Gaire et al. BMC Bioinformatics 2013, 14(Suppl 2):S7
http://www.biomedcentral.com/1471-2105/14/S2/S7

BMC Bioinformatics

PROCEEDINGS Open Access

Discovery and analysis of consistent active sub-networks in cancers

Raj K Gaire[1,2*], Lorey Smith[3], Patrick Humbert[3], James Bailey[1], Peter J Stuckey[1], Izhak Haviv[4,5]

From The Eleventh Asia Pacific Bioinformatics Conference (APBC 2013)
Vancouver, Canada. 21-24 January 2013

Abstract

Gene expression profiles can show significant changes when genetically diseased cells are compared with non-diseased cells. Biological networks are often used to identify *active subnetworks (ASNs)* of the diseases from the expression profiles to understand the reason behind the observed changes. Current methodologies for discovering ASNs mostly use undirected PPI networks and node centric approaches. This can limit their ability to find the meaningful ASNs when using integrated networks having comprehensive information than the traditional protein-protein interaction networks. Using appropriate scoring functions to assess both genes and their interactions may allow the discovery of better ASNs.

In this paper, we present CASNet, which aims to identify better ASNs using (i) integrated interaction networks (mixed graphs), (ii) directions of regulations of genes, and (iii) combined node and edge scores. We simplify and extend previous methodologies to incorporate edge evaluations and lessen their sensitivity to significance thresholds. We formulate our objective functions using mixed integer programming (MIP) and show that optimal solutions may be obtained.

We compare the ASNs obtained by CASNet and similar other approaches to show that CASNet can often discover more meaningful and stable regulatory ASNs. Our analysis of a breast cancer dataset finds that the positive feedback loops across 7 genes, *AR*, *ESR1*, *MYC*, *E2F2*, *PGR*, *BCL2* and *CCND1* are conserved across the basal/triple negative subtypes in multiple datasets that could potentially explain the aggressive nature of this cancer subtype. Furthermore, comparison of the basal subtype of breast cancer and the mesenchymal subtype of glioblastoma ASNs shows that an ASN in the vicinity of *IL6* is conserved across the two subtypes. This result suggests that subtypes of different cancers can show molecular similarities indicating that the therapeutic approaches in different types of cancers may be shared.

Background

The full genome sequencing of cancer cases demonstrates how remarkably heterogeneous cancer cases are [1]. This heterogeneity is consistent with the hypothesis that most mutations are innocent bystander consequences of the failure of cancer cells' intrinsic mechanism to repair and guard the integrity of the genome [2]. However, the observed heterogeneity of the cancer mutations combined with the knowledge of multiple lesions that all could lead to the same phenotypic consequence [3], leads to a new emerging hypothesis. According to this competing hypothesis, intrinsic subtype specific cancer causing mutations are rare, but their biological output is common [4].

The recognition that *cancer stem cells* within a tumour mass uniquely carry the potential for overt malignancy [5,6] and the discovery that these cells can be transformed into and change forms between epithelial or mesenchymal cells, a phenomena known as epithelial-mesenchymal transformation (EMT) [7], has increased our insight into the link between EMT and fatal cancer phenotypes, such as metastasis and resistance to treatments. In addition, the discovery of intrinsic subtypes of breast cancer that express unique groups of genes [8] has advanced its prognosis. Some intrinsic subtypes of breast cancer are

* Correspondence: rgaire@csse.unimelb.edu.au
[1]NICTA, Victoria Laboratory and Department of Computing and Information Systems, University of Melbourne, Parkville, Vic 3010, Australia
Full list of author information is available at the end of the article

associated with elevated susceptibility to specific drugs, such as Herceptin (for amplified *HER*2 cases) and Tamoxifen (for ER+ cases), while other subtypes, such as the mesenchymal basal/triple negative cases remain without a matching therapeutic strategy. Being able to compare subtypes of different cancers may help identify genes causing the specific subtypes of cancers, leading to identify better therapeutic targets. More importantly, this could provide a scientific basis to sharing therapeutic strategies in subtypes of different cancers.

The task of interpreting gene expression profiles in a disease is not only to differentiate the non-random changes from the random and irrelevant changes, but also to identify the disease causing changes, their downstream effects and the cellular responses related to the disease. If such a procedure worked, one would expect to see the intrinsic signature of luminal breast cancer subtype emerging as downstream to gene *ER*. Furthermore, one could identify driver mutations of the mesenchymal/basal subtype for which therapeutic strategies fail to work. Biological interaction networks contain immense amount of knowledge suitable for such analysis [9]. Finding active subnetworks in diseases is a typical analysis which uses such networks to generate meaningful biological contexts from the differentially expressed genes.

An active subnetwork (ASN) is a subnetwork of a biological interaction network in which the significant nodes obtained from an experiment are connected by edges defined in the network [10]. A methodology for finding ASN was initially proposed by Ideker *et al.* [10]. When coining the idea of ASNs, they established the goal of finding ASNs that could answer questions such as *"What are the signalling and regulatory interactions in control of the observed gene expression changes? How is this control exerted?"* To achieve this, several variations of their work have been proposed [11-17] to analyse differentially expressed genes in diseases using protein-protein interaction (PPI) networks.

PPI networks are undirected networks. Therefore, the ASNs obtained by using these networks can show signalling and regulatory information, but without the directionality of edges, they cannot explicitly show how the controls are exerted. Furthermore, PPI networks have two problems. One, individual interaction networks exhibit little overlap [18], suggesting that a single interaction network might not contain complete information. To overcome this problem, different interaction networks are combined in single integrated interaction network as a mixed graph, containing different types of biological interactions, such as activation, inhibition and post-translational modification. Second, these networks contain both false positive and false negative edges [19], suggesting that the qualities of edges may not be consistent across the network. This problem is solved by computing confidence values of the edges from the sources from which the edge information is obtained from (e.g. number of sources). Alternatively, the values are defined by co-relations of genes in the experimental data [16].

With the integrated networks containing more comprehensive information, one would expect to obtain more informative ASNs by using them with the existing methods. In fact, Deshpande *et al.* [17] used the direction of regulation of genes in multiple species and showed that the identified ASNs are more stable and consistent across multiple species. Similarly, other tools such as IPA (Ingenuity® Systems, http://www.ingenuity.com) show directed edges in their outputs. However, these methods do not use node and edge information together which could produce better ASNs from this type of networks.

We have identified the three issues: node centricity, sensitivity to p-value thresholds and inability to compare ASNs, which limit the existing methods from finding ASNs that could explain how the genetic controls are exerted in diseases (see supp text for details). In addition, the existing ASN finding tools require users to have a copy of the entire network database prior to starting any analysis, which can further affect their usability.

