A comparison of mechanistic signaling pathway activity analysis methods

1. Introduction

Pathway Analysis (PA), also called functional enrichment analysis, is becoming more and more important in Omics research, which concerns biological branches that end with the suffix -omics. Pathway Analysis methodologies combine knowledge from gene expression analysis and molecular pathway networks to discover strongly impacted pathways in a given condition and better understand the biological significance of differentially expressed genes and proteins.

The term Pathway refers to the graphical representation of molecular interaction, reaction, and relationship networks. The graph consists of nodes, which correspond to genes, proteins, and/or molecules, and directed edges, which represent relationships and interactions between the nodes. The states of each gene are either on or off, indicating whether the gene is expressed or not expressed respectively. The types of interactions between the nodes vary. Activation, inhibition, and catalysis are some examples of different sorts of interactions between nodes.

Pathway analysis techniques discover the pathways that are strongly impacted in a specific circumstance by combining available pathway databases with gene expression data. KEGG, Reactome, and BioCarta are some of the sources that include thorough information about pathways. The methods analyzed in the present paper mainly use the KEGG, BioCarta and Reactome databases. KEGG is a repository of hand-drawn pathway maps that provide information about genomes, biological pathways, diseases, pharmaceuticals, and chemical compounds, while BioCarta's interactive online services give the means to see how genes communicate in dynamic graphical models. Reactome database’s mission is to provide user-friendly bioinformatics tools for visualizing, interpreting, and analyzing pathway data in support of fundamental and medical trials, genomic analysis, modeling, systems biology, and education.

We can distinguish two main approaches to pathway analysis: the first one takes into consideration only the expression levels of the genes of a pathway, while the second one takes advantage of the pathway topology as well, known as topology-based.

The aim of this study is to identify similarities and differences between distinct pathway activity analysis tools, which are based on statistical and machine learning methods.

2. Pathway analysis tools

2.1. TAPPA

TAPPA is a java-based tool that uses pathway topological measurements to identify phenotype-associated genomic circuits. This is accomplished by calculating a Pathway Connectivity Index (PCI) for each pathway and then assessing its relationship to phenotypic variance.

$$PCI = \sum_{i=1}^{N}\sum_{j=1}^{N} sgn\left(x_{is} + x_{js}\right) * |x_{is}|^{0.5} * \alpha_{ij} * \left|x_{js}\right|^{0.5}$$

The TAPPA tool was developed using JAVA and can handle both binary and numerical attributes. In the case of the binary traits, the Mann–Whitney test is used to assess the significance of the relationship between network PCI and phenotype, while the Spearman correlation is used for continuous attributes. In addition, a permutation test is used to assess the false discovery rate (FDR).

Subsequently, using different zoom ratios, the pathway is visualized and the genes that are strongly related to the phenotype are identified. Eventually, the relationship between the phenotypes of the sub-modules in a pathway is studied and determines the biological significance of the genes involved.

2.2. SPIA

Signaling Pathway Impact Analysis, also known as SPIA, uses data collected from the classical enrichment analysis and combines them with data that evaluate the perturbation of a certain pathway under a specific circumstance.

Impact Analysis considers the overrepresentation of DE genes in a particular pathway and the anomalous alteration of that pathway, as determined by propagating observed expression changes across the pathway topology. These features correspond to two separate probabilities, P_{NDE} and P_{PERT}.

The $P_{NDE} = P(X \geq N_{de}|H_0)$ probability represents the importance of a pathway P_i, based on an overrepresentation study of the number of DE genes (NDE) found on the pathway.

P_{PERT} probability results from the degree of perturbation in each pathway, which is calculated using the following gene perturbation function:

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

The above formula takes into consideration the type of relationship between two genes and is represented by the term β_{ij}. If the value of β equals to +1, then the type of interaction is activation, while -1 corresponds to inhibition and repression.

The net perturbation accumulation at the gene level, Acc_g, is then calculated as the difference between a gene's perturbation factor PF and its observed log fold-change.

$$Acc(g_i) = PF(g_i) - \Delta E(g_i)$$

Subsequently, the total net accumulated perturbation of a pathway is calculated as the sum of the net perturbation accumulation of each gene.

$$t_A = \sum_i Acc(g_i)$$

The possibility of seeing a total accumulated perturbation of the pathway, T_A, greater than t_A, is represented by the PPERT probability:

$$P_{PERT} = P(T_A \geq t_A|H_0)$$

Eventually, P_{NDE} and P_{PERT} are integrated into a global probability value, P_G. Through this probability value the pathways are ranked, and the hypothesis is tested to see if the pathway is significantly disrupted in the study condition.

$$P_G = c_i - c_i \cdot \ln(c_i)\,, c_i = P_{NDE}(i) \cdot P_{PERT}(i)$$

P_G can also set the level of type I error. It is recommended to use the common FDR approach to keep the false discovery rate (FDR) of the pathway analysis at 5%.

2.3. TopologyGSA

This method evaluates the differential expression of a pathway using graphical models. Then it illustrates the components of the pathway that are implicated in the deregulation. Below is a detailed description of the technique.

In this project, KEGG maps are employed, since they provide a good ratio between map accuracy and simplicity. Initially, the paths obtained from the KEGG repository are transformed into a graphical model. This is accomplished by using the following basic steps: i) simple directed edges include inhibition, phosphorylation (+p), and dephosphorylation (-p); ii) BioCarta network provides extensive information that can be used to direct undirected edges and iii) when it comes to complexes (nodes consisted of several gene products), the first principal component is defined as the complex's expression. The data of the same pathway are represented in distinct experimental states as implementations of undirected graphical Gaussian models with the same undirected graph G. For instance, in the case of two scenarios, we employ the Gaussian models

$$\mathcal{M}_1(G) = \{Y \sim N_p(\mu_1, \Sigma_1), \Sigma_1^{-1} \in S^+(G)\},$$

$$\mathcal{M}_2(G) = \{Y \sim N_p(\mu_2, \Sigma_2), \Sigma_2^{-1} \in S^+(G)\}.$$

In this case, p refers to the number of genes (nodes of the graph), while $S^+(G)$ represents the array of symmetric positive definite matrices with null components indicating the missing connections of G.

