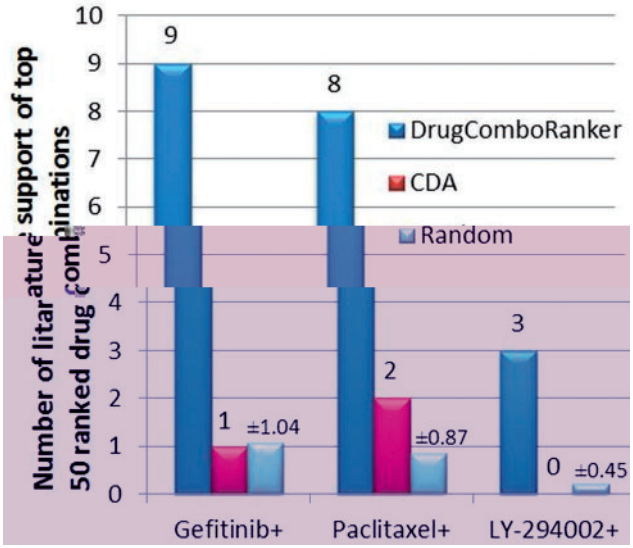

Fig. 5. The comparison results of DrugComboRanker, CDA and RCM in terms of the number of literature supports of the top-ranked 50 drug combinations of lung adenocarcinoma with designated drugs, Gefitinib, Paclitaxel and LY-294002

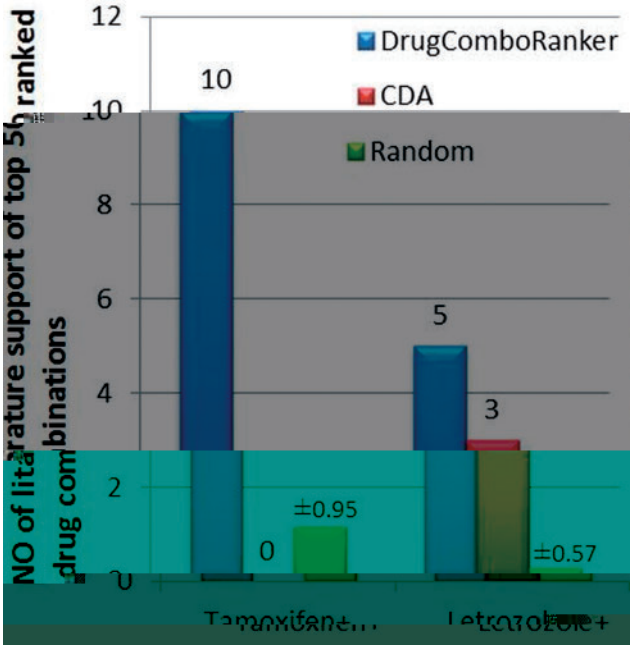

Fig. 6. The comparison results of DrugComboRanker, CDA and RCM in terms of the number of literature supports of the top-ranked 50 drug combinations of ER-positive breast cancer with designated drugs, Tamoxifen and Letrozole