In this paper, we present CASNet (*C*onsistent *A*ctive *Sub*net*works*), our novel methodology that uses (i) an integrated interaction network, STRING [20], (ii) directions of gene regulations, and (iii) combined node and edge evaluations, to find better ASNs. We model the objective function using a mixed integer programming (MIP) model and then solve the model using CPLEX to efficiently discover optimal ASNs. Furthermore, CASNet uses web based APIs to access only relevant parts of the interaction networks and does not require a local copy of the entire database obtained prior to using this tool. We use simulated datasets to show that CASNet can address the above identified limitations. Additionally, we use publicly available datasets to identify and compare ASNs of cancers that provides interesting biological insights. CASNet and supp text of this paper are available at http://www.csse.unimelb.edu.au/~rgaire/CASNet/.

Results and discussion

In this section, firstly we will show that by adding edge information, our method not only selects interaction edges which are consistent with the experimental results, but also helps reduce the sensitivity to p-value significance threshold. Secondly, we will present our analysis of breast cancer and comparison of ASNs in mesenchymal subtypes of breast and glioblastoma cancers.

Comparing with node centric approaches

In order to evaluate CASNet, we used simulated networks of varying nodes and compared our results directly with the result obtained from three different methods

(i) jActiveModule [10], (ii) heinz [15] and (iii) CEZANNE [16]. jActiveModule is a Cytoscape [21] implementation of Ideker *et al.* ASN finding problems being NP hard, it uses the simulated annealing heuristic approach to obtain optimal solutions. In contrast, heinz is an implementation of Dittrich *et al.* which models this problem as a Prize Collecting Steiner Tree problem and uses Integer Programming techniques to obtain the exact solution. Heinz is also available as a Bioconductor package, which was used for this comparison. Finally, CEZANNE is MATISSE [22] module and uses not only the significance of nodes but also the similarity of nodes as well as an assessment of edges based on the correlations of nodes. These three methods cover a broad range of techniques that are currently available for finding ASNs. By comparing our results with the results obtained from these methods, we can understand the performance of our method against other similar methods. In addition to this, we also compared the results obtained by using our method disregarding the directionality of edges to understand the performance differences attributed to the directionality. These experiments were performed for different network sizes to assess the robustness of our method. At this point, we note that some approaches have been developed that use co-expression [23], co-variance [14] and correlation [16] of genes for creating or assessing edges in the networks. Although useful, these methods may not be applicable when only a list of genes is available.

We used the following parameters of the individual methods: jActiveModule was used with the default parameters to obtain only 1 module. For heinz, a false discovery rate (FDR) of 0.05 was used (more stringent FDR thresholds had poor performance). For two nodes a and b, CEZANNE required node similarity scores matrix, $sim[a][b]$. This was calculated from p-values of the nodes, p_a and p_a, as $sim[a][b] = sim[b][a] = min(p_a, p_b)/max(p_a, p_b)$ such that the nodes having similar p-values obtain high similarity scores (≈1) while the nodes having different p-values obtain low similarity scores (≈0). A p-value threshold of 0.05 was used for both CEZANNE and CASNet. Precisions (true positive/all classified as positive) and recalls (true positive/all positives) for both nodes and edges were added for each methods for comparisons.

Figure 1 illustrates the performance of different methods. It shows that jActiveModule finds modules with low precision and high recall. Therefore, the modules obtained by this method are often bigger and may contain large number of false positives. In contrast, heinz produces modules with high precision but low recall, hence discards many true positive nodes and edges. CEZANNE produced ASNs with better recall, but worse precision than heinz. Note that CEZANNE and jActiveModule failed to find modules in small and large networks respectively.

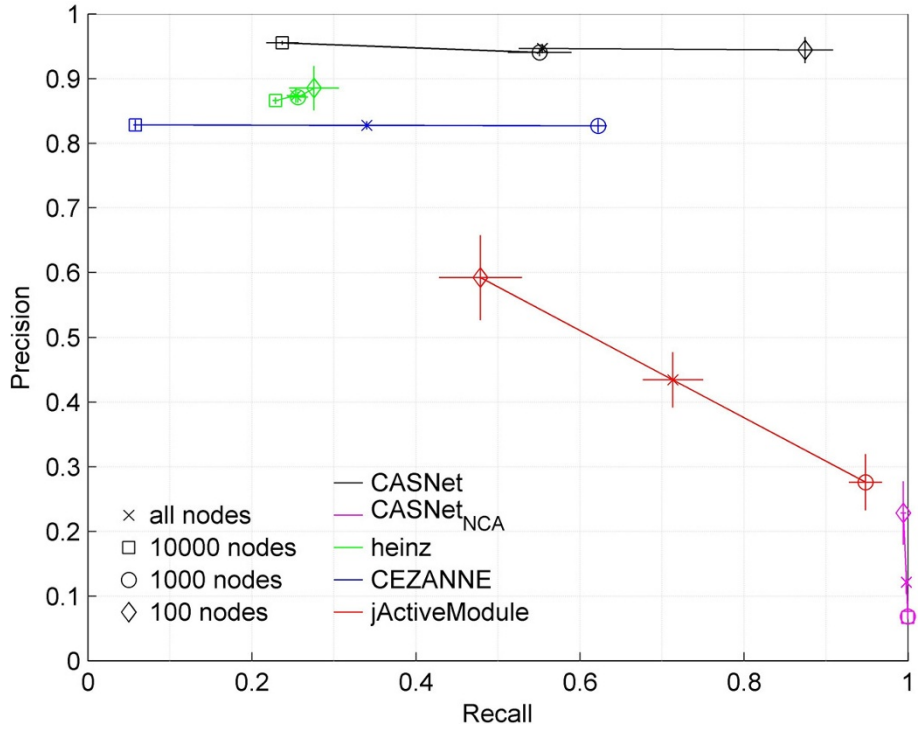

Figure 1 Comparison of ASN finding methods in simulated datasets.

Additionally, jActiveModule produced inconsistent precision and recall whereas CEZANNE produced consistent precision but inconsistent recall for different network sizes. In contrast, heinz performed consistently, though with lower recall rates.

Since CASNet relaxes the significance threshold, it starts by selecting a large number of nodes. The edges connecting these nodes are then assessed based on their directionality and confidence scores. This is also apparent in Figure 1. Here, CASNet without considering directionality of edges ($CASNet_{NCA}$) performs with high precision but low recalls. The recall in CASNet is highly improved by using edge information. This demonstrates that by using additional information of edges, the quality of ASNs can be dramatically improved (see Supp text for further comparisons).

Analysis of cancer datasets

In a recent study, Iliopoulos *et al.* [24] noted that the genes related to inflammatory pathways form stable PFLs regulated by *IL6* causing continuous progression of cancer. Their study involved extensive laboratory based works. By using computational methods instead, we can potentially not only reduce costs and efforts of identifying PFLs, but also identify novel PFLs.