The estimated covariance matrices are calculated by using a technique known as the Iterative Proportional Scaling Technique (IPS) for graph analysis, which ensures that the estimated matrices are positive definite and that their inverse has null elements correlating to the graph's missing edges. The sample covariances matrices can be generated, either from the chip covariance matrices by removing elements referring to all the pairs of genes in the pathway, or by determining the expression levels of the genes on the pathway and measuring the sample covariances.

Then, to compare the gene sets among two experimental conditions, the null hypothesis testing method is employed. The strength of the connections that define a pathway can vary under different situations, causing changes in the pathway's expression. The equivalence of two means is the corresponding hypothesis. The evaluation is determined by whether the models' covariances, which are often unknown, are homogeneous. As a result, the choice of the homogeneity hypothesis has an impact on the analysis of the means.

Eventually, the strength of the gene relationships in two experimental conditions is put to the test to see if they are equal. This is easily accomplished in the context of graphical Gaussian models by comparing the two concentration matrices (opposite of the covariance matrices), which include all the details about the underlying structure. As a result, the focus is on putting the hypothesis $\Sigma_1^{-1} = \Sigma_2^{-1}$ to the test.

The methods for comparing covariance matrices are then applied to the specific instance of graphical Gaussian models in the following methodology. Assume you have $\gamma_1 = (\gamma_1^j)$, $j = 1, \dots, n_1$ observations from $N_p(0, \Sigma_1)$, and $\gamma_2 = (\gamma_2^j)$, $j = 1, \dots, n_2$ observations from $N_p(0, \Sigma_2)$, with $\Sigma_1^{-1} = K_1 \in S^+(G)$ and $\Sigma_2^{-1} = K_2 \in S^+(G)$ without losing generality. The hypothesis to be tested is $H_0: K_1 = K_2$ against $H_1: K_1 \neq K_2$. When the value of W_i is determined using the function $W_i = \sum_{j=1}^{n_i} (\gamma_i^j)(\gamma_i^j)^T$, $i = 1,2$, the likelihood function, $L(K_1, K_2)$, yields as follows:

$$L(K_1, K_2) = \prod_{i=1}^{2} (2\pi)^{-\frac{n_i p}{2}} (\det K_i)^{\frac{n_i}{2}} e^{-\frac{1}{2} tr(K_i W_i)}$$

Starting with the pooled covariance matrix $S = (n_1 + n_2 - 2)^{-1} \cdot \{(n_1 - 1) \cdot S_1 + (n_2 - 1) \cdot S_2\}$ and the null hypothesis, the technique calculates the estimate, $\hat{\Sigma}$, of the common covariance matrix. On the contrary, under the alternative hypothesis, the sample covariance matrices, $S_1 = (n_1 - 1)^{-1} \cdot W_1$ and $S_2 = (n_2 - 1)^{-1} \cdot W_2$, are used, so that the values of $\hat{\Sigma}_1$ and $\hat{\Sigma}_2$ are calculated.

After making several assumptions, the likelihood ratio test, Λ, is computed using the following formula:

$$\Lambda = \frac{L_{H_0}(\hat{K}_1, \hat{K}_2)}{L_{H_1}(\hat{K}_1, \hat{K}_2)} = \frac{L_{H_0}(\hat{K})}{L_{H_1}(\hat{K})}$$

It is also true that by letting $W = W_1 + W_2$ and taking advantage of the fact that $tr(\hat{K}_i W_i) = n_i tr(\hat{K}_i \hat{K}_i^{-1}) = n_i p$ and $tr(\hat{K} W) = (n_1 + n_2) tr(\hat{K} \hat{K}^{-1}) = (n_1 + n_2) p$, the following two formulas emerge.

$$\Lambda = \prod_{i=1}^{2} \left(\frac{\det \hat{K}}{\det \hat{K}_i} \right)^{\frac{n_i}{2}}$$

$$-2 \log \Lambda = \sum_{i=1}^{2} n_i \log \left(\frac{\det \hat{K}_i}{\det \hat{K}} \right)$$

If indeed the null hypothesis is false, the graphical methodology enables us to evaluate the causes of differences between the two concentration matrices. In particular, if the graph is decomposable, it is feasible to break it down into its maximal complete subgraphs (cliques) and perform the preceding test for each clique. Following the standard procedures, the equivalence of the covariance matrices on cliques can be evaluated. However, if the graph is not divisible, more edges can be introduced to create a new triangulated and thus decomposable graph. The graph's cliques can then be used to conduct the test.

Eventually, the pathway's differential expression is examined. The differential expression of the pathway, if the null hypothesis is not rejected, is evaluated by hypothesis

$$H_0\colon \mu_1 = \mu_2 \text{ subject to } \Sigma_1 = \Sigma_2.$$

Exact approaches, such as multivariate analysis of variance, can be used to carry out this test. If the null hypothesis of homogeneity is rejected, the hypothesis that must be tested is

$$H_0\colon \mu_1 = \mu_2 \text{ subject to } \Sigma_1 \neq \Sigma_2.$$

In a two-sample scenario with unequal covariance matrices, this is the standard test for equality of means, also known as the Behrens-Fisher problem.

2.4. PARADIGM

PARADIGM stands for Pathway Recognition Algorithm using Data Integration on Genomic Models and aims to infer the activity of genetic pathways from integrated patient data.

The implementation starts by creating a separate probabilistic model for each pathway. A factor graph was created using a pathway diagram that included both concealed and observable states. Also, we employ variables to characterize the states of entities in a cell to illustrate a biological pathway with a factor graph.