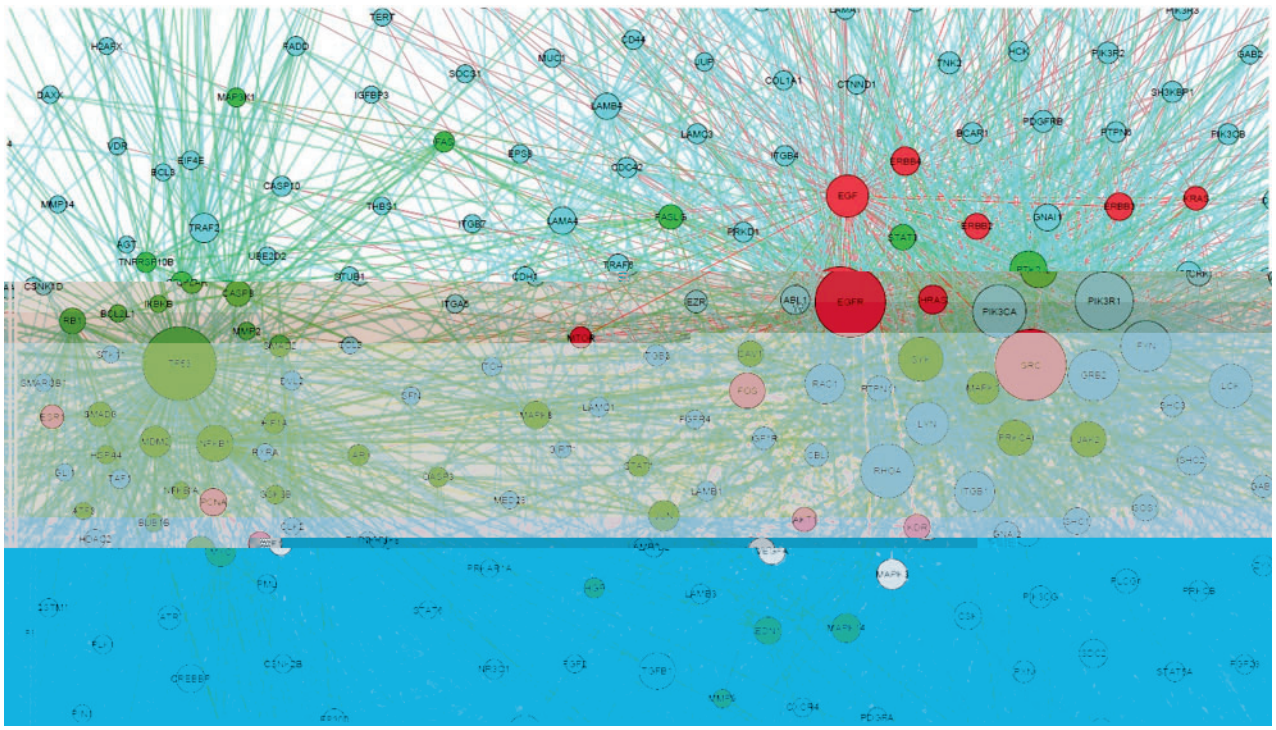

Fig. 7. Drug targets mapped on the disease signaling network. Red and green are the drug targets of Gefitinib and Paclitaxel, respectively

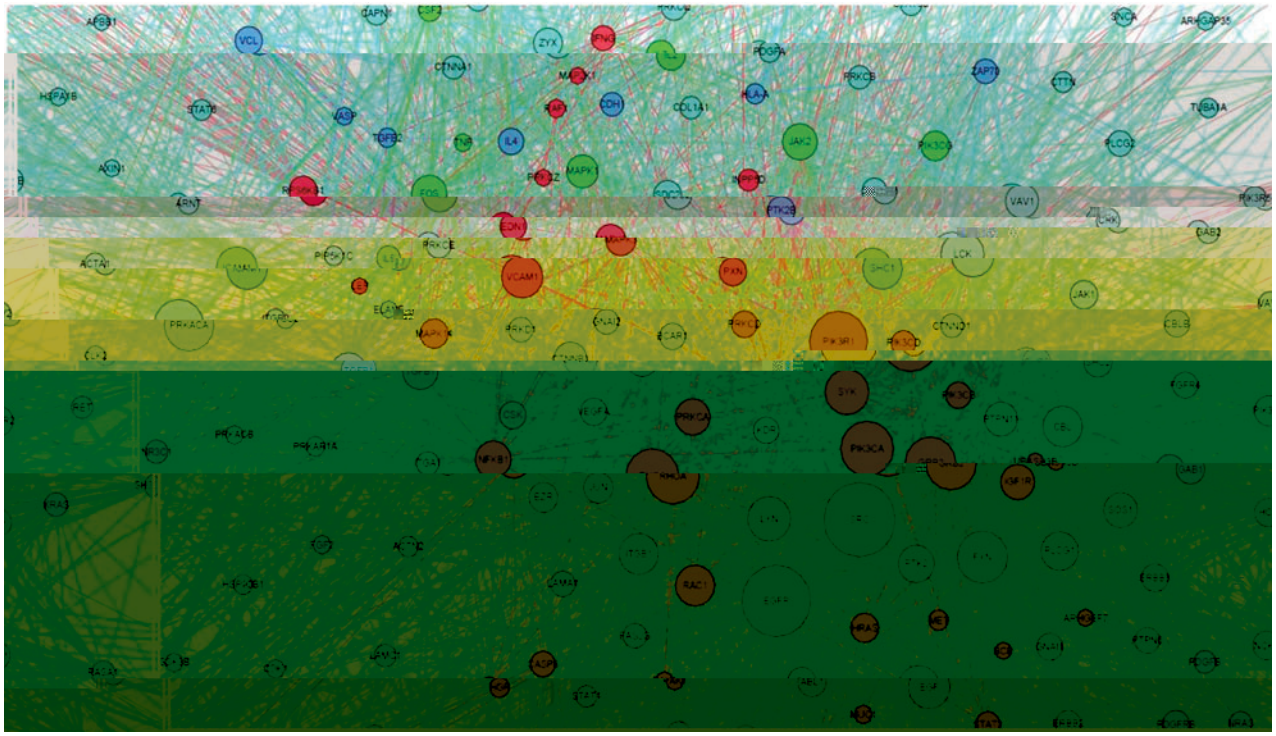

Fig. 8. Drug targets mapped on the disease signaling network. Red and green are the drug targets of LY-294002 and Quercetin, respectively; blue nodes are the weak effected targets of both drugs

