METHODOLOGY

Detection of statistically significant network changes in complex biological networks

Raghvendra Mall[1], Luigi Cerulo[2,3], Halima Bensmail[1], Antonio Iavarone[4] and Michele Ceccarelli[1*]

*Correspondence: mceccarelli@gmail.com
[1]QCRI - Qatar Computing Research Institute, HBKU, Doha, Qatar
Full list of author information is available at the end of the article

Abstract

Background: Biological networks contribute effectively to unveil the complex structure of molecular interactions and to discover driver genes especially in cancer context. It can happen that due to gene mutations, as for example when cancer progresses, the gene expression network undergoes some amount of localised re-wiring. The ability to detect statistical relevant changes in the interaction patterns induced by the progression of the disease can lead to the discovery of novel relevant signatures. Several procedures have been recently proposed to detect sub-network differences in pairwise labeled weighted networks.

Results: In this paper, we propose an improvement over the state-of-the-art based on the Generalized Hamming Distance adopted for evaluating the topological difference between two networks and estimating its statistical significance. The proposed procedure exploits a more effective model selection criteria to generate p-values for statistical significance and is more efficient in terms of computational time and prediction accuracy than literature methods. Moreover, the structure of the proposed algorithm allows for a faster parallelized implementation. In the case of dense random geometric networks the proposed approach is 10-15x faster and achieves 5-10% higher AUC, Precision/Recall, and Kappa value than the state-of-the-art. We also report the application of the method to dissect the difference between the regulatory networks of IDH-mutant versus IDH-wild-type glioma cancer. In such a case our method is able to identify some recently reported master regulators as well as novel important candidates.

Conclusions: We show that our network differencing procedure can effectively and efficiently detect statistical significant network re-wirings in different conditions. When applied to detect the main differences between the networks of IDH-mutant and IDH-wild-type glioma tumors, it correctly selects sub-networks centered on important key regulators of these two different subtypes. In addition its application highlights the role novel candidates that are not detected by standard single network-based approaches.

Keywords: Differential Networks; Gene Regulatory Network Inference; Master Regulators

Background

The omni-presence of complex networks is reflected in wide variety of domains including social networks [1, 2], web graphs [3], road graphs [4], communication networks [5], financial networks [6] and biological networks [7, 8, 9]. Although we focus on biological networks many aspects of the method proposed in this paper can also be applied for networks in other contexts. In cancer research, the comparison between gene regulatory networks, protein interaction networks, and DNA methy-

lation networks is performed to detect differences between two conditions, such as, healthy and disease [10, 11]. This can lead to discovery biological pathways related to the disease condition, and, in case of cancer, the gene regulatory changes as the disease progresses [12, 13, 14].

A central problem in cell biology is to model functional networks underlying interactions between molecular entities from high throughput data. One of the main question is how the cell globally changes its behavior in response to external stimuli or what is the effect of alterations such as, driver somatic mutations and changes in copy number. Signatures of differentially expressed and/or methylated genes are the downstream effect of global cell de-regulation in different conditions such as cancer subtypes. Therefore, it is argued that driver mutations activate functional pathways described by different global re-wiring of the underlying gene regulatory network.

The identification of significant changes induced by the presence or the progression of disease can help to discover novel molecular diagnostics and prognostic signatures. For example, it is known that, according to the mutation of the gene IDH [15, 16], the majority of malignant brain tumors can be divided two main macro-categories, which can be further divided in seven molecular and clinically distinct subtypes [17]. These two macro-groups are characterized by highly different global expression and epigenomic profiles. Hence, one of the main questions to understand the molecular basis of diseases is how to identify significant changes in the regulatory structure in different conditions.

Various techniques have been developed to compare two graphs including graph matching and graph similarity algorithms [18, 19, 20]. However, the problem addressed in this paper is different from popular graph theory problems including graph isomorphism [21] and sub-graph matching [22]. Here the goal is to identify statistically significant differences between two weighted networks (with or without labels).

One common statistic used to distinguish one graph, A from another B, having the same number of nodes N, is the Mean Absolute Difference (MAD) metric, defined as: $d(A,B) = \frac{1}{N(N-1)} \sum_{i \neq j} |a_{ij} - b_{ij}|$, where a_{ij} and b_{ij} are edge weights corresponding to the topology of networks A and B. This distance measure is equivalent to the Hamming distance [23] and has been extensively used in literature to compare networks [24, 25]. Another statistic used to test association between networks is the Quadratic Assignment Procedure (QAP) defined as: $Q(A,B) = \frac{1}{N(N-1)} \sum_{i=1} \sum_{j=1} a_{ij} b_{ij}$. The QAP metric is used in a permutation-based procedure to differentiate two networks [26, 27]. Ruan *et al.* showed that these metrics are not always sensitive to subtle topological variations [28].

Our aim is to detect statistically significant differences between two networks under the premise that any true topological difference between the two networks would involve only a small set of edges when compared to all the edges in the network. Recently, a Generalized Hamming Distance (GHD) based method was introduced to measure the distance between two labeled graphs [28], where it was shown that the GHD statistic is more robust than MAD and QAP metrics for identifying subtle variations in the topology of paired networks. In particular the authors showed that GHD permutation distribution follows a normal distribution

with closed-form expression for first two moments under the null hypothesis that networks A and B are independent. Utilizing the moments, corresponding p-values were obtained in closed-form. They also propose a differential sub-network identification technique namely dGHD. The advantage of this technique is that – unlike previous differential network analysis techniques [25, 29, 30] – it provides a closed-form solution for p-values for the differential sub-network left after iterative removal of the least differential nodes. We propose an extension of dGHD, namely Closed-Form approach that exploits the conditions for asymptotic normality which is computationally cheaper and attains better prediction performance than the Original (dGHD) algorithm. Computational efficiency and prediction accuracy is crucial in cancer contexts where networks have a large number of nodes and the topological difference is associated to few driver genes.

Methods

Preliminaries on Generalized Hamming Distance

The Generalized Hamming Distance is a way to estimate the distance between two weighted graphs [28]. Let $A = (V, E_A)$ and $B = (V, E_B)$ be two graphs, with the same set of nodes $V = \{1, \ldots, N\}$, and different sets of edges, E_A and E_B. The Generalized Hamming Distance (GHD) is defined as:

$$\mathrm{GHD(A,B)} = \frac{1}{\mathrm{N(N-1)}} \sum_{\mathrm{i,j,i\neq j}} (\mathrm{a'_{ij}} - \mathrm{b'_{ij}})^2, \tag{1}$$

where a'_{ij} and b'_{ij} are mean centered edge-weights defined as:

$$a'_{ij} = a_{ij} - \frac{1}{N(N-1)} \sum_{i,j,i\neq j} a_{ij}, \;\; b'_{ij} = b_{ij} - \frac{1}{N(N-1)} \sum_{i,j,i\neq j} b_{ij}$$

The edge weights, a_{ij} and b_{ij}, depend on the topology of the network and provide a measure of connectivity between every pair of nodes i and j in A and B. Different metrics have been adopted to measure the connectivity between pairs of nodes, including: topological overlap (TO) [31, 32], cosine similarity and Pearson correlation [33]. In our experiments, we used the cosine similarity to capture first order interactions between the nodes in the network. Cosine similarity computation scales well for large sparse networks and can be used in place of TO, as it has nearly perfect correlation with it.

Given two networks A and B, a permutation π of the labels of the vertices of A (keeping the edges unchanged) generates a permuted network A_π. The quantity $GHD_\pi(A_\pi, B)$ represents the test statistics of an inferential problem having as null hypothesis $\mathcal{H}_o$: *Graphs A and B are independent* [28]. The distribution of GHD_π can be obtained through an exhaustive calculation which can be approximated by a Monte Carlo approach. The authors of [28], indeed, simplified this calculation showing that under the null hypothesis it can be approximated well by a normal distribution with moments that can be obtained analytically.