The factor graph uses a random variable $X = x_1, x_2, \dots, x_n$ for each entity and a set of m non-negative functions, or factors, to restrict the entities' ability to take on biologically meaningful values as functions of one another to represent the status of a cell. A probability distribution over a subset of entities $X_j \subset X$ is defined by the j-th factor ϕ_j. The joint probability distribution over all the entities is encoded in the whole graph of entities and factors as follows:

$$P(X) = \frac{1}{Z}\prod_{j=1}^{m} \phi_j(X_j)$$

Each entity can be active, nominal, or deactivated in relation to a control level, and these states are encoded as 1, 0 or -1 correspondingly.

To make factor building easier, we turn the pathway into a directed graph, with each edge annotated with a positive or negative influence. In the directed graph, every interaction in the pathway is turned to a single edge. We next create a list of factors to define the factor graph using this directed graph. Eventually, we complete the integration of pathway and multi-dimensional functional genomics data by adding observation variables and factors to the factor graph.

Subsequently, we want to know if a particular hidden entity x_i is likely to be in state α based on patient data.

$$P(x_i = \alpha \parallel \Phi) = \frac{1}{Z}\prod_{j=1}^{m} \sum_{S \sqsubset_{A_i(\alpha)} X_j} \phi_j(S)$$

Similarly, the likelihood that x_i is in state α, based on all the patient's observations, is:

$$P(x_i = \alpha \parallel \Phi) = \frac{1}{Z}\prod_{j=1}^{m} \sum_{S \sqsubset_{A_i(\alpha) \cup D} X_j} \Phi_j(S)$$

The expectation-maximization (EM) technique is used to estimate the parameters of the observation factors. A factor graph for each patient is generated, the patient's data are applied, and EM runs until the likelihood changes by less than 0.1 percent for each pathway. The factors learned from each pathway were averaged, and the final posterior estimates for each variable were calculated using these parameters.

Afterwards, a matrix of Integrated Pathway Activities (IPA) is generated for each variable with an ‘active’ molecular type after inference. A log-likelihood ratio that describes the level to which a patient's data boosts our opinion that entity *i*'s activity is up or down is calculated based on the following formula.

$$L(i,\alpha) = \log\left(\frac{P(D, x_i = \alpha|\Phi)}{P(D, x_i \neq \alpha|\Phi)}\right) - log\left(\frac{P(x_i = \alpha|\Phi)}{P(x_i \neq \alpha|\Phi)}\right) = log\left(\frac{P(D|x_i = \alpha, \Phi)}{P(D|x_i \neq \alpha, \Phi)}\right)$$

Based on the log-likelihood ratio, a single IPA for gene *i* is computed as follows:

$$IPA(i) = \begin{cases} L(i,1)\,, & L(i,1) > L(i,-1) and\, L(i,1) > L(i,0) \\ -L(i,-1)\,, & L(i,-1) > L(i,1)\; and\; L(i,-1) > L(i,0) \\ 0\,, & otherwise \end{cases}$$

The IPA score is a signed equivalent of the log-likelihood ratio, L. The IPA is set to L if the gene is more likely to be activated. If the gene is more likely to be inhibited, the IPA is set to -L; otherwise, it is set to 0.

Two alternative permutations of the data are used to measure the significance of IPA scores. A permuted data sample is constructed for the 'within' permutation by selecting a random tuple of data first from a random real sample, and then from a random gene within the same network, until tuples have been selected for each gene in the pathway, while the technique for the 'any' permutation is much like the 'within' permutation method, except the random gene selection stage could pick a gene from anywhere in the genome. In both cases, 1000 permuted samples are constructed, and perturbation scores are calculated for each permuted sample. To evaluate the significance of real samples, the distribution of perturbation scores from permuted samples is employed as a null distribution.

2.5. GGEA

Gene Graph Enrichment Analysis (GGEA) is a method that uses previous information acquired from directed gene regulation networks to find consistently and coherently enriched gene sets.

The method on which GGEA is based consists of three critical stages. Initially, to create an induced subnetwork, the gene set is first mapped into the fundamental regulatory network. This is the part of the network that is impacted, which is made up of edges that involve members of the gene set. Next, each edge of the induced network is evaluated for consistency with the expression data, i.e., the signals of two interaction partners' expression changes are compared to the regulatory type (activation/inhibition) of the link connecting both genes. Finally, using a permutation process, the edge consistencies are aggregated over the induced network, normalized, and significance evaluated.

Consistency is calculated using the following formula:

$$C(t) = cons(de_O, f_t(de_i))$$

The raw GGEA consistency score *S* is induced by summing the consistency of all gene regulatory network (GRN) transitions and then normalized by the number of transitions, to compensate for the GRN's size.

$$S := \sum_{t \epsilon T_u} C(t)$$

$$\bar{S} := \frac{S}{|T_u|}$$

Eventually, for each gene set we estimate the consistency P-value and rank the gene sets based on the adjusted P-values. Gene sets that are significantly and persistently enriched fall below the predefined significance level.

2.6. HotNet

HotNet is another approach for detecting significantly altered subnetworks in a large gene interaction network, that was initially designed for cancer mutation data. The method for identifying cancer pathways that have been significantly mutated is presented below.

First, the model that will be used for the interpretation of the data is defined. Graph $G = (V, E)$ is used to model the interaction network, while $T \subseteq V$ corresponds to a subgroup of the genes that were tested. Each g gene is classified as either mutant or normal. The notation M_i is used to describe a subset of mutant genes in group *T*, while S_j denotes the samples under which the gene $g_j \in T$ has been altered, and m indicates the total number of mutant genes detected across all samples. A linked subgraph of *G* is defined as the resultant pathway.

After that, the influence graph, that encodes the knowledge in the interaction network, is constructed. The importance of a subnetwork is determined by (i) the number of samples with mutations in the subnetwork's genes, and (ii) the linkages among genes in the subnetwork in the context of the overall network's topology. On the interaction network, a diffusion process is employed to create a strict level of influence across all network nodes. The procedure outlined by Qi et al. (2008) is used to calculate the effect of node *s* on all other nodes in the network, and therefore, the influence graph $G_I = (T, E_I)$ with the collection of nodes belonging to the subset of tested genes is obtained. The weight of each edge $w(g_j, g_k) = \min[i(g_k, g_j), i(g_j, g_k)]$ is also considered.