ASNs and PFLs in breast cancer

We used a publicly available breast cancer dataset to explore ASNs and PFLs in the basal subtype of breast cancer. SAM analysis [25] was performed to obtain a list of significant genes. This list was used with STRING network [20] to find ASNs and PFLs.

Figure 2 is the ASN obtained from the gene list of basal breast cancer. This ASN contains PFLs as shown in Figure 3. Besides other genes, we found that the PFLs across *MYC*, *E2F1*, *AR*, *ESR1*, *CCND1*, *PGR* and *BCL2* were conserved across independent breast cancer dataset as shown in Figure 4. However, we did not find any PFL when the differentially expressed genes from the entire dataset without discriminating the cancer subtypes were used. *ESR* (i.e. *ER*) is one of the genes which differentiates luminal and basal subtypes [8]. The breast cancer patients with low *ER* expression levels have poorly survival rates. Our identification of these PFL in ER- samples with oncogene like *MYC* and tumour suppressor gene like *CCND1* with the breast cancer discriminating gene *ESR1* is a novel finding. Existence of such PFLs probably explains the reasons behind the resistance to the therapeutic in this cancer subtype.

In order to obtain independent evidence for the CASNet choice of expression-based nodes and edges, we used the TCGA query portal [26] to find genes included in the ASN which had consistent expression and copy number changes. Here, we assumed that a gene could be causal if it is not affected by other genes and have mRNA expression level changes consistent with the copy numbers of the genes. *FOXA1*, *NCOA7* and *DOCK7* were the top three genes in this result. Since *DOCK7* is a downstream signalling intermediate of *ERBB2*, identification of *DOCK7* aberrations in the absence of *HER2* over-expression may implicate *DOCK7* in Transtuzumab drug resistance. Moreover, when *FOXA1* and *NOCA7* were considered with the PFL forming genes, all the samples had at least one gene that had consistently changed expression levels. This further suggests that PFLs and their neighbouring genes could be important in understanding the nature of complex cancer cases.

ASNs in glioblastoma vs breast cancer

Here, we used Glioblastoma (GBM) dataset from Verhaak *et al.* [27] to obtain an ASN for mesenchymal subtype cases and compare it with the ER- breast cancer ASN, since both of these subtypes of cancers show mesenchymal signatures, have poor patient survival and are likely to be related to the EMT event.

Figure 5 shows the similar components of ASN in these subtypes of cancers. Here, *PPARG*, *SREBF2*, *E2F1*, *MYBL2* and *MYCN* are the genes which are differently regulated while several other genes are similarly regulated in the two cancer subtypes. *PPARG* is up-regulated in mesenchymal GBM, but down-regulated in ER- breast cancer. This gene can both enhance and suppress the expression of *PTGS2* and *ICAM1*, which are up-regulated in both subtypes. *PTGS2* (*COX2*, an aspirin target) is a key mediator of inflammation [28]. This behaviour of *PPARG* is likely expressed by stromal adipocytes (fat cells), which are known to accentuate inflammatory processes in a number of human pathologies, including cancer.

IL6 is the most connected node in this similarity network. It is associated with acute inflammation, suggesting that higher expression of *IL6* could be a cellular response to inflammation. *MYC*, which is found to be forming PFLs in basal/triple negative breast cancer cases, is seen to be regulated by *IL6*, *IL1B*, *IFNG* and *HIF1A*, and is conserved in both mesenchymal GBM and ER- breast cancer. This independently confirming the role of *IL6* in cancers and suggests that maintaining high level of inflammation may be a conserved feature of mesenchymal subtypes of cancers. More generally, our finding of the common pathways in the different subtypes of cancers suggests that even though the genetic signatures among different cancers may not be similar, the cancer subtypes might not only have similarities in their genetic signatures but also have similarly affected pathways and could potentially be treated in the same manners.

Materials and methods

In this section, we describe the datasets and our methodology for finding ASNs.

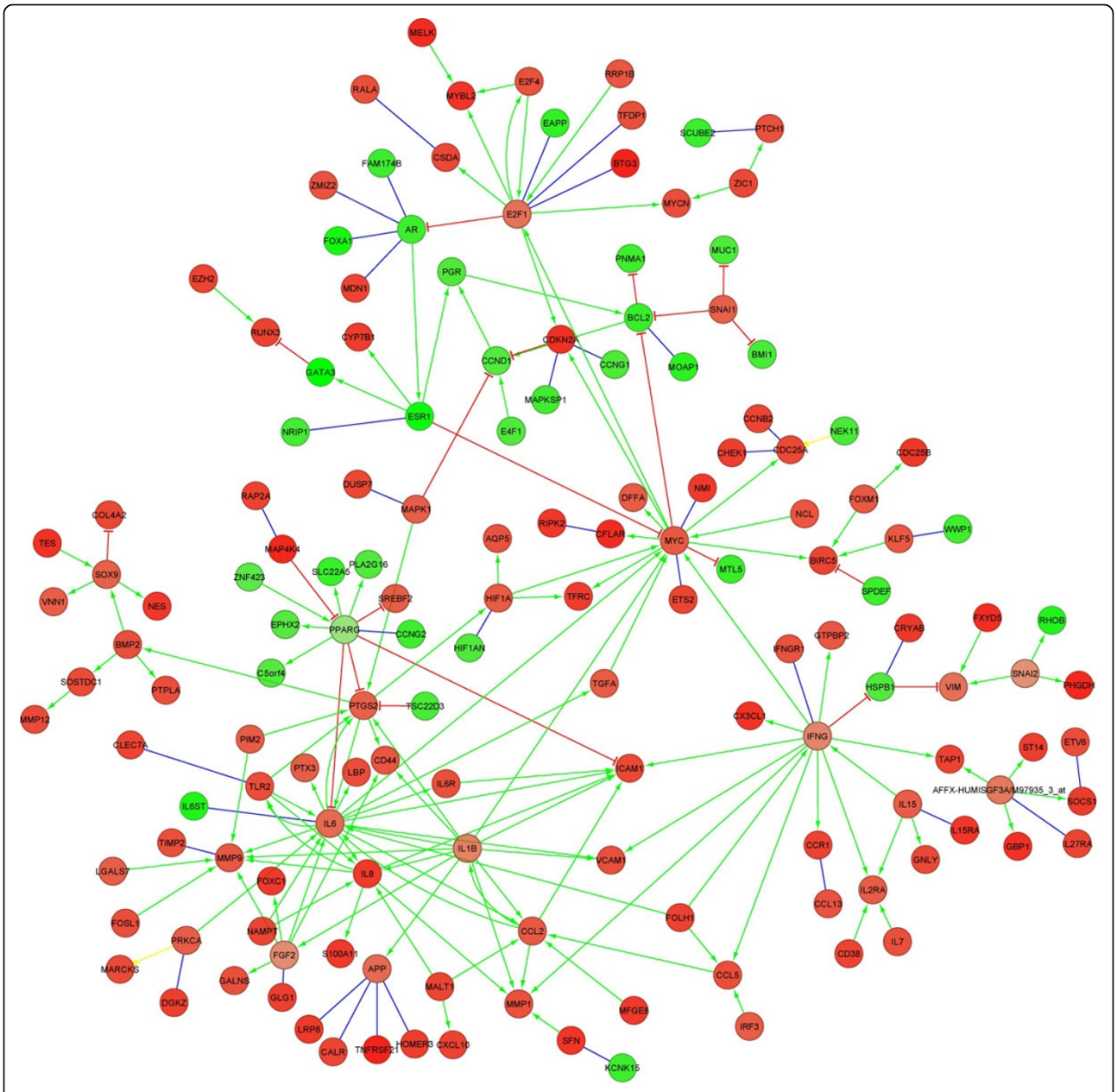

Figure 2 ASN in ER- subtype of breast cancer. Here, the red and green circles represent the genes that are respectively over- and under-expressed in ER-cases.