This can be shown as:

$$\frac{\mathrm{GHD(A_\pi, B)} - \mu_\pi}{\sigma_\pi} \sim N(0,1) \tag{2}$$

where μ_π is the asymptotic value of the mean GHD statistic and σ_π is the asymptotic value of the standard deviation of GHD statistic computed between A_π and B. In order to calculate the μ_π and σ_π values we define:

$$S_a^t = \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} a_{ij}^t, t = 1, 2 \text{ and } T_a = \sum_{i=1}^{N} (\sum_{j=1, j\neq i}^{N} a_{ij})^2$$
$$S_b^t = \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} b_{ij}^t, t = 1, 2 \text{ and } T_b = \sum_{i=1}^{N} (\sum_{j=1, j\neq i}^{N} b_{ij})^2$$

Here a_{ij}^t and b_{ij}^t are the edge weights with the power t. Furthermore, we require the following terms:

$$A_a = (S_a^1)^2, \quad B_a = T_a - (S_a^2) \text{ and } \mathrm{C_a} = \mathrm{A_a} + 2(\mathrm{S_a^2}) - 4\mathrm{T_a}$$
$$A_b = (S_b^1)^2, \quad B_b = T_b - (S_b^2) \text{ and } \mathrm{C_b} = \mathrm{A_b} + 2(\mathrm{S_b^2}) - 4\mathrm{T_b}$$

Using these definitions the closed-form expression for mean μ_π and variance σ_π^2 are expressed as:

$$\begin{aligned} \mu_\pi &= \frac{S_a^2 + S_b^2}{N(N-1)} - \frac{2(S_a^1)(S_b^1)}{N^2(N-1)^2}, \\ \sigma_\pi^2 &= \frac{4}{N^3(N-1)^3}[2(S_a^2)(S_b^2) + \frac{4(B_a)(B_b)}{N-2} + \\ &\quad \frac{(C_a)(C_b)}{(N-2)(N-3)} - \frac{(A_a)(A_b)}{N(N-1)}] \end{aligned} \tag{3}$$

Given a significance threshold α (e.g. 0.01), p-values $> \alpha$ indicate that there is no sufficient evidence to reject the null hypothesis ($\mathcal{H}_o$) that graphs A and B are independent. Hence, higher p-values indicate more probability that the two graphs under consideration are independent.

Differential sub-network detection with GHD

The GHD distance is able to tell us to what extent are two graphs different but is not able to identify which parts of the graph are similar and which are different. In this work, we are interested in detecting which part of the graphs contribute to make the two graphs different. We call such different sub-graphs *differential sub-networks*.

The notion of differential sub-networks is based on the idea that when comparing two networks only a subset of edges would present altered interaction. The goal is to identify the set of nodes, namely V^*, associated with such a subset of edges and the p-values p^* corresponding to the nodes in V^*. This goal, formulated as a statistical test, requires that for such a subset V^* there is no sufficient evidence to reject the null hypothesis that the corresponding sub-networks $A^*(V^*, E_{A^*})$ and $B^*(V^*, E_{B^*})$ are statistically independent.

The idea here is to adopt an iterative technique to identify the set of nodes V^* which contributes more to the difference. We start from the dGHD algorithm proposed in [28]. The algorithm measures the edge connectivity with topological overlap metric and benefits from the closed-form solution of p-value (Equations (3)).

In the dGHD algorithm, an iterative procedure is followed where at each iteration the change in centralized GHD (cGHD) i.e. cGHD = GHD$(A, B) - \mu_\pi$ is estimated after the removal of one node. The node where the change in cGHD (i.e. difference in cGHD before and after removal of a node) is maximum is removed. The GHD statistic is computed for remaining sub-networks and the p-value is estimated. This process is repeated till a user specified minimal set size is reached or it is no-longer possible to have closed-form representation for p-values which happens for $N \leq 3$ as shown in equation 3. The p-values are then adjusted for multiple testing by controlling the false discovery rate [34].

The dGHD algorithm suffers from the following limitations: a) During the i^{th} iteration, the GHD measure is calculated $N - i$ times on different sub-graphs with an overall time complexity $\sim O(N^2 \times |E|)$ where $E = E_A \cup E_B$; b) The algorithm is prone to discovery more false positives since it uses the change in cGHD (ΔcGHD) as a model selection criterion. We overcome such limitations by proposing the following improvements:

1 *Remove nodes by exploiting the Closed-Form.* We use the idea that nodes which have similar topology in networks A and B will contribute the least to cGHD. So, we first calculate the closed-form contribution of each node in cGHD once using equation 4 and then iteratively remove nodes with least contributions. However, this process is continued till we observe that the p-value of the remaining sub-network becomes greater than a threshold θ.

2 *Using a different model selection criterion.* Once the p-value reaches θ, we follow a procedure similar to the dGHD algorithm but use the more intuitive criterion of selecting the node that when removed makes the cGHD value maximum rather than using the change in the cGHD value (before and after removal of a node) as a model selection criterion. By using this model selection criterion, we iteratively identify and remove that node whose contribution is least in the cGHD.

The advantage of the Closed-Form approach is that we significantly reduce the computational complexity and improve the predictive performance. A simple alternative to the Closed-Form approach would be to sort all the nodes based on their contribution to cGHD and thus rank all the nodes based on their capability to differentiate the two networks with complexity ($O(N \log N)$). However, then we will not be able to identify statistically different sub-networks between the two graphs as indicated in [28].

Closed-Form Approach

We propose a fast approach to perform differential sub-network analysis taking into consideration the contribution of each node to GHD and μ_π. Using equations (1)

and (3) this can mathematically be represented as:

$$\begin{aligned}\mathrm{GHD}(A,B)(i) = \frac{1}{N(N-1)}\Big(&\sum_{j=1,j\neq i}^{N}(a'_{ij})^2 + \sum_{j=1,j\neq i}^{N}(b'_{ij})^2 \\ &- \sum_{j=1,j\neq i}^{N}(2a'_{ij}\times b'_{ij})\Big) \\ \mu_\pi(i) = \frac{(\sum_{j=1,j\neq i}^{N}(a_{ij})^2 + \sum_{j=1,j\neq i}^{N}(b_{ij})^2)}{N(N-1)} &- \frac{2(\sum_{j=1,j\neq i}^{N}a_{ij})(S_b^1)}{N^2(N-1)^2} \\ -\frac{2(\sum_{j=1,j\neq i}^{N}b_{ij})(S_a^1)}{N^2(N-1)^2} &+ \frac{2(\sum_{j=1,j\neq i}^{N}a_{ij})(\sum_{k=1,k\neq i}^{N}b_{ik})}{N^2(N-1)^2}\end{aligned} \tag{4}$$

We observe that if we sum $\mathrm{GHD}(A,B)(i)$ and $\mu_\pi(i)\ \forall i \in V$, we obtain $\mathrm{GHD}(A,B)$ and μ_π. We use the idea that nodes which have similar topology in networks A and B will contribute the least to centralized GHD, i.e. $\mathrm{GHD}(A,B) - \mu_\pi$. We calculate the Closed-Form contribution of each node in the centralized GHD (cGHD) once using equation (4) and then iteratively remove nodes with least contribution to the cGHD, i.e. nodes having similar topology in graphs A and B. Thus, we calculate cGHD once and sort all the nodes based on their contribution to the cGHD metric.

This process is continued till we observe that the p-value of the remaining sub-network becomes greater than a threshold θ. Once the p-value reaches θ, we estimate $\Delta_{V_K} = \mathrm{GHD}(A(V_K,E_A),B(V_K,E_B)) - \mu_{V_K}$ where μ_{V_K} is the mean of the permutation distribution for the nodes (V_K) of the remaining sub-network. Furthermore, we define $\Delta_{V_{K|i}}$ as the value of cGHD after removal of node i. We adopt a different model selection criterion than that proposed in [28] to remove non-differential nodes. We use the intuitive criterion of selecting that node after removal of which the cGHD value becomes maximum, i.e. the node which was most similar in terms of topology for the paired-graphs. Finally, the obtained p-values are adjusted for multiple testing by controlling the false discovery rate [34]. Provided the paired-graphs A and B, the calculation of $\Delta_{V_{K|i}}$ can be done independently for each i. Details of the Closed-Form method is provided in Algorithm 1. The sensitivity of the Closed-Form approach with the parameter θ is demonstrated in Experimental Results section. Table 1 summarizes the improvements with respect to the dGHD algorithm in terms of time complexity.

Alternative Procedure (Fast Approximation)

We propose an alternative procedure to the Closed-Form approach namely the Fast Approximation method where we first calculate the cGHD value without including the i^{th} node, $\forall i \in V$ once. This helps to estimate the cGHD value after removal of the i^{th} node and can be performed in parallel. Our aim is to quickly discard those nodes after removal of which the cGHD value becomes large thereby removing nodes which were contributing least to the cGHD value. This helps to reduce the dependence between the two sub-networks by removing nodes which have similar topology in graphs A and B. Again, the idea is motivated by the premise that only a subset of nodes will form the differential sub-networks in graph A and B.

Algorithm 1: Closed-Form

Data: Graphs A and B with N vertices V.
Result: Subset V^* representing the set of nodes which comprise the differential sub-network & p-values for GHD measure.
$V^* = \{\}$ // Empty Set for differential sub-network nodes.
$V_K = V$ // Initialize a copy of the set of vertices V.
$p^* = \{\}$ // Empty Set for p-values.
Calculate contribution of each node i in centralized GHD using equation 4.
Sort all nodes based on their contribution in ascending order and keep in $\mathcal{O}$.
while $N > 3$ **do**
  $z = \frac{\mathsf{GHD}(A(V_K,E_A),B(V_K,E_B))-\mu_{V_K}}{\sigma_{V_K}}$.
  Calculate p-value using z and append p-value to p^*.
  if $p\text{-}value > \theta$ **then**
    $\Delta_{V_K} = \{\}$ **forall the** $i \subset V_K$ **do**
      $t = (\mathsf{GHD}(A(V_{K|i}, E_A), B(V_{K|i}, E_A)) - \mu_{V_{K|i}})$.
      Add t to Δ_{V_K} // Perform in parallel.
    $n^* = \mathsf{max}_i \Delta_{V_K}$
    // Select that node after removal of which cGHD becomes maxmimum.
    Remove node n^* from V_K i.e $V_K = V_K \setminus n^*$ and $\mathcal{O} = \mathcal{O} \setminus n^*$
  else if $p\text{-}value < \theta$ **then**
    $n^* = \mathbf{min}_i(\mathcal{O})$ // Select node in the sub-network with least contribution.
    Remove node n^* from $\mathcal{O}$.
    // $\mathcal{O}$ is sorted so remove 1^{st} node.
  if $p\text{-}value > 0.01$ **then**
    Append n^* to V^*.
  $N = N - 1$.
Adjust the p-values for false-discovery rate [34].