Then, to discover altered paths, a combinatorial model is developed. First, collections of nodes in the influence graph G_I, that are (1) related by high-influence links and (2) relate to mutated genes in many samples, are selected. A threshold δ is determined and, by deleting all edges with $w(g_i, g_j) < \delta$ and all nodes belonging to genes in the sample data with no modifications, a reduced impact graph $G_I(\delta)$ of G_I is constructed. Consequently, the size of the identified related subgraphs is determined by a threshold δ, which is entirely reliant on the null hypothesis. The connected maximum coverage problem, which is an NP-hard problem, is analogous to discovering the linked subgraph of k genes that is mutated in the maximum number of samples. To make the algorithm run properly, a modified version of the combinatorial algorithm shown in Figure 1 is used, in which for each pair of nodes (u, v), all the shortest paths between u and v are evaluated, and the one that optimizes $\frac{|\boldsymbol{P_v(u)}|}{|\boldsymbol{\ell_v(u)}|}$ is preserved.

Combinatorial Algorithm

Input: Influence graph G_I and parameters δ and k

Output: Connected subgraph C of $G_I(\delta)$ with k vertices

1. Construct G_I (δ) by removing from G_I all edges with weight < δ;
2. $\boldsymbol{C} \leftarrow \emptyset$;
3. **for** each node $\boldsymbol{v} \in \boldsymbol{V}$ **do**
4. $\quad \boldsymbol{C_v} \leftarrow \{\boldsymbol{v}\}$;
5. $\quad$ **for** each $\boldsymbol{u} \in \boldsymbol{V}\{\boldsymbol{v}\}$ **do** $\boldsymbol{p_v(u)} \leftarrow$ shortest path from $\boldsymbol{v}$ to u in $G_I(\delta)$;
6. $\quad$ **while** $|\boldsymbol{C_v}| < \boldsymbol{k}$ **do**

   $//\boldsymbol{\ell_v(u) = set\ of\ nodes\ in\ p_v(u);\ P_v(u) = elements\ of\ I\ covered\ by\ \ell_v(u);\ P_{C_v} = elements\ covered\ by\ C_v;\ P_C = elements\ covered\ by\ C}$

7. $\quad\quad \boldsymbol{u \leftarrow arg\ max_{u \in V \setminus C_v : |\ell_v(u) \cup C_v| \leq k} \left\{ \frac{|P_v(u) \setminus P_{C_v}|}{|\ell_v(u) \setminus C_v|} \right\}}$;
8. $\quad\quad \boldsymbol{C_v \leftarrow \ell_v(u) \cup C_v}$;
9. $\quad$ **if** $|\boldsymbol{P_{C_v}}| > |\boldsymbol{P_C}|$ **then** $\boldsymbol{C \leftarrow C_v}$
10. **Return** C;

Figure 1: Pseudocode for the combinatorial model's algorithm

Subsequently, to detect mutated subnetworks, a computationally efficient enhanced influence model is generated. The Enhanced Influence Model is rooted in the idea of increasing the influence measure among genes by the number of mutations found in each of these genes, and then breaking the resulting enhanced influence graph into linked components.

H refers to the enhanced influence graph. All genes g_j having at least one mutation in the data make up the set V_H of H's vertices, while the improved influence

$$h(g_j, g_k) = w(g_j, g_k) \times \max\{|S_j|, |S_k|\}$$

determines the weight of the edge (g_j, g_k). Then, to produce a graph $H(\delta)$, whose linked components represent the significant subnetworks, any edges with a weight less than a threshold δ are eliminated.

Eventually, a statistical analysis is performed to determine the network's significance. There are two null hypothesis distributions considered:

i. H_0^{sample} in which $m = \sum_i |M_i|$ mutations are randomly distributed throughout the nodes correlating to the $|T|$ tested genes
ii. H_0^{gene} which is obtained by permuting the identities of the network's evaluated genes, using a random permutation σ

A two-stage multi-hypothesis test is executed and the Family Wise Error Rate (FWER), that is the probability of making at least one Type I error in any of the tests, is used as the rigorous indicator of its significance level. The False Discovery Rate is a less conventional alternative to minimizing errors in multiple testing (FDR). It is denoted as $FDR = E[V/R]$, where V represents the number of Type I errors and R represents the total amount of null hypotheses excluded. The two-stage test identifies several subnetworks in the data as statistically significant with low FDR values.

Using a similar approach to that used in the Combinatorial model, it is shown in this study how the number of hypotheses can be limited to merely $K = |T|$ hypotheses. The first stage of evaluating each hypothesis with confidence level α/K determines the smallest size s, such that the null hypothesis that the number of linked components of size $\geq s$ detected in the graph H(d), r_s, may be rejected with confidence level α. The test also includes a second criterion that ensures that the FDR is kept within a certain range.

A Monte-Carlo simulation ("permutation test") or analytical bounds can be used to calculate the null hypothesis distributions. Two properties of the Monte-Carlo simulation approach considerably minimize the cost of the estimations. The Monte Carlo simulation must be done on the graph G_I. The *p*-value of the distribution of the number of connected linked subgraphs/components of a particular size is used in the statistical test. As a result, it is essential to determine *p*-values that are a magnitude larger for this test, using vastly fewer simulation rounds.

Using analytical bounds, the null hypothesis can be approximated for a greater number of tested genes. For any node g_i in G_I, the maximum δ is set such that the weight of less than $\alpha M/|T|$ connected edges gratify $s_{max} w(g_i, g_j) \geq \delta$, for any given $\alpha < 1$.

2.7. PRS

Pathway Regulation Score, or else PRS, is a method which distinguishes between essential processes in real-world biological datasets. The procedure is as follows.