Simulated networks

Here, we created simulated experimental datasets and consistent gold standard ASNs. These ASNs were combined to obtain simulated network. The experimental datasets and the simulated networks were then used to obtain ASNs from different methods and compared against the gold standard ASNs. More specifically, we fist created nodes of sizes N = 100; 1000; 10000. p-values were then assigned to these nodes such that n = 0:1 x N randomly selected nodes obtained values smaller than 0:001 (considered as the significant nodes), while other nodes were assigned uniformly distributed values between 0 and 1. 2 x n random pairs of significant nodes were assigned directed up- and down-regulating edges consistent with the direction of regulation of nodes. The confidence scores of all the edges were assigned a constant value, 1. Each of the above experiment with different node sizes was repeated for 10; 15 and 20 times by randomly reassigning p-values but changing the directions of regulations of a fraction (0:5) of significant nodes to add randomness in

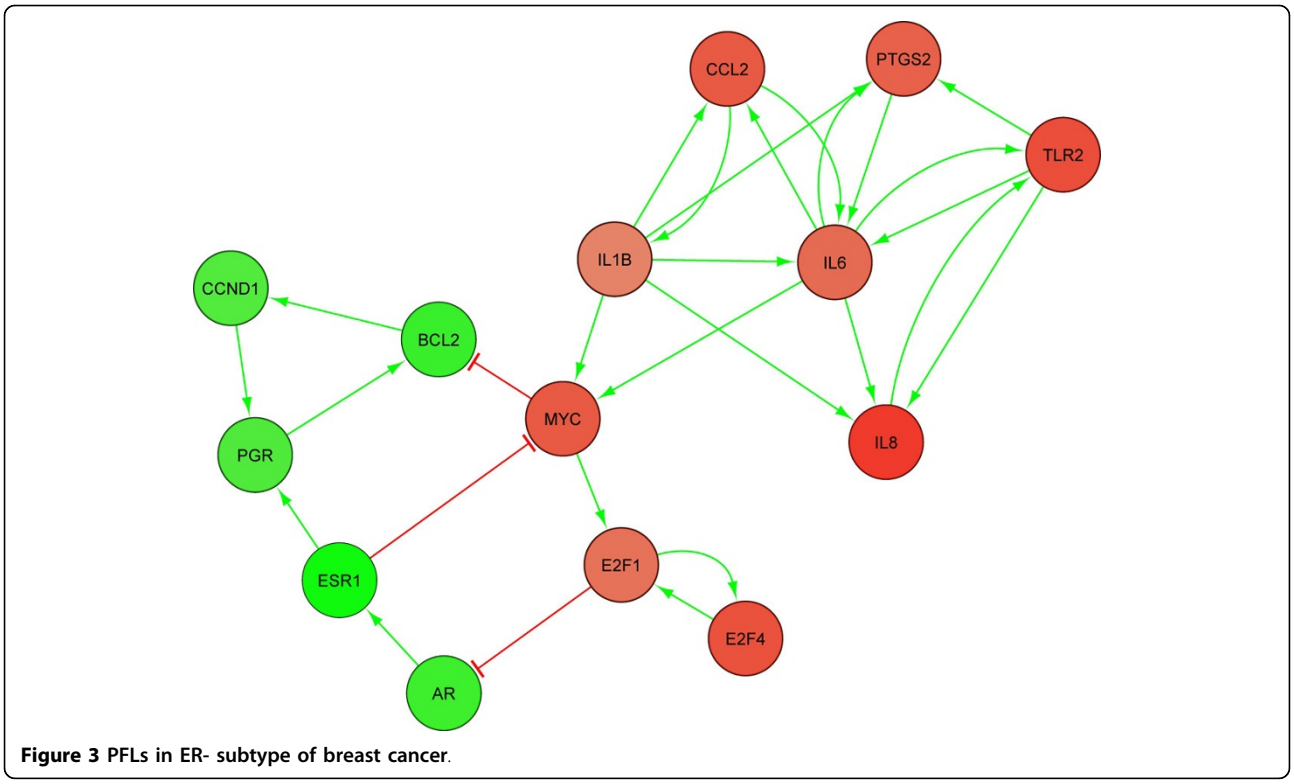

Figure 3 PFLs in ER- subtype of breast cancer.

the experimental results. Since the edges are reassigned in each experimental data, this variation in node regulation can create different edges among the same two nodes as found in real biological networks.

Biological network data

We used the action networks of STRING [20] as our network data source. The network was access at run time via the publicly available web based APIs. The benefits of this approach are that downloading and maintaining the entire database is not required, making CASNet usable in a computer with internet connection. The assessed parts of STRING were saved locally for future use, thereby avoiding excessive internet usage.

Experiment data

The breast cancer samples GSE2034 [29,30] were obtained from obtained from GEO [31]. The GEO dataset contained 282 samples, with 206 were ER+ and 75 ER- cases. The TCGA BRCA dataset contained 465 samples. TCGA [32] (BRCA) dataset was used as an independent validation dataset.

The GBM dataset was obtained from a TCGA publication [27] containing 206 samples. It categorised the samples into 4 subtypes: Proneural (PN), Neural (NL), Classical (CL) and Mesenchymal (MES) based on their genetic signatures. Their SAM analysis [25] results of MES subtype were taken as a basis for the significance measurement of the genes (see supp text for details).

Interaction networks

Let V_{exp} be a set of biological molecules being investigated in an experiment. The differential expression analysis finds molecules $M = (V_{exp}, P, D)$ with p-value significances P, and directions of regulations (up- or down-regulation) D.