In this approach, we iteratively discard those nodes after removal of which the cGHD value becomes maximal till the p-value for the remaining sub-network reaches a threshold θ. Once the p-value reaches θ, we return back to the procedure of estimating $\Delta_{V_{K|i}}$ $\forall i \in V_K$ as described in the Closed-Form approach. We use the same model selection criterion of selecting that node after removal of which the cGHD value becomes maximum as used in the Closed-Form approach. We then adjust the obtained p-values for multiple testing by controlling the false discovery rate [34]. We refer to this technique as a Fast Approximation to the Original technique (dGHD [28]). We explain the Fast Approximation technique in detail in Algorithm 2.

From our experiments, we observe that the results of the Closed-Form approach and the Fast Approximation technique are identical. Although, in the case of Closed-Form approach, we calculate closed-form contribution of each node in the cGHD value and remove the node with least contribution, while in case of Fast Approximation we select that node after removal of which cGHD value becomes maximum, the ordered list $\mathcal{O}$ obtained for both the methods is identical. Moreover, the computational complexity of the Fast-Approximation technique is the same as that of Closed-Form approach.

Inference of the Glioma networks and Master Regulator Analysis

We used the TCGA pan-glioma samples dataset including 1250 samples (463 IDH-mutant and 653 IDH-wild-type), 583 of which profiled with Agilent microarray and 667 with RNA-Seq Illumina HiSeq (REF) downloaded from the TCGA portal. The batch effects between the two platform were corrected using the COMBAT algorithm [35]. The final gene expression data matrix includes 12,985 genes and 1250 samples. We re-constructed two gene regulatory networks belonging to two

Algorithm 2: Fast Approximation

Data: Graphs A and B with N vertices V.
Result: Subset V^* representing the set of nodes which comprise the differential sub-network & p-values for GHD measure.

$V^* = \{\}$ // Empty Set for differential sub-network nodes.
$V_K = V$ // Initialize a copy of the set of vertices V.
$p^* = \{\}$ // Empty Set for p-values.
$\Delta_{V_K} = \{\}$ **forall the** $i \subset V_K$ **do**
  $t =$**GHD(**$A(V_{K|i}, E_A), B(V_{K|i}, E_A)) - \mu_{V_{K|i}}$.
  // Estimate cGHD value after removal of node i.
  Add t **to** Δ_{V_K}. // Perform in parallel.
Sort Δ_{V_K} **in descending order and keep in** $\mathcal{O}$.
while $N > 3$ **do**
  $z = \frac{\mathsf{GHD}(A(V_K, E_A), B(V_K, E_B)) - \mu_{V_K}}{\sigma_{V_K}}$.
  Calculate p-value using z and append p-value to p^*.
  if $p\text{-}value > \theta$ **then**
    $\Delta_{V_K} = \{\}$ **forall the** $i \subset V_K$ **do**
      $t = (\mathsf{GHD}(A(V_{K|i}, E_A), B(V_{K|i}, E_A)) - \mu_{V_{K|i}})$.
      Add t to Δ_{V_K} // Perform in parallel.
    $n^* = \max_i \Delta_{V_K}$
    // Select that node after removal of which cGHD becomes maximum.
    Remove node n^* from V_K and $\mathcal{O}$
  else if $p\text{-}value < \theta$ **then**
    $n^* = \mathbf{max}_i(\mathcal{O})$ // Select node in the sub-network with least contribution.
    Remove node n^* from $\mathcal{O}$.
  if $p\text{-}value > 0.01$ **then**
    Append n^* to V^*.
  $N = N - 1$
Adjust the p-values for false-discovery rate.

different glioma subtypes: IDH-mutant and IDH-wild-type. Both networks were reconstructed with a four step procedure that follows ARACNE [36]: i) Computation of mutual information between gene expression profiles to determine interaction between Transcription Factors (TFs) and target genes [37]; ii) data processing inequality to filter out indirect relationships [36], iii) permutation test with $1,000$ re-samplings to keep only statistically significant relationships. We also assembled a global glioma network using all the available 1250 transcriptional profiles using the aforementioned method. In this last case we also used intersection with transcription factor (TF) binding sites to keep only relationships due to promoter binding.

Master Regulator Analysis (MRA) algorithm [38] was applied to the global glioma network in order compute the statistical significance of the overlap between the regulon of each TF (i.e. its ARACNe inferred targets) and the differentially expressed gene list (Wilcoxon-Mann-Whitney test FDR ≤ 0.05) between IDH-mutant and IDH-wild-type samples. Given a gene interaction network, generated by ARACNe and a gene phenotype signature (ex. a set of differentially expressed genes), the MRA algorithm computes for each TF the enrichment of the phenotype signature in the regulon of that TF. The regulon of a TF is defined as its neighborhood in the gene interaction network. There are two different methods to evaluate the enrichment of the signature in the regulon. One method uses the statistical Fisher's exact test, while the other approach uses Gene Set Enrichment Analysis (GSEA). In this work, the enrichment was evaluated using the Fisher's exact test and corrected using the Benjamini and Hochberg (BH) false discovery rate (FDR) for multiple-testing. A Master Regulator (MR) gene is a TF which regulon exhibit a statistical significant enrichment of the given phenotype signature.

Moreover, *master regulators activity* allows to computationally infer protein post-transcriptional activity of the TFs selected by the MRA, on an individual sample basis, from the gene expression profiles. It uses the expression of genes that are most directly regulated by a given protein, such as the targets of a transcription factor (TF), as an accurate reporter of its activity, taking into account the regulator mode of action of target genes. The activity of an MR is defined as an index that quantify the activation of the transcriptional program of a given MR in each sample S_i and is calculated as:

$$Act(S_i, MR) = \frac{1}{N}\sum_{k=1}^{N} x_{ki}^{+} - \frac{1}{M}\sum_{j=1}^{M} x_{ji}^{-} \tag{5}$$

where x_{ki}^{+} is the standardized expression level of the k-th positive MR target, *i.e.* directly correlated with the expression of the MR, in the i-th sample, x_{ji}^{-} is the standardized expression level of the j-th negative MR target, , *i.e.* inversely correlated with the expression of the MR, in the i-th sample, and $N(M)$ is the number of positive (negative) targets present in the regulon of the considered MR. If $Act(S_i, MR) > 0$, the MR is activated positively in that particular sample. If $Act(S_i, MR) < 0$, the MR is activated negatively in that particular sample. However, if $Act(S_i, MR) \approx 0$ then it is deactivated in that particular sample. Activated positively means that a high expression is needed to maintain the model of action of its regulator, while activated negatively means that a lower expression is needed to maintain the model of action of its regulator.

Validation in the Rembrandt dataset using Transcription Factor Motif Enrichment Analysis

We used an independent dataset and Transcription Factor Motif Enrichment Analysis (TFMEA) to validate the differential TFs reported in the Glioma case study. Raw gene expression (Affymetrix U133 Plus 2.0) from the publically available Repository for Molecular Brain Neoplasia Data (Rembrandt) (https://caintegrator.nci.nih.gov/rembrandt/) included 444 samples divided in 218 Glioblastoma, 148 Astrocytoma, 67 Oligodendrogliomas and 11 mixed histologies. Expression subtype and IDH status was inferred from gene expression following the procedure in [39]. Supervised differential analysis on the IDH status resulted in 1774 differentially expressed probesets (log fold change ≥ 1 and corrected p-value less than 0.05). 1000 bp 5' promoter sequences of 534 genes corresponding to the differential probesets were downloaded using Ensembl biomart and used to perform TFMEA using two tools: MEME-AME [40] and PSCAN [41] with default parameters. The results in Table 3 report for each TF the minimum enrichment p-value between these two methods.

Results and Discussion

For all our experiments, we used the Closed-Form approach (since results obtained from Closed-Form and Fast-Approximation techniques are identical) and compare it with the original dGHD method [28].