The data were first pre-processed using the Robust Multiarray Average (RMA) approach, and the DEG lists were produced using simple fold change and *p*-value calculations. The pathways retrieved by the KEGG database were represented in the form of a graph. Due to redundancies in KEGG pathways, fold-change values for a node may be assigned to a route several times, resulting in a skewed PRS calculation. Consequently, a new structure emerged, in which duplicated genes were unified into a single term with a unique ID.

In order to implement the PRS algorithm, the pathways were represented as networks, so that each pathway is characterized by a unique identity, definition name, and its corresponding nodes. Particularly, a pathway's nodes are described by the following attributes. *Node_genes* correspond to a distinct function that maps to one or more transcripts and *Node_value* (*NV*) represents a value based on expression data. *Node_weight* (*NW*) concerns only the significant nodes and indicates their structural strength. The *Node_Score* (*NS*) is calculated by combining the *NV* and *NW* values.

$$NS = \begin{cases} NV * NW & if\ NV > 1 \\ 0 & if\ NV \leq 1 \end{cases}$$

Subsequently, using the following formula the PRS is determined:

$$PRS(p_i) = \sum_{j=1}^{n_i} NS_j \ .$$

Prior to rating the paths, a normalization step is performed to account for two crucial features.

i. Pathway size: the bias caused by pathway size was reduced as seen in the following equation.

$$PRS(p_i) = \left(\sum_{j=1}^{n_i} NS_j\right) * \left(\frac{NDEGs_i}{NEGs_i}\right)$$

ii. Pathway-specific PRS score null distributions that contribute to statistical bias: a nonparametric permutation approach is employed to determine the null distributions of raw PRS values acquired for each pathway.

$$nPRS_i = \frac{PRS_i - mean(pPRS_i)}{STD(pPRS_i)}$$

$$npPRS_{ij} = \frac{pPRS_{ij} - mean(pPRS_i)}{STD(pPRS_i)}$$

To determine the significance, the PRS values were recalculated using the equation used to reduce the bias caused by the pathway size, after the fold-change values for the full gene list were permuted and mapped back onto pathways. To construct a null distribution of each raw score, this procedure was repeated 1000 times. Then, the normalized scores were compared, and the p-values were determined as shown below:

$$P(nPRS_i) = \frac{\sum_{j=1}^{n} I\left(npPRS_{ij} \geq nPRS_i\right)}{n}$$

Finally, a multiple test adjustment was implemented, and the FDR modified P_{final} to account for type I errors.

2.8. DEGraph

DEGraph is yet another pathway analysis tool, that uses modern hypothesis testing approaches to predict whether a specific gene network is differentially expressed between two scenarios and is very useful in cancer research.

First, a lower-dimension basis is constructed, after which the multivariate test of means is used. The testing question of whether two sets of random vectors of gene expression measures are expected to have emerged from equal-mean distributions, can be directly formulated, and solved using multivariate statistics.

A network of p genes is depicted as graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, having $|\mathcal{V}| = p$ nodes and edge set $\mathcal{E}$, while δ refers to the mean shift, to wit, the vector of differences between the p genes' mean expression values among the two study populations. Afterwards, a lower-dimensional $k \ll p$ space is constructed, retaining most of the low-energy functions $E_{\mathcal{G}}(\delta)$. To accomplish this, the function with the least potential energy is identified, followed by the function with the lowest possible energy in the orthogonal space of the last one, and so on, up to the *k*th function with the minimum energy in the orthogonal subspace of the first $k-1$ functions.

$$u_i = \begin{cases} \arg\min_{f \in \mathbb{R}^p} E_{\mathcal{G}}(f) \\ such\ that\ u_i \perp u_j, j < i \end{cases}, i \leq k$$

The following energy function states that if the variation in mean expression of any gene among the two populations is equivalent to the (signed) average of the difference between the mean expression for the genes that either activate or inhibit it, an expression shift will have limited power:

$$E_{\mathcal{G}}(\delta) = \sum_{i:d_i^- \neq 0}^{p} \left(\delta_i - \frac{1}{d_i^-} \sum_{(j,i)\epsilon\mathcal{E}} a_{ji}\delta_j \right)^2$$

The number of directed edges leading from any node to u_i is denoted by d_i^-.

Then, to achieve orthonormal functions with low intensity, the first few eigenvectors of $M_{\mathcal{G}}$ are employed to construct a lower dimension space.

Following that, a graph-structured two-sample test statistic is demonstrated. Hotelling's T^2-test, a classic location shift test, is a consistently most powerful invariant against global-shift alternatives for multivariate normal distributions. The statistical test $T^2 = \frac{n_1 n_2}{n_1+n_2}(\bar{x}_1 - \bar{x}_2)^{\mathrm{T}}\hat{\Sigma}^{-1}(\bar{x}_1 - \bar{x}_2)$ is predicated on the sample mean shift's squared *Mahalanobis norm*. In this work, T^2-statistics follow the nominal *F*-distribution, while Hotelling's test in the new area limited to its first k components is said to generate greater power than testing in the complete new space.

Subsequently, a systematic way for identifying nonhomogeneous subgraphs, or subgraphs of a large graph with a significant shift in means, is to examine each one individually. Due to the huge combinatorial issue posed by large sizes of graphs, it's critical to rapidly discover groups of subgraphs that all fit the null hypothesis of equal means. This is achieved by using a threshold on the value of the test statistic for every subgraph containing a particular network. The corresponding algorithm is described below.