A network (or graph) $G = (V, E)$ consists of vertexes (or nodes) V which are connected by edges E. Given two nodes $v_1, v_2 \in V$, and edge $e \in E$; $e = (v_1, v_2, t, c)$ connects the nodes v_1 and v_2 by an edge with a type t and a confidence value c. The confidence value of an edge is in the range of (0, 1] where 0 is the least confidence and 1 is the most confidence value. Two nodes may be connected by multiple edges with different values of t. The following four types of edges (t) are defined between nodes: (i) physical binding of two nodes, denoted by v_1 - v_2, (ii) catalytic and post-transcription modification of a node by another node, denoted by $v_1 \multimap v_2$, (iii) activation of a node by another node, denoted by $v_1 \rightarrow v_2$, and (iv) inhibition of a node by another node, denoted by $v_1 \dashv v_2$. For an edge e, a confidence value c is defined. Now, the problem of finding an ASN can be stated as follows:

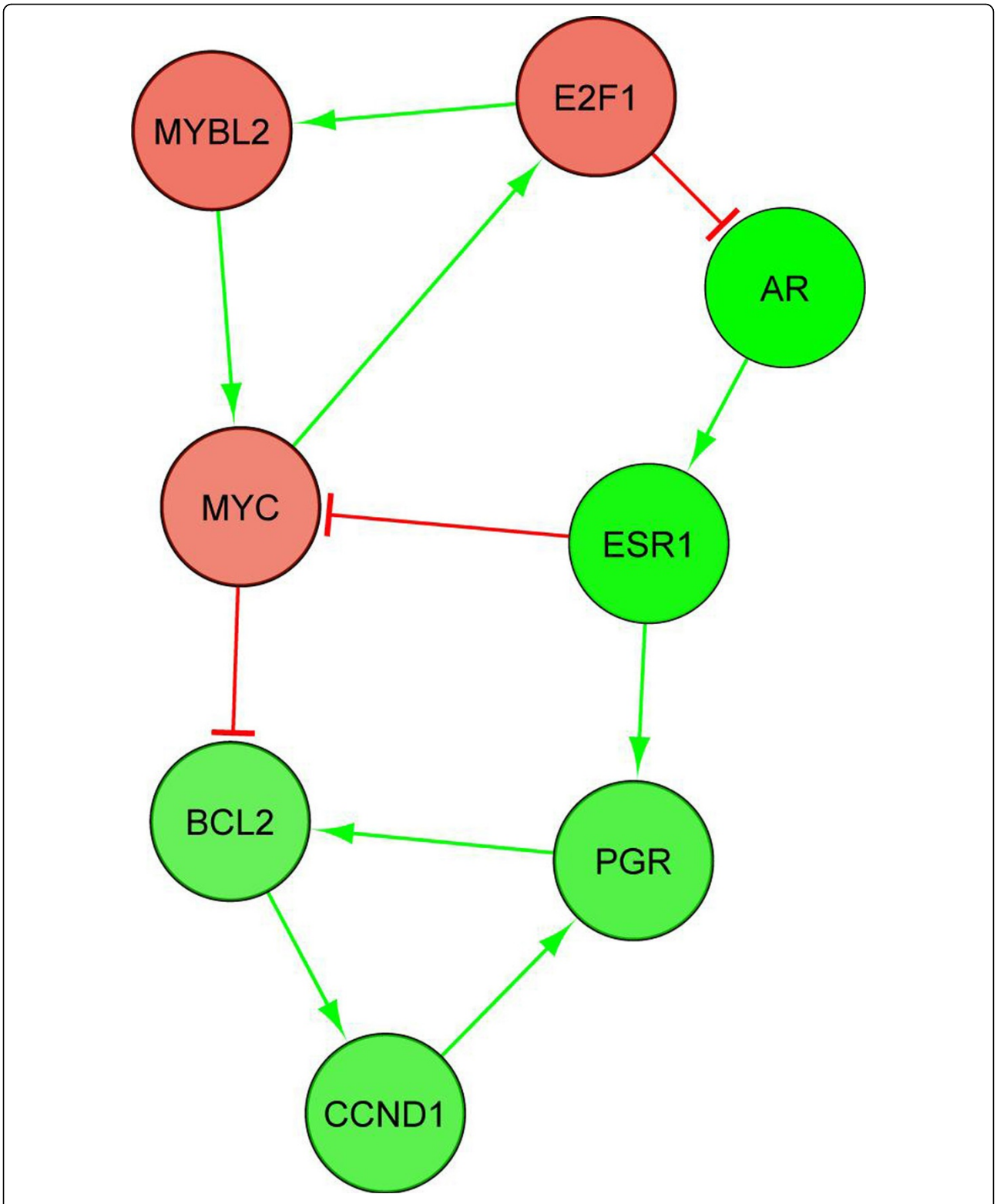

Figure 4 Conserved PFLs in basal/triple negative subtype of breast cancer in two independent datasets.