Cosine similarity and topological overlap

The one-step topological overlap measure used to estimate the edge weights is defined as:

$$a_{ij} = \frac{\sum_{l \neq i,j} A_{il}A_{lj} + A_{ij}}{\min(\sum_{l \neq i} A_{il} - A_{ij}, \sum_{l \neq j} A_{lj} - A_{ij}) + 1} \tag{6}$$

In this work we use the cosine similarity to calculate the edge weights a_{ij}. The cosine similarity takes into consideration one-step neighbourhood of nodes i and j while constructing the edge weight and is very efficient to calculate for sparse matrices. The weights a_{ij} are estimated as follows:

$$a_{ij} = \frac{\sum_l A_{il}A_{jl}}{\sqrt{\sum_l A_{il}^2}\sqrt{\sum_l A_{jl}^2}} \tag{7}$$

where A_{ij} represents the adjacency matrix.

We perform an experiment to calculate the correlation between the one-step topological measure and the cosine similarity measure. For this experiment, we generated 250 random geometric networks using $N = 250$ and the connectivity parameter $d = 0.15$.

Figure 1 shows that the cosine similarity metric is nearly perfectly correlated (pearson correlaton = 0.952) to the topological overlap measure.

Sensitivity to θ

In this experiment, we check the sensitivity of the proposed Closed-Form approach w.r.t. the heuristic θ. For this experiment, we first generated 100 random geometric (RG) networks. In a RG network nodes are generated by uniformly sampling N points on $[0, 1]^2$. An edge is then drawn between these points if the Euclidean distance between the points is less than a parameter d. This parameter d controls the density of the RG network where smaller values of d result in sparse networks while larger values of d generates dense networks. In our case, we conducted experiments using two different settings. In the first case, we use $d = 0.15$, while in the second setting, we use $d = 0.3$. For both experiments we fix $N = 250$. For each value of d and for each generated RG network A, we permute the first 50 rows and columns of the network to generate network B. Therefore, the first 50 nodes in networks A and B form the gold-standard.

In order to test the sensitivity of the proposed approach w.r.t. θ, we estimate the fraction of permuted nodes correctly identified by the Closed-Form method for various values of θ. We used a grid of θ values varying from $\Theta = \{10^{-50}, \ldots, 10^{-300}\}$ in multiplicative steps of 10^{-20}. **The goal of this experiment is to show that the fraction of correctly identified nodes w.r.t. various $\theta \in \Theta$ remains nearly constant for smaller values of θ.** Figure 2 shows the result for RG networks with density parameter $d = 0.15$ and $d = 0.3$. From Figure 2, we observe that the median fraction of permuted nodes identified by the proposed approaches increases slowly before it converges to a nearly constant value as we decrease the threshold θ (i.e. increase absolute log of threshold θ).

From this experiment, we conclude that the fraction of truly differential nodes identified by the proposed methods increases as we decrease the threshold θ before it starts to converge for smaller values of threshold θ.

We performed further experiments using different θ for various values of N and observed that threshold θ behaves similarly independent of the value of N. We used the $\theta = 10^{-50}$ as heuristic cut-off for future experiments.

Predictive performance comparison

Experimental Setup: The next simulation study that we carried out was to compare the predictive performance of the proposed approach w.r.t. the dGHD [28] technique. For this experiment, we generate 100 RG networks with $N = 1,000$. For the first experiment we fix the density parameter $d = 0.15$ and permute first 100 nodes in network A to obtain network B. Thus, these first 100 nodes form the differential sub-network for the paired networks A and B.

In the second case, we use the density parameter $d = 0.3$ to generate the edges for network A. We then generate a small RG network with 100 nodes using density parameter $d' = 0.5$. This small dense sub-network is then used to replace the network formed by first 100 nodes in the original network A to form network B. Thus, in the second experiment, these 100 nodes form the differential sub-network for the paired networks A and B. This kind of mechanism can appear in real-life networks, for example, in case of cancer the transcription activity of some set of genes might get enhanced or suppressed in patients resulting in more or fewer edges in a sub-network of the gene or DNA methylation network. Hence, the networks generated in the first case are much sparser in comparison to the networks generated in the second case.

Evaluation Metrics: We define the following terms to be used in our analysis:

- True Positives (TP) - Refers to the nodes that are correctly identified as part of a differential network.
- False Positives (FP) - Refers to the nodes that are incorrectly identified as part of a differential network.
- False Negatives (FN) - Refers to the nodes that are part of the differential sub-network but are not identified correctly as part of the sub-network.
- True Negatives (TN) - Refers to the nodes that are correctly identified as nodes which are not part of the differential sub-network A^* and B^*.

ROC and PR curve comparisons: We generate two set of plots including the receiver operating characteristic (ROC) curves and the precision-recall (PR) curves. To generate the plots as shown in Figure 3, we use the 'ROCR' [42] package in R. It generates relatively smooth curves by automatically using different thresholds to estimate the true positive rate i.e. $\frac{n(TP)}{n(TP)+n(FN)}$ and the false positive rate i.e. $\frac{n(FP)}{n(FP)+n(TN)}$ for ROC plot and precision i.e. $\frac{n(TP)}{n(TP)+n(FP)}$ and recall i.e. $\frac{n(TP)}{n(TP)+n(FN)}$ for the PR plot. Here we use the true positive rate (TPR) and Recall interchangeably. Here $n(\cdot)$ represents the total number of nodes. For generating the plots we used the adjusted p-value lists as obtained from the Closed-Form and dGHD approaches without specifying any threshold to generate smooth curves.

The data in Figure 3A and Figure 3C shows that Closed-Form approach achieves better performance in case of differential sub-networks formed by permuted nodes

and sub-networks with higher density. One of the reasons for relatively poor performance of the dGHD approach is that it has low true positive rate (TPR) and a high false positive rate (FPR) when the network has more edges. This is also reflected by the relatively low Recall and Precision values for the dGHD algorithm in Table 2 when $d = 0.3$ and $d' = 0.5$. From Figure 3C, we can observe that the performance of both the dGHD and Closed-Form algorithm improves w.r.t. ROC when the differential sub-network is denser than the remaining network. However, the gap between the PR curves of Closed-Form and dGHD methods increases when the differential sub-network is denser.

AUC comparision: For all further simulated experiments, we use p-value 0.01 as cut-off in order to determine TP, TN, FP and FN respectively. We also evaluated the area under the ROC curve (AUC_ROC [43]) and area under PR curve (AUC_PR [43]) for 100 runs of Closed-Form and dGHD methods (using p-value 0.01 as cut-off) as shown in Figure 4.

We observe from Figures 4A and 4B that the dGHD method has lower variance w.r.t. AUC_ROC and AUC_PR metrics in comparison to Closed-Form approach in the case of permuted differential sub-network. However, in case of denser differential sub-network, the Closed-Form approach has much smaller variance in comparison to dGHD algorithm w.r.t. AUC_ROC and AUC_PR metrics as depicted in Figure 4C and Figure 4D respectively. This suggests that the performance of Closed-Form technique is better than dGHD method when differential sub-networks are formed either using permuted nodes or higher density. In order to test for significance we performed the Student's t-test under the null that the difference in the mean values of the two ROC distributions is zero i.e. $\mu_{AUC\_ROC_A} - \mu_{AUC\_ROC_B} = 0$. At a significance level of 5%, we obtain p-value of 0.48 in case of permuted sub-network, thereby accepting the null i.e. the difference between the two distributions is not significant. In the case of paired networks with a denser differential sub-network (i.e. $d' = 0.5$), we obtain p-value of 3.42×10^{-14} for the Student's t-test, thereby rejecting the null. Similarly for the two PR distributions we obtained p-value of 0.42 in case of permuted sub-network and p-value of 2.64×10^{-20} for the denser differential sub-network.

Comparison with Community Detection techniques

The task of identifying differential sub-networks can also be rephrased as one of finding heavy sub-networks on a single network (say C) constructed by considering the absolute difference in the edge weights between the topological graph of network A and the topological graph of network B i.e. $C_{ij} = \|a_{ij} - b_{ij}\|$, $\forall i, j \in V$. This problem can then be construed as one of identifying dense modules in the network C i.e. from the previous experiments we want to discover a module corresponding to the set of nodes which have permuted or identify the denser sub-network forming the differential sub-network as a module.

The task of identifying dense/heavy modules in a network (C) is often referred as community detection or graph partitioning or graph clustering. There is a plethora of research associated with the problem of community detection including [44, 45, 46, 47, 48, 49, 50, 51]. Several of these methods such as jActiveModules [52] and Spinglass algorithm [47] have also been applied to identify biologically meaningful

modules (like functional modules, protein complexes, disease associated genes etc.) in biological networks as shown in [53, 54]. For our task of identifying dense modules in network C we applied 3 different community detection methods namely Louvain [45], Infomap [46] and Spinglass [47] techniques to have a comprehensive comparison with the proposed Closed-Form approach. We used the implementation of these methods available in the 'igraph' package in R and run each of these methods at their default settings.