Nonhomogeneous subgraph discovery algorithm

Input: $\mathcal{G}$, X_1, X_2, α, q

Output: selectedSubgraphs

selectedSubgraphs = ∅;
previousSubgraphs = nodes ($\mathcal{G}$);
prunedSubgraphs = ∅;
foreach $s \in \{1 \ldots q-1\}$ **do**
  checkedSubgraphs = ∅;
  foreach previousSubgraph **do**
    foreach subgraph ∈ subgraphBoundary(previousSubgraph) **do**
      if subgraph *has been checked or has a pruned subgraph* **then** next**;**
      if $s < q-1$ **then**
        if $\boldsymbol{upperBound(subgraph, \mathcal{G}, X_1, X_2, q) < T^2_{\alpha,k}}$ **then**
          add subgraph to prunedSubgraphs;
        else
          add subgraph to currentSubgraphs;
        end
      else
        foreach q-subgraph ∈ subgraphBoundary(subgraph) **do**
          if q-subgraph *has been checked or has a pruned subgraph* **then** next
          else
            if $\boldsymbol{\widetilde{T}^2_k(q - subgraph, X_1, X_2) > T^2_{\alpha,k}}$ **then**
              add q-subgraph to selectedSubgraphs
            end
            add q-subgraph to checkedSubgraphs
          end

```
                end
            end
            add subgraph to checkedSubgraphs
        end
    end
    set previousSubgraphs to currentSubgraphs
end
```

Figure 2: Pseudocode for the nonhomogeneous subgraph discovery algorithm

In the case of "limited" graphs over a certain level of connectivity and q large enough, $\nu(g', q-s)$, the (q - s)-neighborhood of g, increases at the initial stage of the above exact process, while the number of tests being conducted may not reduce significantly considering the number of feasible tests. As a result, a faster, approximation algorithm is introduced. The main idea is to find subgraphs with sample mean shifts in the first k components of a new space, where the Euclidean norm $\left\|\hat{\hat{\delta}}_{[k]}(g)\right\| = \left\|U_{[k]}^T\left(\bar{x}_1(g) - \bar{x}_2(g)\right)\right\|$ is greater than a specified threshold. The output of substituting the *upperBound* with the following inequality in the Nonhomogeneous subgraph discovery algorithm produces an upper bound on $\tilde{T}_k^2(g)$.

$$
\begin{aligned}
\left\|U_{[k]}^T\left(\bar{x}_1(g) - \bar{x}_2(g)\right)\right\|^2 \le \left\|U^{\mathrm{T}}\left(\bar{x}_1(g) - \bar{x}_2(g)\right)\right\|^2 &= \|\bar{x}_1(g) - \bar{x}_2(g)\|^2 \\
&\le \|\bar{x}_1(g') - \bar{x}_2(g')\|^2 \\
&+ \max_{u_1,\dots,u_{q-s} \in \nu(g', q-s)} \Big\|\bar{x}_1\left(u_1, \dots, u_{q-s} \in \nu(g', q-s)\right) \\
&- \bar{x}_2\left(u_1, \dots, u_{q-s} \in \nu(g', q-s)\right)\Big\|^2
\end{aligned}
$$

This specifies a technique for detecting all subgraphs whose sample mean shift's Euclidean norm exceeds a certain threshold. Employing the T^2-test on these preselected subgraphs can also predict the group of subgraphs produced by the Nonhomogeneous subgraph discovery procedure.

A major issue with DE genes is the classification of non-significant differences as significant. The approaches presented by Lönnstedt and Speed (2002) can be used to solve such a problem.

Finally, the problem of multiple testing is raised due to the huge number of subgraphs assessed for homogeneity. This issue can be resolved by employing a permutation technique, which minimizes the amount of false positive subgraphs. Initially, the $n_1 + n_2$ observations' class/population labels are permuted, and then the nonhomogeneous subgraph discovery process is implemented to the permuted data to give a specific amount of false positive subgraphs. This technique is repeated several times to provide an approximation of the distribution of Type I error rates.

The procedures' performance is first assessed on synthetic data, and subsequently on breast cancer microarray data examined using KEGG pathways.

2.9. TEAK

Topology Enrichment Analysis frameworK (TEAK) was developed to discover active subpathways that underpin biological processes making use of the KEGG pathway database. Nodes represent gene products and/or complexes of gene products, while edges denote relationships between proteins or enzymes and are exploited to generate a set of unweighted adjacency matrices, which illustrate the KEGG pathways.

The subpathways extracted by the previous procedure can be either linear or nonlinear. Subpathways that consist of root to leaf linear paths are identified as linear. On the contrary, nonlinear subpathways are distinguished by feed-forward loops that are adjacent and overlap.

To evaluate the linear and nonlinear subpathways, TEAK initially fits a context specific Gaussian Bayesian network for each subpathway using the Bayes Net Toolbox. A Gaussian Bayesian network is a Bayesian network, that is a probabilistic graphical model, with all its nodes being linear Gaussians. Specifically, the Conditional Probability Distribution of Y for a continuous node Y with m continuous parents $X_1,..., X_m$ is:

$$p(Y|x_1, \dots, x_m) = N(\beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m;\ \sigma^2)$$

Subsequently, for subpathways consisting of context specific data, TEAK uses the Bayesian Information Criterion (BIC) provided in the Bayes Net Toolbox and scores each Bayesian network.

$$Score_{BIC} = \log P\left(D|\hat{\theta}\right) - 0.5d \log N.$$

Finally, since BIC is capable of breaking, which means that each node's score is computed separately and then added together to get the final score, each subpathway's value is normalized by the amount of nodes in order to make the scores equivalent.

2.10. PATHiWAYS

PATHiWAYS is another tool for pathway analysis. More specifically, it is a web server that can interpret the consequences of multiple changes in gene expression levels when it comes to signaling pathways.

It is based on a probabilistic model of the pathway, in which the probabilities of signal transmission are calculated. Gene expression values represent the gene activity and, therefore, the presence or absence of a protein. The 90^{th} percentile of the distribution of the probe activation probabilities is used in order to reduce the number of false positives caused by faulty probe observations.

After calculating the individual probability of each node, depending on the number of proteins they are composed of, a simple product of probabilities is computed to estimate the probability of signal transmission through the pathway.

Eventually, the final probabilities are compared and detect which stimulus–response circuits had their probabilities of signal transmission significantly altered.