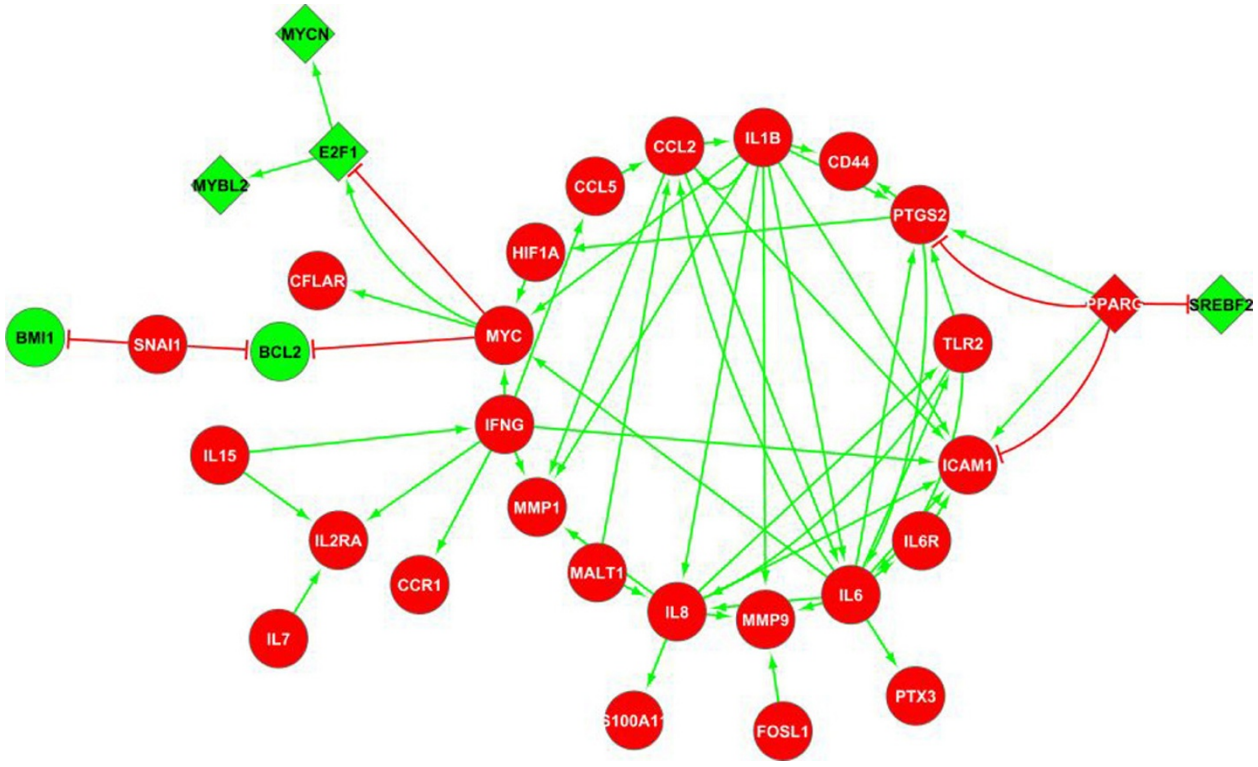

Figure 5 Conserved subnetwork in mesenchymal glioblastoma and ER- breast cancers. Here, the red and green circles represent the genes that are respectively over- and under-expressed in the subtypes of both cancers. The red and green diamonds represent the genes which are respectively over-expressed in glioblastoma but low in ER- cases and vise-versa. It shows that not only the inflammation regulating genes such as *IL6* and *IL8*, but also the mesenchymal marker genes such as *CD44* and *SNAI1* are conserved across the two subtypes.

Problem statement

Given an interaction network, $G = (V, E)$, and a differentially expressed (DE) gene dataset, $M = (V_{exp}, P, D)$, find the best (where the notion of best is defined by scoring functions) subnetwork $G' = (V', E') \subset G$ which connects the highly DE genes $v \in V'$, $V_{exp} \subset V$ with edges $E' \subset E$ that are consistent with the data.

Consistent edge

An edge $e = (n_1, n_2, t, c) \in E$ between two nodes $n_1, n_2 \in V$ is considered to be consistent if the direction of regulation of gene n_2 can be explained by the direction of regulation of n_1 by using the edge type t with a high confidence c.

For example, if n_1 promotes n_2, i.e. $n_1 \rightarrow n_2$, and both n_1 and n_2 are up- or down-regulated, then the edge $e = (n_1, n_2, \rightarrow, c)$ is considered to be consistent with the data with a confidence c. In contrast, if a node n_1 is up-regulated and the other node n_2 is down-regulated and vice-versa, then the edge is considered to be inconsistent with the data.

Based on the problem statement, we now define the node, edge and subnetwork scores that will be used to evaluate ASNs.

Node scores

In an experiment, DE genes are often associated with p-values obtained from some statistical tests. A subnetwork with nodes having low p-values is a desirable feature of an ASN [10]. Additionally, it is often desirable to exclude highly connected nodes in an ASN [33-35]. Here we adopt Dittrich *et al*'s [15] node scoring method to include low p-values in the networks (profit). At the same time, we use a simple approach to penalise highly connected nodes (cost).

Let p_n be the p-value of a node n. Then the node's profit score W_n is given as [15],

$$W_n = (a - 1) \times (log(p_n) - log(\tau)) \quad (1)$$

where $a = (0, 1]$ is a shape parameter of the beta distribution fitted for a dataset representing a signal to noise ratio, and τ is a threshold to control the size of the ASN (interpreted as the false discovery rate). In this case, the value of $(a - 1)$ acts as a scaling factor. This factor is useful to compare the scores of a node obtained from different datasets having different p-value distributions. On the other hand, τ can be a selected threshold which can make a node score positive or

negative, and thereby controls the size of the ASNs. Since the value of a does not affect the sign of the scores, and if a single p-value dataset is used, a can be assigned a constant value without affecting the resulting ASN. Practically, a is close to 0. Therefore, the node score can be simplified by assigning $a = 0$ as $W_n = -1 \times (log(p_n) - log(\tau))$. Furthermore, if a set of significant genes is used in an experiment after applying a p-value threshold, the Eq. 1 can be further reduced to, $W_n = 1 \times log(p_n)$ for all the genes in the set and $W_n = -1 \times |Constant|$ for the genes not in the list.

If D is the degree of a node n (i.e. it is connected to D other nodes), we assign a cost C_n to the node to penalise highly connected nodes, as, $C_n = log(D)$.

Since the profit and the cost scores of a node do not have the same scale, we scale these values to a range of [-1, 1] to obtain *standard* scores as, $W'_n = \frac{W_n}{max_{m \in V}(|W_m|)}$; $C'_n = \frac{C_n}{max_{m \in V}(|C_m|)}$

Edge scores

It is desirable to assign a high *positive* score to an edge which is consistent with the data, so that including such edges will increase the subnetwork scores. Similarly, the inconsistent edges should be penalised by assigning *negative* scores to them. Additionally, the confidence value of an edge should be used to scale these scores. Therefore, we use the following scheme to obtain the edge scores:

1. If n_1 promotes n_2, then the consistency score W_e is equal to (i) 2 if both n_1 and n_2 are changed in the same direction, (ii) -1 if either n_1 or n_2 is unchanged and (iii) -2, if n_1 and n_2 are changed in opposite direction.
2. Similarly, if n_1 suppresses n_2, then the consistency score W_e is equal to (i) 2 if n_1 and n_2 are changed in an opposite direction, (ii) -1 if only one of n_1 and n_2 is changed and (iii) -2, if n_1 and n_2 are changed in the same direction.
3. For edges $n_1 \multimap n_2$ and n_1 - n_2, a constant value can be assigned depending on whether including such edges is desirable or not. For example, $W_e = 1$ could be used in PPI networks, while $W_e = -1$ could be used in the networks where having these edges is not desirable. By default, we use $W_e = -1$ to lessen emphasis on undirected edges.

Now, an edge score is defined as, $S_e = W_e \times C_e$ where $C_e = (0, 1]$ is the confidence score of the edge obtained from the interaction network.

Finally, the *standard* edge score is obtained as, $S'_e = \frac{S_e}{max_{f \in E}|S_f|}$

Sub-network score

Based on the standard node and edge scores, the score of a subnetwork G'-is obtained by-using the linear equation,

$$S = \sum_{n \in V} x_n \times (\alpha \times W'_n - \beta_n \times C'_n) + \gamma \times \sum_{e \in E} x_e \times S'_e \quad (2)$$

where α, β and γ are the scaling factors of the node weight profits, the node connectivity costs and the edge consistency scores respectively and x_n and x_e are boolean variables with values 1 if $n \in V'$, $e \in E'$ and 0 otherwise. The values of these scaling factors could be obtained by using gold standard network and experiment datasets. In the absence of such datasets, we use $\alpha = \gamma = 1$ and $\beta = 0$ for the *positive* scoring nodes and $\beta = 1$ for *negative* scoring nodes in our experiments. This is because a large number of edges around cancer related genes exist in the biological networks, since a high number experiments have been performed in those genes. Penalising them at the same rate as others eliminates the highly DE genes (results not shown).