We used the same set of RG networks as in the previous experiments to have a comparison with the community detection techniques. Since we are considering the difference in the topology of networks A and B in network C, we remove all the similarity between the two networks and the module with the maximum internal volume (i.e. total weight of edges within the community) is the one capturing the maximum difference between the topologies of networks A and B. Hence, we consider the densest inferred module as the one comprising the differential sub-network and label all the nodes belonging to this cluster as differential while all the other modules are considered non-differential. Using this notion to label the inferred communities, we compare the results obtained for the 3 different community detection techniques w.r.t. the gold standard (i.e. the actual set of labelled nodes which either belong to the permuted sub-network or belong to the denser sub-network) in a binary classification framework [55, 56]. These results are integrated in Table 2 along with the results of dGHD technique and the proposed Closed-Form (CF) approach. We assess the results obtained from the 3 community detection methods w.r.t. several quality metrics commonly used for binary classification including Precision, Recall, Kappa, Accuracy, Specificity, AUC_ROC and computational time. From Table 2, we observe that the Louvain method clearly outperforms the Infomap and Spinglass techniques in correctly identifying the differential sub-network as a module with respect to the various evaluation metrics.

Simulated Result Analysis

Finally, the summary Table 2 highlights the computational efficiency and better predictive capabilities of the proposed technique in comparison to dGHD algorithm. For this comparison, we report the results obtained on 100 random runs of RG networks with $N = 1000$, $d = 0.15$ and $d = 0.3$ respectively, where the first 100 nodes are permuted. We also report results when the first 100 nodes form the denser differential sub-networks i.e. in experiments where $d = 0.15$ use $d' = 0.3$ to form denser sub-network and where $d = 0.3$ use $d' = 0.5$ to form denser sub-network. We also conducted experiments on undirected Power Law (PL) graphs using $N = 1000$ and $E = 10,000$ with power law exponents $\alpha = \{2, 3\}$ respectively. We permuted the first 100 nodes of each PL network (B) to form the permuted network (A). We performed 100 random runs and report the mean values for various evaluation metrics.

Table 2 compares the Closed-Form, Louvain, Infomap, Spin-glass and dGHD techniques w.r.t. various standard evaluation metrics like AUC, Precision, Recall, Accuracy, Specificity, Kappa statistic and computational Time for all the simulation experiments. Higher values of these evaluation metrics represents better quality results. Here the time required by dGHD algorithm is normalized to 1 and the time required by the other algorithms is scaled by the same normalization factor.

We observe from Table 2 that the Closed-Form approach performs exceedingly well in case of experiments on denser RG networks ($d = 0.3$) and PL graphs. It emerges as the best method on these networks for various evaluation metrics. For this configuration, in case of both permuted and denser differential sub-networks, the mean AUC_ROC of Closed-Form approach is at least 10% higher than the dGHD algorithm. This is also reflected in higher values of Precision (0.714 and 0.771) and Recall (0.789 and 0.930) metrics for Closed-Form approach in comparison to low values of Precision (0.645 and 0.7) and Recall (0.577 and 0.731) for the dGHD algorithm in case of these experiments.

However, in case of sparse networks where its relatively easier to identify differential sub-networks ([28]), both Closed-Form and dGHD method have similar predictive performance. For sparse networks, the Louvain method nearly outperforms all other methods for the task of identifying the differential sub-network as a module. From Table 2, we observe that the 3 community detection techniques have nearly perfect Recall scores but usually have relatively low Precision values. This indicates that these methods correctly identify all the nodes forming the differential sub-network but also detect a large quantity of false-positives in the densest module, thereby reducing the Precision values. The Louvain and Infomap methods are extremely fast and interestingly the Louvain method has highest Precision (0.887) which is at least 10% higher than dGHD algorithm and 5% higher than Closed-Form approach while identifying the dense differential sub-network in a sparse network ($d = 0.15$, $d' = 0.3$) as shown in Table 2.

We observe that among the community detection techniques the Louvain method is the most efficient and is highly competitive with the dGHD algorithm but cannot outperform the Closed-Form approach on denser networks and Power Law graphs.

Case study in Glioma

As a case study, we performed the differential sub-networks analysis of two gene regulatory networks re-constructed from the glioma dataset available on the TCGA. It is well known that the majority of gliomas are divided into two main macro-categories according to the mutation of the gene IDH1 [17, 15, 57]. Therefore, an important biological question, that motivated the development of the reported methodology, was to identify the sub-networks of differentially activated transcription factors (TFs) in these two major conditions. We re-constructed two gene regulatory networks belonging to two different glioma subtypes: IDH-mutant and IDH-wild-type as reported in the Methods Section.

In our final networks we have 457 TFs and $4,085$ targets. We observe that these networks consist of $13,683$ unique connections for IDH-mutant and $14,158$ for IDH-wild-type between TF-TF and TF-target. Using these networks, we construct two unipartite topological graphs as described in the Methods section for the 457 TFs. We then perform the proposed differential sub-network analysis to identify the TFs which are part of differential sub-networks in these topological graphs.

Figure 5 shows the significant differential sub-networks and Table 3 reports the topmost TFs which are part of differential sub-networks as detected by our algorithm. In the table, GHD and μ_π represent the generalized Hamming Distance and its asymptotic mean beetween the subgraphs after removing the specific transcription factor in each row of the table.

To asses the biological validity, we also assembled a global glioma network using all the available transcriptional profiles using the same method described above and performed a master regulator analysis [38] with respect to the molecular phenotype under investigation, *i.e.* genes differentially expressed between IHD mutant and wild type. Master regulator analysis is extensively adopted to identify TFs that act as principal regulators in driving the phenotype from one condition to another. The last three columns of the table show the master regulator analysis results for each TF (in boldface the most significant master regulators).

Interestingly, among the topmost TFs (out of 457) forming the differential subnetworks, we found several genes known to have a central role in controlling specific glioma subtypes as well as novel candidates that deserve further biological validation. In particular, differential network analysis reveals that the sub-network of STAT3 is one the most different between IDH-mutant and IDH-wild-type networks and a particularly significant Master Regulator of this wild-type phenotype. Members of our group have previously shown that STAT3, together with C/EBPβ, is a key regulator of the mesenchymal differentiation and predicts the poor clinical outcome of IDH-wild-type gliomas [38]. Another key regulator of the IDH-wild-type gliomas was recently reported by using an integrative functional copy number analysis is the set of HOXA genes [17]. Moreover, another key network hub that the algorithm detects as different is SOX10 which appears to be an active master regulator of the IDH-mutant phenotype. We recently reported that the GCIMP-low subgroup in the IDH-mutant cohort can mediated by loss of CpG methylation and binding of SOX factors [17]. Furthermore, our algorithm identifies methyl-CpG-binding domain protein 2 (MBD2) as a differential network hub. In particular, MBD2 has no links in the IDH-wild-type network, which is consistent with its activity nearby zero in IDH-wild-type, whereas it is highly connected in the IDH-mutant network where it is characterized by the CpG island methylator phenotype (GCIMP) [58]. The activity of MBD2 in IDH-mutant network is revealed as negative suggesting that to maintain the gene expression phenotype in tumor samples its expression is lower. Further investigation is needed to claim such an hypothesis as MBD2 is known also as a mediator of the epigenetic gene regulation and its role in glioblastoma is being studied as its over-expression may drive tumor growth by suppressing the anti-angiogenic activity of key tumor suppressors [59].

In addition to these TFs which are also identified by the standard Master Regulator Analysis, the differential network method highlights several other TFs as hubs of differential subnetworks. For example ETV1 and ETV4 which are overexpressed in gliomas of the Codel subtype carrying the mutation of the CIC gene [60]. Another differential subnetwork hub not detected by standard MRA is the tumor suppressor RFX1 whch has been identified as an important target/regulator of the malignancy of glioblastoma. [61] where as the cell cycle regulators such as E2F1 and E2F1 which play a role in progression of IDH mutant glioma are also detected by the Closed-Form algorithm [62].

Finally, as a further independent validation we used another dataset and a different sequence-based methodology to validate the TF in Table 3. In particular, we evaluated whether the TFs identified by the differential network analysis were significant in another dataset using motif enrichment analysis. For this purpose

we used 444 expression profiles of gliomas from the Rembrandt dataset to identify differential expressed genes (DEGs) between IDH mutant and IDH wild type samples. Motif enrichment was then performed on the promoters (1000 bp) of the DEGs using AME [40] and PSCAN [41] as reported in the Methods section. Interestingly, twenty-one of the 45 selected had a significant overrepresentation of the corresponding binding site on the promoter of the DEGs in the Rembrandt dataset.