2.11. DEAP rewrite: no p-value , the DEAP algorithm returns the scores

DEAP, which is short for Differential Expression Analysis for Pathways, is a pathway analysis method for identifying relevant regulatory patterns from differential expression data that takes advantage of information about biological pathways. Unlike previous methods of analysis, DEAP takes advantage of existing knowledge about pathway structure and recognizes the path that is the most differentially expressed. This technique calculates the scores of each subpathway with the use of the DEAP algorithm.

Initially, to estimate the null distribution of the test statistics and compute the *p*-values, a random rotation technique was employed. Rotation testing asserts that pathway and set data come from independent random samples of a multivariate normal distribution with mean zero under the null hypothesis.

Subsequently, the DEAP algorithm is applied. The algorithm handles expression data that are composed using the following multivariate normal distribution:

$$E = d(\mu + g) + e,$$

where d signifies if a gene is 'on' or 'off', μ represents the 'pathway effect' and g and e are generated using a normal distribution with both means equal to 0 and variances $\sigma^2{}_g$ and 1 respectively.

The algorithm is based on the following discrete steps. At first, the expression data are overlayed onto the network and each path from the graph is separately examined. Afterward, a recursive function is implemented and estimates the differential expression for each pathway considering the type of relationship between nodes by adding or subtracting all downstream nodes. The path with the maximal differential expression is detected by using and comparing the absolute value of the differential expression calculated for each pathway on the previous step.

Finally, the data were rotated *n* times to simulate a null distribution of the test statistic, s*, and the DEAP score was recalculated for every rotation sample. The random rotation approach helps resolve difficulties, such as not directly comparable DEAP scores for different paths, due to variances in size and structure among pathways, and determines the statistical significance.

Finally, the *p*-value is determined as the proportion of simulated DEAP scores, whose value is greater than or equal to the observed DEAP score, divided by the number of scores that are at least as extreme as the observed DEAP score:

$$p = \frac{\#(s_i \geq s*)}{n}$$

2.12. GraphiteWeb

Another option is GraphiteWeb, which is an innovative web tool for network analyses and visualization for gene expression data from both microarray and RNA-seq studies. It integrates topological and multivariate pathway studies with an efficient model of interactive network representations for simple comprehension of the results and uses a variety of multivariate gene set techniques. In addition, it uses multivariate gene set analysis including conventional hypergeometric enrichment, global test, GSEA, SPIA, and CliPPER, as well as the KEGG and Reactome pathway databases. In this paper, we will focus solely on the Enrichment Analysis (competitive and non-topological) approach.

Enrichment analysis uses the Fisher Exact test to estimate the odds of finding a certain number of genes in a specific pathway among the DEGs, denoted as $n_{G,deg}$. Within a set of N_{deg} genes, the likelihood P of seeing at least $n_{G,deg}$ genes is calculated by

$$P\left(N_{G,deg} \geq n_{G,deg}\right) = \sum_{i=n_{G,deg}}^{N_{deg}} \frac{\binom{N_G}{i}\binom{N-N_G}{N_{deg}-i}}{\binom{N}{N_{deg}}},$$

where N is the actual population of genes tested, G denotes the pathway, and N_i and n_i measure the frequency of genes within every table cell.

Subsequently, using the Benjamini and Hochberg technique, Ps are modified.

The statistical methodology typically employed to identify DEGs in RNA-seq count data is built on the negative binomial distribution. Given the strong relationship between read count and gene length, the read count specifies the test's power in this scenario. The P-value correction for gene length is an option in graphite web for adjusting for this bias.

2.13. PATHOME

PATHOME stands for Pathway and Transcriptome Information and is a computational approach for detecting differentially expressed subpathways. Its methodology is based on gene expression profiles of two control groups and relevant biochemical pathways.

At first, PATHOME divides the pathways into subpathways and then uses statistical tests to assess the significance of differential expression profiles alongside the pathway. The type of interaction is also considered.

The decomposition of the main pathway into linear paths is achieved using a depth-first search algorithm. Due to the huge number of possible paths resulting from the previous step, a selection step is used prior to the statistical significance test step to avoid such difficulties.

In order to select which segment of the subpathway will be statistically reviewed in the test step, the following rule is applied:

$$I^k = argmin_m \left\{ -\sum_{i=1}^{m} I\left(sgn\left(r_{i,i+1}^k \times e_{i,i+1}\right) = 1\right) + \sum_{i=1}^{m} R\left(sgn\left(r_{i,i+1}^k \times e_{i,i+1}\right)\right) \right\} + 1, m\epsilon\{1, \dots, p-1\}, R(x) = \begin{Bmatrix} 0, if\ x \in \{1\} \\ \infty, otherwise \end{Bmatrix}$$

A subpathway is chosen and continuous to the test step if the association rule between the expression correlation and the edge information for the neighboring items along the path is agreed upon by both experimental groups, and both consecutive segments include at least four components.

The final step determines which subpathway has a statistically significant difference in correlation between two subsequent segments for the two studies. The significance is examined under the null hypothesis, in which the alternative hypothesis represents the case in which the global mean of the correlations between the two groups are different.

Finally, to determine significance, we employed the z-test statistic, considered multiple comparisons, and the FDR was set at 0.05.

2.14. SubSPIA

The SPIA approach, mentioned above, was paired with a current subpathway analysis method to create the sub-SPIA method, which was used to find cancer-related pathways. To avoid problems resulting from the k-clique structure, used to define subpathways in the original subpathway analysis, the sub-SPIA method uses the minimal-spanning-tree structure.

A minimal-spanning tree is a tree-like subgraph, in which all nodes are connected, without forming a cycle. Because of the sparse connections between genes and the indirect connectivity of DEGs, this technique outperforms the k-clique notion.

The implementation of sub-SPIA was done using the R programming language. The following steps outline the main idea behind this method. Initially, we reassemble the gene network based on the signaling pathway. Then, in the gene network that has been created, the DEGs are mapped, and, finally, the subpathways are identified and their statistical and perturbation significance is evaluated. The

Kruskal algorithm is used to create the minimal-spanning tree and then remove any non-signature nodes remaining in the leaves of the MST.