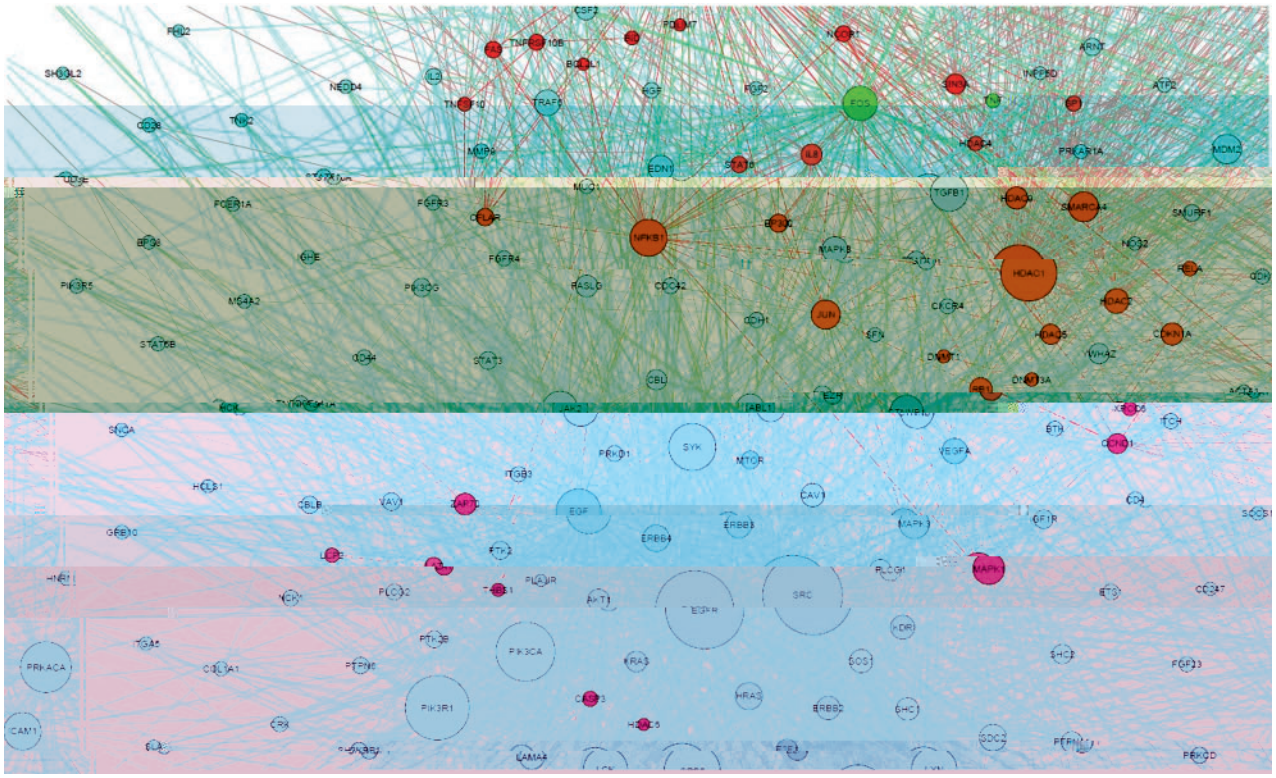

Fig. 9. Drug targets mapped on the disease signaling network. Red and green are the drug targets of Gefitinib and Vorinostat, respectively

biological processes, such as cell cycle, apoptosis and the hub genes, i.e. EGFR, TP53, SRC, FOS, JUN. Distinctly the LY-294002 and Quercetin combination affects the alternative EGFR, PI3K-AKT and JAK-STAT3 pathways, as can be seen in Figure 8. In addition, the drug combination, Gefitinib and Celecoxib, targets the EGFR and COX-2 signaling pathways, respectively. The Gefitinib and Celecoxib have distinct transcriptional responses that indicate EGFR and COX-2 signaling pathways are complementary, and have cross talks. Another example is Gefitinib and Vorinostat combination. Vorinostat is a Histone deacetylases inhibitor, as shown in Figure 9, which interacts with CTNNB1, and CTNNB1 interacts with E-cadherin, ERBB2 and EGFR, whereas Gefitinib targets on EGFR. Thus, this combination forms a double inhibition on growth factors.

4 DISCUSSION AND CONCLUSION