Conclusion

The comparison of gene expression profiles across different phenotypes is enabling the discovery of novel biomarkers for prognosis or diagnosis. They hold the key to identify novel targets for therapeutical intervention. In this paper, we proposed an improvement to the state-of-the-art for comparing two labeled/unlabeled graphs that are representative of two conditions (e.g. the macro-categories according to the mutation of the gene IDH1 in our case study) and identifying statistically significant differences in their topology. We used the centralized GHD (cGHD) metric [28] to calculate the distance between the two labeled networks. We proposed a Closed-Form approach, an improvement to the dGHD algorithm, to detect localized topological differences between paired networks. The Closed-Form approach calculates the closed-form contribution of each node in the cGHD metric and efficiently removes nodes with the smaller contributions in the cGHD value. From our experiments on scale free random geometric networks, we discovered that the Closed-Form approach was 10-15x faster than Original method from a computational complexity point of view. For differential sub-network analysis in very sparse paired graphs, both the Closed-Form and Original methods had good predictive performance. They reached mean AUC values of ≈ 0.935 and ≈ 0.926 respectively for 100 random runs of simulation experiments. However, for relatively denser networks, the Closed-Form approach outperformed the Original method. The proposed method achieved a mean AUC of ≈ 0.877 while the Original technique reached a mean AUC of ≈ 0.724. The Closed-Form approach also achieved much higher Precision, Recall and Kappa values in comparison to the Original method for relatively denser networks. We applied our algorithm to detect the main differences between the networks of IDH-mutant and IDH-wild-type glioma tumors and show that it correctly selects sub-networks centered on important key regulators of these two different subtypes.

Competing interests
The authors declare that they have no competing interests.

Author's contributions
RM conceived the methodology, developed the algorithms and drafted the manuscript. LC generated the data on glioma and helped to draft the manuscript. HB performed the statistical analysis. AI participated in the design of the study and to the critical analysis of the results. MC conceived of the study, participated in its design and co-ordination and helped to draft the manuscript.

Availability of data and materials
The scripts implementing the proposed algorithms are available in R at https://sites.google.com/site/raghvendramallmlresearcher/codes.

Acknowledgements
This work was funded by Qatar Foundation.

Author details
[1]QCRI - Qatar Computing Research Institute, HBKU, Doha, Qatar. [2]Department of Science and Technology, University of Sannio, Benevento, Italy. [3]BioGeM, Institute of Genetic Research "Gaetano Salvatore", Ariano Irpino (AV), Italy. [4]Department of Neurology, Department of Pathology, Institute for Cancer Genetics, Columbia University Medical Center, New York, USA.

References

1. Jin L, Chen Y, Wang T, Hui P, Vasilakos AV. Understanding user behavior in online social networks: a survey. Communications Magazine, IEEE. 2013 September;51(9):144–150.
2. Mislove A, Marcon M, Gummadi KP, Druschel P, Bhattacharjee B. Measurement and Analysis of Online Social Networks. In: Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement. IMC '07. ACM; 2007. p. 29–42.
3. Broder A, Kumar R, Maghoul F, Raghavan P, Rajagopalan S, Stata R, et al. Graph Structure in the Web. Comput Netw. 2000;33(1-6):309–320.
4. Erath A, Löchl M, Axhausen K. Graph-Theoretical Analysis of the Swiss Road and Railway Networks Over Time. Networks and Spatial Economics. 2009;9(3):379–400.
5. Kesidis G. An Introduction to Communication Network Analysis. Hoboken, NJ: Wiley; 2007.
6. Boginski V, Butenko S, Pardolas PM. Statistical analysis of financial networks. Computational Statistics and Data Analysis. 2005;48(2):431–443.
7. Ideker T, Ozier O, Schwikowski B, Siegel AF. Discovery regulartory and signalling circuits in molecular interaction networks. Bioinformatics. 2002;18.
8. Keller A, Bakes C, Gerasch A, Kaufmann M, Kohlbacher O, Meese E, et al. A novel algorithm for detecting differentially regulated paths based on gene enrichment analysis. Bioinfomatics. 2009;25(21):2787–2794.
9. Nacu S, Critchley-Throne R, Lee R, Holmes S. Gene expression network analysis and applications to immunology. Bioinformatics. 2007;23(7):850–858.
10. Dehmer M, Emmert-Streib F. Analysis of Microarray Data: a network-based appraoch. Weinheim: John Wiley & Sons; 2008.
11. D'haeseleer P, Liang S, Somogyi R. Genetic network inference: From co-expression clustering to reverse engineering. Bioinformatics. 2000;16(8):707–726.
12. Wallace TA, Martin DN, Ambs S. Interaction among genes, tumor biology and the environment in cancer health disparities: examining the evidence on a national and global scale. Carcinogenesis. 2011;32(8):1107–1121.
13. Ahern TP, Horvath-Puho E, Spindler KLG, Sorensen HT, Ording AG, Erichsen R. Colorectal cancer, comorbidity, and risk of venous thromboembolism: assessment of biological interactions in a Danish nationwide cohort. British Journal of Cancer. 2016;114(1):96–102.
14. Ceccarelli M, Cerulo L, Santore A. De novo reconstruction of gene regulatory networks from time series data, an approach based on formal methods. Methods. 2014 Oct;69(3):298–305.
15. Turcan S, Rohle D, Goenka A, Walsh LA, Fang F, Yilmaz E, et al. IDH1 mutation is sufficient to establish the glioma hypermethylator phenotype. Nature. 2012;483(7390):479–483.
16. Network CGAR, et al. Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. N Engl J Med. 2015;2015(372):2481–2498.
17. Ceccarelli M, Barthel FP, Malta TM, Sabedot TS, Salama SR, Murray BA, et al. Molecular Profiling Reveals Biologically Discrete Subsets and Pathways of Progression in Diffuse Glioma. Cell. 2016 Feb;164(3):550–563.
18. Brandes U, Eriebach T. Network Analysis: Methodological Foundations. Springer. 2005;3418.
19. Lena PD, Wu G, Martelli P, Casadio R, Nardini MC. An efficient tool for molecular interaction maps overlap. BMC Bioinforma. 2013;14(1):159.
20. Yang Q, Sze S. Path matching and graph matching in biological networks. Journal of Computational Biology. 2007;14(1):56–67.
21. Ramana MV, Scheinerman ER, Ullman D. Fractional isomorphism of graphs. Discrete Mathematics. 1994;132(1):247–265.
22. Shervashidze N, Schweitzer P, van Leeuwen EJ, Mehlhorn K, Borgwardt KM. Weisfeiler-Lehman Graph Kernels. Journal of Machine Learning Research. 2011;12:2539–2561.
23. Hamming RW. The unreasonable effectiveness of mathematics. American Mathematical Monthly. 1980;87(2):81–90.
24. Butts C, Carley KM. Canonical labeling to facilitate graph comparison; 1998.
25. Gill R, Datta S, Datta S. A statistical framework for differential network analysis from microarrya data. BMC: Bioinformatics. 2010;11(1):95.
26. Mantel N. The detection of disease clustering and a generalized regression approach. Cancer Research. 1967;27(2):209.
27. Hubert LJ. Assignment methods in combinatorial data analysis. Marcel Dekker. 1987;1.
28. Ruan D, Young A, Montana G. Differential analysis of biological networks. BMC Bioinformatics. 2015;16:327.
29. Fuller TF, Ghazalpour A, Aten JE, Drake TA, Lusis AJ, Horvath S. Weighted Gene Co-expression Network Analysis Strategies Applied to Mouse Weight. Mammilian Genome. 2007;18(6):463–472.
30. Ha MJ, Baladandayuthapani V, Do KA. DINGO: differential network analysis in genomics. Bioinformatics. 2015;31(21):3413–20.
31. Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol. 2005;4(1):1128.
32. Allen JD, Xie Y, Chen M, Girad L, Xao GH. Comparing statistical methods for constructing large scale gene networks. PLoS ONE. 2012;7(1:e29348).
33. Deshpande R, Vandersluis B, Myers CL. Comparison of profile similarity measures for genetic interaction networks. PLoS ONE. 2013;8(7:e68664).
34. Benjamini Y, Yekutieli D. The control of false discovery rate in multiple testing under dependency. Annals of Statistics. 2001;29:1165–1188.

35. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007;8(1):118–127.
36. Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Favera RD, et al. ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. BMC Bioinformatics. 2006;7(S-1).
37. Sales G, Romualdi C. parmigene - a parallel R package for mutual information estimation and gene network reconstruction. Bioinformatics [ISMB/ECCB]. 2011;27(13):1876–1877. Available from: `http://dblp.uni-trier.de/db/journals/bioinformatics/bioinformatics27.html#SalesR11`.
38. Carro MS, Lim WK, Alvarez MJ, Bollo RJ, Zhao X, Snyder EY, et al. The transcriptional network for mesenchymal transformation of brain tumours. Nature. 2010;463(7279):318–325.
39. Guan X, Vengoechea J, Zheng S, Sloan AE, Chen Y, Brat DJ, et al. Molecular subtypes of glioblastoma are relevant to lower grade glioma. PLoS One. 2014;9(3):e91216.
40. McLeay RC, Bailey TL. Motif Enrichment Analysis: a unified framework and an evaluation on ChIP data. BMC bioinformatics. 2010;11(1):1.
41. Zambelli F, Pesole G, Pavesi G. Pscan: finding over-represented transcription factor binding site motifs in sequences from co-regulated or co-expressed genes. Nucleic acids research. 2009;37(suppl 2):W247–W252.
42. Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCR: visualizing classifier performance in R. Bioinformatics. 2005;21(20):3940–3941.
43. Mankiewicz R. The Story of Mathematics. Princeton, NJ: Princeton University Press; 2004.
44. Girvan M, Newman ME. Community structure in social and biological networks. Proceedings of the national academy of sciences. 2002;99(12):7821–7826.
45. Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment. 2008;2008(10):P10008.
46. Rosvall M, Bergstrom CT. Multilevel compression of random walks on networks reveals hierarchical organization in large integrated systems. PloS one. 2011;6(4):e18209.
47. Reichardt J, Bornholdt S. Statistical mechanics of community detection. Physical Review E. 2006;74(1):016110.
48. Orman GK, Labatut V. A comparison of community detection algorithms on artificial networks. In: International Conference on Discovery Science. Springer; 2009. p. 242–256.
49. Mall R, Langone R, Suykens JA. Multilevel hierarchical kernel spectral clustering for real-life large scale complex networks. PloS one. 2014;9(6):e99966.
50. Mall R, Langone R, Suykens JA. Kernel spectral clustering for big data networks. Entropy. 2013;15(5):1567–1586.
51. Mall R, Langone R, Suykens JA. Self-tuned kernel spectral clustering for large scale networks. In: Big Data, 2013 IEEE International Conference on. IEEE; 2013. p. 385–393.
52. 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(13):i223–i231.
53. West J, Beck S, Wang X, Teschendorff AE. An integrative network algorithm identifies age-associated differential methylation interactome hotspots targeting stem-cell differentiation pathways. Scientific reports. 2013;3:1630.
54. Jiao Y, Widschwendter M, Teschendorff AE. A systems-level integrative framework for genome-wide DNA methylation and gene expression data identifies differential gene expression modules under epigenetic control. Bioinformatics. 2014;30(16):2360–2366.
55. Steinwart I, Hush D, Scovel C. A classification framework for anomaly detection. Journal of Machine Learning Research. 2005;6(Feb):211–232.
56. Kumar A, Niculescu-Mizil A, Kavukcuoglu K, Daume III H. A binary classification framework for two-stage multiple kernel learning. arXiv preprint arXiv:12066428. 2012;.
57. Eckel-Passow JE, Lachance DH, Molinaro AM, Walsh KM, Decker PA, Sicotte H, et al. Glioma groups based on 1p/19q, IDH, and TERT promoter mutations in tumors. New England Journal of Medicine. 2015;372(26):2499–2508.
58. Noushmehr H, Weisenberger DJ, Diefes K, Phillips HS, Pujara K, Berman BP, et al. Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma. Cancer cell. 2010;17(5):510–522.
59. Zhu D, Hunter SB, Vertino PM, Van Meir EG. Overexpression of MBD2 in glioblastoma maintains epigenetic silencing and inhibits the antiangiogenic function of the tumor suppressor gene BAI1. Cancer research. 2011;71(17):5859–5870.
60. Gleize V, Alentorn A, Connen de Kérillis L, Labussière M, Nadaradjane AA, Mundwiller E, et al. CIC inactivating mutations identify aggressive subset of 1p19q codeleted gliomas. Annals of neurology. 2015;78(3):355–374.
61. Feng C, Zhang Y, Yin J, Li J, Abounader R, Zuo Z. Regulatory factor X1 is a new tumor suppressive transcription factor that acts via direct downregulation of CD44 in glioblastoma. Neuro-oncology. 2014;16(8):1078–85.
62. Bai H, Harmancı AS, Erson-Omay EZ, Li J, Coşkun S, Simon M, et al. Integrated genomic characterization of IDH1-mutant glioma malignant progression. Nature genetics. 2016;48(1):59–66.

Figures

Figure 1: **Correlation between topological overlap and cosine similarity on 250 random networks.**

Figure 2: **Sensitivity Analysis of Parameter θ.** The boxplots represents the distribution of True Positive Rate (TPR) identified by Closed-Form approach for 100 random runs of the experiment.

Figure 3: **Comparison of proposed Closed-Form approach with dGHD algorithm.** Figures A and B correspond to the ROC and PR plot for permuted sub-network ($d = 0.15$) respectively. Figure C and D represents the ROC and PR plot corresponding to denser sub-network ($d = 0.3$ and $d' = 0.5$) respectively. Clearly, the Closed-Form technique has better performance than the dGHD algorithm.

Figure 4: **Comparison of proposed Closed-Form approach with dGHD method w.r.t. AUC$_{ROC}$ and AUC$_{PR}$ for 100 random runs of the experiment.** These metrics are calculated using p-value 0.01 as cut-off. Figures A and B correspond to the AUC$_{ROC}$ and AUC$_{PR}$ for permuted sub-network ($d = 0.15$) respectively. Figures C and D represents the AUC$_{ROC}$ and AUC$_{PR}$ corresponding to denser sub-network ($d = 0.3$ and $d' = 0.5$) respectively.

Figure 5: **Differential sub-networks between IDH-mutant and IDH wild-type detected by the closed form approach.** In red the connection present only in the IDH-mutant sub-network, while in green those present only in the IDH-wild-type sub-network. In black are represented common connections.

Tables

Table 1: **Time complexity comparison** Here K represents the number of nodes for which p-value is greater than θ and generally $K \ll N$. An important remark is that the cGHD calculation after removal of each node can be done independently in parallel. So, in case we have T processors, the complexity of the proposed approach will reduce $\approx$ linearly w.r.t. T.

dGHD	Closed-Form
$O(N^2\|E\|)$	$O(N\|E\| + N\log(N) + K^2\|E\|)$

Table 2: **Comparison of proposed Closed-Form (CF) approach with dGHD algorithm** We compared the proposed Closed-Form approach with dGHD, Louvain, Infomap and Spinglass techniques w.r.t. various evaluation metrics for random geometric (RG) and power law (PL) networks. Bold represents the best results.