The objective function for finding ASNs is to obtain a sub-network which maximises the subnetwork score S.

The MIP model

Here, we model the problem of finding an ASN by using the mixed integer linear programming (MIP) model in CPLEX which maximises the objective function in Eq. 2. x_n and x_e are defined as boolean variables (i.e. $x \in 0, 1$). Further to this, the following additional constraints are imposed: (a) $x_{n(i)} \rightarrow \exists x_{e(i,\ j)}$ i.e. $x_{n(i)} \leq \Sigma_j x_{e(i,\ j)}$. (b) $x_{e(i,\ j)} \rightarrow x_{n(i)}$; i.e. $x_{e(i,\ j)} \leq x_{n(i)}$. (c) $x_{e(i,\ j)} \rightarrow x_{n(i)}$; i.e. $x_{e(i,\ j)} \leq x_{n(i)}$. where $n(i)$ is the i^{th} node and $e(i, j)$ is an edge connecting the nodes $n(i)$ and $n(j)$.

Conclusion

A large number of datasets that are currently being produced, such as TCGA and ICGC, include definitive genome wide mutational status of many samples, making the task of interpreting the results, and identifying common features even more formidable. Adapting to these high dimensional datasets, biological interaction databases are being integrated into single databases to provide more comprehensive information. Since all the interactions among the nodes might not be active at the same point of time or environmental conditions, the current methodologies of enrichment assessment for candidate networks or pathways can fall short in their ability to discriminate between real biological inferences from false positive ones. Better methodologies are required to use these networks and systematically find interesting and meaningful interactions in disease conditions.

In this paper, we presented a methodology to analyse gene expression results of diseases with STRING network, to not only find a connected subset of nodes that are observed to be highly differentially expressed, but also use the edges in the network to generate hypotheses regarding the reason behind the observed changes. Our methodology, CASNet, enhances existing methodologies by introducing edge scores to solve node centricity, p-value sensitivity problems and subnetwork comparability problems.

We demonstrated that complicated regulatory ASNs and PFLs exist in low surviving cancer cases such as the mesenchymal subtype of GBM and the basal/triple negative subtype of breast cancer. Finally, we showed that by comparing ASNs of different disease types, molecular similarities of the disease can be identified that can be useful in their treatments. In this way, CASNet has widened the possibilities of network analysis in generating biologically significant hypotheses and directing the future researches.

Limitations

Firstly, the current state of directed interaction network has low coverage. Additionally, a large number of edges around the genes of well studied diseases, including cancer, exist due to the large number of experiments that have been carried out with those genes. This creates a bias toward some parts of networks. The parts of network where directionality information exists are most likely the pathways that are already well-understood. As such, the discovery of new genes and their interactions from these networks may be less likely. Secondly, we found several instances where the edges are incorrectly defined in STRING. The edges in our ASNs are only as good as the curation and literature mining methodologies used in creating the networks. Finally, the interaction networks do not differentiate wild type/mutant and active/inactive molecules in their nodes. In the absence of this level of sensitivity in the existing biological networks, precise conclusions cannot be made. Consequently, the results obtained from ASN finding methods are susceptible to the problems associated with the underlying networks and the datasets being used, and hence require independent validations.

Authors' contributions

RKG, JB, PJS, IH - conception, design, development, data acquisition, analysis, interpretation of results, draft manuscript; LS, PH - conception, preliminary interpretation of results.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

This work is supported by National ICT Australia (NICTA), which is funded by the Australian Government's Backing Australia's Ability initiatives, in part through the Australian Research Council (ARC), Komen for the cure, National Health and Medical Research Council (NHMRC) Cancer Australia, Nation Breast Cancer Foundation (NBCF) and Cancer Council Victoria (CCV).

Declarations

The publication costs for this article were funded by the University of Melbourne.

This article has been published as part of *BMC Bioinformatics* Volume 14 Supplement 2, 2013: Selected articles from the Eleventh Asia Pacific Bioinformatics Conference (APBC 2013): Bioinformatics. The full contents of the supplement are available online at http://www.biomedcentral.com/bmcbioinformatics/supplements/14/S2.

Author details

[1]NICTA, Victoria Laboratory and Department of Computing and Information Systems, University of Melbourne, Parkville, Vic 3010, Australia. [2]Metabolomics, Population Studies and Profiling, Baker IDI Heart and Diabetes Institute, Melbourne, Vic 3004, Australia. [3]Cell Cycle & Cancer Genetics, Peter MacCallum Cancer Centre, Melbourne, Vic 3002, Australia. [4]Department of Pathology, School of Medicine, University of Melbourne, Parkville, Vic 3010, Australia. [5]Faculty of Medicine in Galilee, Bar Ilan University, Israel.

Published: 21 January 2013

References

1. McDermott U, Downing JR, Stratton MR: **Genomics and the Continuum of Cancer Care.** *N Engl J Med* 2011, **364(January 27)**:340-50.
2. Lai Y, Eckenrode SE, She J: **A statistical framework for integrating two microarray data sets in differencial expression analysis.** *BMC Bioinformatics* 2009, **10(Suppl 1)**:S23.
3. Wood LD, Parsons DW, Jones S, Lin J, Sjoblom T, Leary RJ, Shen D, Boca SM, Barber T, Ptak J, Silliman N, Szabo S, Dezso Z, Ustyanksky V, Nikolskaya T, Nikolsky Y, Karchin R, Wilson PA, Kaminker JS, Zhang Z, Croshaw R, Willis J, Dawson D, Shipitsin M, Willson JKV, Sukumar S, Polyak K, Park BH, Pethiyagoda CL, Pant PVK, Ballinger DG, Sparks AB, Hartigan J, Smith DR, Suh E, Papadopoulos N, Buckhaults P, Markowitz SD, Parmigiani G, Kinzler KW, Velculescu VE, Vogelstein B: **The Genomic Landscapes of Human Breast and Colorectal Cancers.** *Science* 2007, **318(5853)**:1108-13.
4. Bignell GR, Greenman CD, Davies H, Butler AP, Edkins S, Andrews JM, Buck G, Chen L, Beare D, Latimer C, Widaa S, Hinton J, Fahey C, Fu B, Swamy S, Dalgliesh GL, Teh BT, Deloukas P, Yang F, Campbell PJ, Futreal PA, Stratton MR: **Signatures of mutation and selection in the cancer genome.** *Nature* 2010, **463(18 February)**:893-8.
5. Jordan CT, Guzman ML, Noble M: **Cancer Stem Cells.** *N Engl J Med* 2006, **355**:1253-61.
6. Reya T, Morrison SJ, Clarke MF, Weissman IL: **Stem cells, cancer, and cancer stem cells.** *Nature* 2001, **414**:105-11.
7. Thiery JP, Acloque H, Huang RY, Nieto MA: **Epithelial-Mesenchymal Transitions in Development and Disease.** *Cell* 2009, **139(5)**:871-90.
8. Perou CM, Jeffrey SS, van de Rijn M, Rees CA, Eisen MB, Ross DT, Pergamenschikov A, Williams CF, Zhu SX, Lee JCF, Lashkari D, Shalon D, Brown PO, Botstein D: **Distinctive gene expression patterns in human mammary epithelial cells and breast cancers.** *PNAS* 1999, **96(16)**:9212-17.
9. Ideker T, Sharan R: **Protein networks in disease.** *Genome Res* 2008, **18(4)**:644-52.
10. Ideker T, Ozier O, Schwikowski B, Siegel AF: **Discovering regulatory and signaling circuits in molecular interaction networks.** *Bioinformatics* 2002, **18**:S233-40.
11. Sohler F, Hanisch D, Zimmer R: **New methods for joint analysis of biological networks and expression data.** *Bioinformatics* 2004, **20**:1517-21.
12. Rajagopalan D, Agarwal P: **Inferring pathways from gene lists using a literature-derived network of biological relationships.** *Bioinformatics* 2005, **21(6)**:788-793.
13. Scott MS, Perkins T, Bunnell S, Pepin F, Thomas DY, Hallett M: **Identifying Regulatory Subnetworks for a Set of Genes.** *Molecular & Cellular Proteomics* 2005, **4**:683-692.
14. Guo Z, Li Y, Gong X, Yao C, Ma W, Wang D, Li Y, Zhu J, Zhang M, Yang D, Wang J: **Edge-based scoring and searching method for identifying condition-responsive protein-protein interaction sub-network.** *Bioinformatics* 2007, **23(16)**:2121-2128.