Parameters	Method	AUC_ROC	Precision	Recall	Accuracy	Specificity	Kappa	Time
		Mean ± Sd	Mean ± Sd	Mean ± Sd	Mean ± Sd	Mean ± Sd	Mean ± Sd	Mean
$d = 0.15$ (RG)	**CF**	0.935 ± 0.051	**0.849 ± 0.037**	0.846 ± 0.102	**0.969 ± 0.011**	**0.983 ± 0.004**	0.828 ± 0.068	0.078
$d = 0.15$ (RG)	**dGHD**	0.926 ± 0.018	0.793 ± 0.021	0.878 ± 0.036	0.965 ± 0.005	0.974 ± 0.003	0.813 ± 0.026	1.0
$d = 0.15$ (RG)	**Louvain**	**0.980 ± 0.016**	0.767 ± 0.052	**1.0 ± 0.0**	0.965 ± 0.028	0.960 ± 0.031	**0.841 ± 0.113**	**0.012**
$d = 0.15$ (RG)	**Infomap**	0.843 ± 0.012	0.262 ± 0.015	**1.0 ± 0.0**	0.718 ± 0.022	0.685 ± 0.024	0.304 ± 0.024	0.018
$d = 0.15$ (RG)	**Spinglass**	0.832 ± 0.011	0.249 ± 0.012	**1.0 ± 0.0**	0.699 ± 0.018	0.665 ± 0.021	0.285 ± 0.020	0.85
$d = 0.15$, $d' = 0.3$	**CF**	0.927 ± 0.048	0.839 ± 0.031	0.862 ± 0.098	0.969 ± 0.008	**0.982 ± 0.005**	0.825 ± 0.054	0.081
$d = 0.15$, $d' = 0.3$	**dGHD**	0.922 ± 0.022	0.806 ± 0.027	0.868 ± 0.045	0.966 ± 0.006	0.977 ± 0.004	0.816 ± 0.032	1.0
$d = 0.15$, $d' = 0.3$	**Louvain**	**0.978 ± 0.018**	**0.887 ± 0.137**	0.974 ± 0.042	**0.982 ± 0.018**	**0.982 ± 0.023**	**0.916 ± 0.083**	**0.013**
$d = 0.15$, $d' = 0.3$	**Infomap**	0.849 ± 0.008	0.269 ± 0.009	**1.0 ± 0.0**	0.728 ± 0.015	0.698 ± 0.016	0.316 ± 0.016	0.020
$d = 0.15$, $d' = 0.3$	**Spinglass**	0.859 ± 0.009	0.284 ± 0.013	**1.0 ± 0.0**	0.747 ± 0.016	0.719 ± 0.017	0.339 ± 0.019	0.92
$d = 0.3$ (RG)	**CF**	**0.877 ± 0.067**	**0.714 ± 0.075**	0.789 ± 0.135	**0.947 ± 0.016**	**0.975 ± 0.011**	**0.716 ± 0.099**	0.083
$d = 0.3$ (RG)	**dGHD**	0.724 ± 0.029	0.645 ± 0.049	0.577 ± 0.059	0.921 ± 0.007	0.971 ± 0.006	0.504 ± 0.051	1.0
$d = 0.3$ (RG)	**Louvain**	0.866 ± 0.019	0.406 ± 0.061	**1.0 ± 0.0**	0.850 ± 0.034	0.833 ± 0.038	0.505 ± 0.072	**0.013**
$d = 0.3$ (RG)	**Infomap**	0.677 ± 0.011	0.147 ± 0.004	**1.0 ± 0.0**	0.419 ± 0.019	0.354 ± 0.022	0.100 ± 0.008	0.021
$d = 0.3$ (RG)	**Spinglass**	0.678 ± 0.011	0.148 ± 0.004	**1.0 ± 0.0**	0.420 ± 0.018	0.355 ± 0.021	0.100 ± 0.008	0.90
$d = 0.3$, $d' = 0.5$	**CF**	**0.979 ± 0.005**	**0.771 ± 0.061**	0.930 ± 0.082	**0.965 ± 0.012**	**0.969 ± 0.011**	**0.821 ± 0.062**	0.09
$d = 0.3$, $d' = 0.5$	**dGHD**	0.848 ± 0.071	0.700 ± 0.038	0.731 ± 0.148	0.941 ± 0.010	0.964 ± 0.009	0.672 ± 0.078	1.0
$d = 0.3$, $d' = 0.5$	**Louvain**	0.932 ± 0.029	0.478 ± 0.118	**1.0 ± 0.0**	0.879 ± 0.054	0.866 ± 0.059	0.582 ± 0.128	**0.014**
$d = 0.3$, $d' = 0.5$	**Infomap**	0.674 ± 0.010	0.145 ± 0.004	**1.0 ± 0.0**	0.413 ± 0.018	0.348 ± 0.020	0.097 ± 0.008	0.023
$d = 0.3$, $d' = 0.5$	**Spinglass**	0.711 ± 0.007	0.162 ± 0.003	**1.0 ± 0.0**	0.481 ± 0.013	0.423 ± 0.014	0.128 ± 0.006	0.94
$\alpha = 2$ (PL)	**CF**	**0.797 ± 0.046**	**0.307 ± 0.307**	0.792 ± 0.099	**0.801 ± 0.018**	**0.349 ± 0.051**	**0.802 ± 0.022**	0.09
$\alpha = 2$ (PL)	**dGHD**	**0.797 ± 0.013**	0.294 ± 0.009	0.809 ± 0.027	0.787 ± 0.008	0.333 ± 0.015	0.784 ± 0.009	1.0
$\alpha = 2$ (PL)	**Louvain**	0.780 ± 0.014	0.212 ± 0.010	**1.0 ± 0.0**	0.703 ± 0.018	0.272 ± 0.016	0.690 ± 0.011	**0.015**
$\alpha = 2$ (PL)	**Infomap**	0.665 ± 0.013	0.141 ± 0.004	**1.0 ± 0.0**	0.603 ± 0.018	0.162 ± 0.012	0.484 ± 0.019	0.026
$\alpha = 2$ (PL)	**Spinglass**	0.687 ± 0.014	0.153 ± 0.006	**1.0 ± 0.0**	0.645 ± 0.021	0.194 ± 0.011	0.527 ± 0.016	0.90
$\alpha = 3$ (PL)	**CF**	**0.825 ± 0.019**	**0.345 ± 0.015**	0.825 ± 0.035	**0.826 ± 0.007**	**0.402 ± 0.024**	**0.826 ± 0.004**	0.085
$\alpha = 3$ (PL)	**dGHD**	0.808 ± 0.027	0.327 ± 0.018	0.799 ± 0.050	0.816 ± 0.008	0.375 ± 0.031	0.817 ± 0.004	1.0
$\alpha = 3$ (PL)	**Louvain**	0.774 ± 0.015	0.233 ± 0.011	**1.0 ± 0.0**	0.736 ± 0.019	0.301 ± 0.009	0.732 ± 0.019	**0.015**
$\alpha = 3$ (PL)	**Infomap**	0.670 ± 0.014	0.168 ± 0.005	**1.0 ± 0.0**	0.635 ± 0.017	0.210 ± 0.014	0.532 ± 0.014	0.027
$\alpha = 3$ (PL)	**Spinglass**	0.694 ± 0.013	0.179 ± 0.007	**1.0 ± 0.0**	0.670 ± 0.023	0.232 ± 0.012	0.571 ± 0.017	0.94

Table 3: **The top most different transcription factors subnetworks detected between IDH-mutant and IDH-wild-type networks.** The first four columns report differential measures in terms of Z-score of the proposed differencing test (equation (2)), GHD computed between the two networks, the mean of the null GHD distribution. The last three columns report the False Discovery Rate of the Fisher exact test obtained with a Master Regulator Analysis, and the mean of transcription factor activity in IDH-mutant and IDH-wild-type cases. Transcription factor activity explains whether the transcription factor regulates directly (> 0) or inversely (< 0) its targets in the given condition.

TF	Z-score	GHD	μ_π	FDR	ActivityIDH	ActivityWT	TFMEA
FOXJ3	0.000	1.000	1.000	6.777E-09	-0.740	0.522	2.76E-34
NFIA	0.000	1.000	1.000	4.447E-02	0.545	0.487	-
MLX	0.000	1.000	1.000	1.083E-01	-1.900	-0.648	5.39E-03
FOXD3	0.000	1.000	1.000	1.000E+00	-1.071	-0.047	6.63E-30
ETV1	0.062	0.058	0.058	1.000E+00	1.670	1.161	3.6E-02
E2F1	0.085	0.058	0.058	1.000E+00	-2.018	-1.204	1.50E-07
CREB1	0.208	0.058	0.058	1.000E+00	1.097	0.924	3.92E-05
SOX10	0.234	0.058	0.058	1.573E-07	0.858	-1.130	-
KLF13	0.338	1.000	0.278	8.086E-04	1.014	-0.255	4.9E-02
STAT3	0.354	0.058	0.058	1.112E-31	-0.318	1.335	-
RUNX3	0.387	0.058	0.059	7.192E-05	-1.503	-0.029	3.28E-02
IRF3	0.406	0.840	0.455	2.356E-13	-1.505	0.142	1.24E-10
ZNF354C	0.498	0.058	0.057	1.000E+00			2.95E-04
HOXD13	0.540	0.059	0.059	8.705E-07	-1.840	-0.223	6.96E-11
ZIC1	0.622	0.058	0.058	3.319E-19	-2.752	0.475	1.17E-03
HOXA2	0.700	0.059	0.059	2.541E-02	-1.388	0.201	-
FOXO1	0.743	0.058	0.058	2.572E-02	-2.344	-0.687	3.95E-09
MAFG	0.817	0.862	0.467	1.000E+00	0.739	-0.100	-
RFX1	0.865	0.059	0.059	1.768E-01	-0.060	0.958	-
NR1H2	0.871	0.058	0.058	1.000E+00	-2.363	-0.399	-
PAX6	1.003	0.058	0.057	9.057E-01	2.209	1.416	1.60E-03
GLIS2	1.035	0.058	0.059	4.905E-01	0.332	-0.699	8.16E-21
NR4A2	1.118	0.058	0.058	1.000E+00	-0.169	-0.318	-
STAT4	1.137	0.848	0.486	9.025E-01	-0.929	-1.049	-
DLX6	1.208	0.058	0.059	1.000E+00			-
SIX4	1.232	0.058	0.058	5.592E-03	2.040	0.004	-
MEF2D	1.379	0.058	0.059	1.567E-01	0.406	-0.583	5.38E-20
MTF1	1.388	0.058	0.057	1.000E+00	0.401	0.389	4.25E-32
MBD2	1.480	0.820	0.495	2.330E-10	-1.488	0.070	-
OTP	1.493	0.058	0.057	2.156E-05	-1.017	0.911	-
ETV4	1.529	0.059	0.059	1.661E-01	-0.874	0.782	5.0E-02
ZBTB12	1.566	0.194	0.189	4.556E-03	1.304	-0.362	-
HOXB4	1.595	0.058	0.057	6.416E-03	-2.019	-0.345	-
PLAG1	1.622	0.195	0.190	7.846E-05	-2.224	-0.727	1.18E-04
E2F6	1.668	0.197	0.192	6.927E-01	-0.677	0.305	7.66E-05
CREM	1.674	0.765	0.506	1.408E-01	-1.594	-0.324	-
IRF9	1.700	0.058	0.057	1.000E+00	0.302	0.675	-
KLF6	1.709	0.059	0.059	9.536E-07	-1.378	0.333	-
TFE3	1.716	0.199	0.193	1.000E+00	0.523	1.448	-
HSF2	1.759	0.201	0.195	1.802E-09	1.145	-0.669	-
NR2C1	1.800	0.058	0.058	1.000E+00	-0.147	-0.380	-
ONECUT2	1.804	0.202	0.196	4.162E-06	0.709	-1.050	4.22E-05
HOXD3	1.847	0.204	0.198	4.984E-02	-1.535	-0.701	-
BACH1	1.888	0.058	0.059	1.000E+00	-0.565	0.223	-
GSX1	1.895	0.207	0.200	1.000E+00			-