15. Dittrich MT, Klau GW, Rosenwald A, Dandekar T, Müller T: **Identifying functional modules in protein-protein interaction networks: an integrated exact approach.** *Bioinformatics* 2008, **24(ISMB 2008)**:i223-i231.
16. Ulitsky I, Shamir R: **Identifying functional modules using expression profiles and confidence-scored protein interactions.** *Bioinformatics* 2009, **25(9)**:1158-64.
17. Deshpande R, Sharma S, Verfaillie CM, Hu WS, Myers CL: **A scalable approach for discovering conserved active subnetworks across species.** *PLoS Computational Biology* 2010, **6(12)**:e1001028.
18. Prieto C, Rivas JDL: **APID: Agile Protein Interaction DataAnalyzer.** *Nucleic Acids Research* 2006, **34(Web Server)**:W298-W302.
19. Finley R: In *A Guide to Yeast Two-Hybrid Experiments. Volume 24.* Cambridge, MA: Cell Press; 2007:17-21, chap. 5.
20. Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, Muller J, Doerks T, Julien P, Roth A, Simonovic M, Bork P, von Mering C: **STRING 8-a global view on proteins and their functional interactions in 630 organisms.** *Nucleic Acids Res* 2009, **37**:D412-D416.
21. Shannon P, Markiel A, Ozier O, Baliga N, Wang J, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: a software environment for integrated models of biomolecular interaction networks.** *Genome Research* 2003, **13(11)**:2498-504.
22. Ulitsky I, Shamir R: **Identification of functional modules using network topology and high-throughput data.** *BMC Syst Biol* 2007, **1**:8, open access.
23. Zhang B, Horvath S: **A General Framework for Weighted Gene Co-Expression Network Analysis.** *Stat Appl Genet Mol Biol* 2005, **4**:Article17.
24. Iliopoulos D, Hirsch HA, Struhl K: **An Epigenetic Switch Involving NF-kB, Lin28, Let-7 MicroRNA, and IL6 Links Inammation to Cell Transformation.** *Cell* 2009, **139**:693-706.
25. Tusher VG, Tibshirani R, Chu G: **Significance analysis of microarrays applied to ionizing radiation response.** *PNAS* 2001, **98**:5116-21.
26. The Cancer Genome Atlas Group: **The Cancer Genome Atlas data browser.** 2011 [http://tcga-portal.nci.nih.gov/tcga-portal/AnomalySearch.jsp].
27. Verhaak RG, Hoadley KA, Purdom E, Wang V, Qi Y, Wilkerson MD, Miller CR, Ding L, Golub T, Mesirov JP, Alexe G, Lawrence M, O'Kelly M, Tamayo P, Weir BA, Gabriel S, Winckler W, Gupta S, Jakkula L, Feiler HS, Hodgson JG, James CD, Sarkaria JN, Brennan C, Kahn A, Spellman PT, Wilson RK, Speed TP, Gray JW, Meyerson M, Getz G, Perou CM, Hayes DN: **Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in PDGFRA, IDH1, EGFR, and NF1.** *Cancer Cell* 2010, **17**:98-110.
28. The UniProt Consortium: **The Universal Protein Resource.** *Nucl Acids Res* 2010, **38(4)**:D142-8.
29. Wang Y, Klijn JG, Zhang Y, Sieuwerts AM, Look MP, Yang F, Talantov D, Timmermans M, van Gelder MEM, Yu J, Jatkoe T, Berns EM, Atkins D, Foekens JA: **Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer.** *Lancet* 2005, **365(9460)**:671-9.
30. Carroll JS, Meyer CA, Song J, Li W, Geistlinger TR, Eeckhoute J, Brodsky AS, Keeton EK, Fertuck KC, Hall GF, Wang Q, Bekiranov S, Sementchenko V, Fox EA, Silver PA, Gingeras TR, Liu XS, Brown M: **Genome-wide analysis of estrogen receptor binding sites.** *Nat Genet* 2006, **38(11)**:1289-97.
31. Edgar R, Domrachev M, Lash A: **Gene Expression Omnibus: NCBI gene expression and hybridization array data repository.** *Nucleic Acids Research* 2002, **30**:207-10.
32. The Cancer Genome Atlas Group: **The Cancer Genome Atlas website.** 2011 [http://tcga-data.nci.nih.gov/tcga/].
33. Croesa D, Couche F, Wodak SJ, van Helden J: **Inferring Meaningful Pathways in Weighted Metabolic Networks.** *J Mol Biol* 2006, **356**:222-236.
34. Faust K, Dupont P, Callut J, van Helden J: **Pathway discovery in metabolic networks by subgraph extraction.** *Bioinformatics* 2010, **26(9)**:1211-1218.
35. Dezső Z, Nikolsky Y, Nikolskaya T, Miller J, Cherba D, Webb C, Bugrim A: **Identifying disease-specific genes based on their topological significance in protein networks.** *BMC Syst Biol* 2009, **3**:36.

doi:10.1186/1471-2105-14-S2-S7
Cite this article as: Gaire *et al.*: Discovery and analysis of consistent active sub-networks in cancers. *BMC Bioinformatics* 2013 **14**(Suppl 2):S7.