To determine the statistical significance of each subpathway, the hypergeometric test and anomalous perturbation are employed. As in the SPIA tool, the present method contains two probabilities, P_{NDE} and P_{PERT}. The *p*-value can then be used to determine the pathway's enrichment significance using the following equation:

$$p = 1 - \sum_{x=0}^{r-1} \frac{\binom{t}{x}\binom{m-t}{n-x}}{\binom{m}{n}}$$

In the equation above, *m* represents the total number of genes in the genome, while t is the number of genes involved in the studied pathway. The variable t denotes the number of genes provided for study, of which r are participating in the same pathway. The P_{PERT} is calculated the same way as shown in the SPIA method and is combined with P_{NDE} in order to form a new global probability, PG.

2.15. MinePath

Another tool for pathway analysis is MinePath, which combines knowledge from gene expression profiles and molecular pathways. MinePath recognizes functionally differential sub-paths among different phenotype classes.

MinePath's general technique consists of five modular components that must be implemented. Initially, the gene expression data must be discretized so that the domain dimensionality is reduced. The supervised Entropy-based global discretization approach was implemented to convert gene expression values into binary equivalents that are either high (expressed) or low (non-expressed). In addition, because of the differences in the nomenclature of pathways and gene expression data, MinePath examines each gene's various probesets and indicates a combined expression value by employing a logic OR to the probeset values.

Then, following a depth-first search technique, each pathway is broken into all its sub-paths. The sub-paths that emerged are compared to the binary gene expression sample profiles provided as input.

Subsequently, the most discriminant sub-paths are identified using a multi-parametric sub-path selection technique, which is implemented by the employment of feature selection and classification techniques. MinePath includes three independent filters to analyze the phenotype differential power of sub-paths and identify the most discriminant among them: coverage, p-value, and polarity, each with its own customizable threshold. Those sub-paths that pass all the filters are chosen and maintained as the most discriminant.

Finally, MinePath evaluates the relevance of the pathways and ranks them according to their *p*-value, which is calculated based on the following formula:

$$p - value_p = \frac{\left(\frac{(a_p - 1) + b_p}{a_p - 1}\right)\left(\frac{c_p + b_p}{c_p}\right)}{\left(\frac{n}{(a_p - 1) + c}\right)}$$

MinePath's final output is a *p*-value ranked list of pathways from which the user can choose one to visualize and study.

2.16. HiPathia

HiPathia is a method that uses transcriptome data to calculate signal transduction along signaling pathways. To model the various cell functions in detail, each pathway is first broken down into circuits. The algorithm used by the HiPathia method models signal propagation by considering the level of activity of the proteins that make up the circuit. The simultaneous presence of the chain of proteins that connect the receptor to the effector, as well as the absence of inhibitor proteins that could compromise the signal's transduction along the circuit, in order to be active and thus transduce the signal to eventually trigger a function, is necessary for a circuit. The signal generated by the input node is communicated along the pathway as in the direction of the interactions and the output is collected by an output node, which activates a cell function. The signal is transmitted along the path according to the following recursive formula:

$$S_n = \upsilon_n \cdot \left(1 - \prod_{S_a \epsilon A} (1 - S_a)\right) \cdot \prod_{S_i \epsilon I} (1 - S_i)$$

S_n and υ_n represent the signal intensity for the current node *n* and its normalized gene expression value respectively. *A* describes all the activation signals (S_a) that are collected at the current node *n* from activation edges, and *I* describes the corresponding inhibitory signals (S_i).

Afterwards, a recursive technique based on the Dijkstra algorithm is used to determine the signal's propagation over the network. When the signal value across a node is updated in an iteration and the difference between the previous value and a threshold is exceeded, every node that the current updated node can lead to are also updated, until the updated values are less than the threshold.

Table 1. List of methods for Pathways Analysis

Method	**Date**	**Formula**
TAPPA	2007	$PCI = \sum_{i=1}^{N}\sum_{j=1}^{N} sgn(x_{is} + x_{js}) * \lvert x_{is}\rvert^{0.5} * \alpha_{ij} * \lvert x_{js}\rvert^{0.5}$
SPIA	2008	$P_G = c_i - c_i \cdot \ln(c_i)\,, c_i = P_{NDE}(i) \cdot P_{PERT}(i)$
TopologyGSA	2010	$\Lambda = \frac{L_{H_0}(\hat{K}_1, \hat{K}_2)}{L_{H_1}(\hat{K}_1, \hat{K}_2)} = \frac{L_{H_0}(\hat{K})}{L_{H_1}(\hat{K})}$
PARADIGM	2010	$IPA(i) = \begin{cases} L(i,1)\,, & L(i,1) > L(i,-1) and\ L(i,1) > L(i,0) \\ -L(i,-1)\,, & L(i,-1) > L(i,1)\ and\ L(i,-1) > L(i,0) \\ 0\,, & otherwise \end{cases}$
GGEA	2011	$\bar{S} \coloneqq \frac{S}{\lvert T_u\rvert}$
HotNet	2011	$FDR = E[V/R]$
PRS	2012	$PRS(p_i) = \sum_{j=1}^{n_i} NS_j$
DEGraph	2012	$T^2 = \frac{n_1 n_2}{n_1 + n_2}(\bar{x}_1 - \bar{x}_2)^{\mathrm{T}}\hat{\Sigma}^{-1}(\bar{x}_1 - \bar{x}_2)$
TEAK	2012	$Score_{BIC} = \log P(D\lvert\hat{\theta}) - 0.5d \log N$
PATHiWAYS	2013	Probabilistic model
DEAP	2013	$p = \frac{\#(s_i \geq s\,*)}{n}$
GraphiteWeb	2013	$P(N_{G,deg} \geq n_{G,deg}) = \sum_{i=n_{G,deg}}^{N_{deg}} \frac{\binom{N_G}{i}\binom{N-N_G}{N_{deg}-i}}{\binom{N}{N_{deg}}},$
PATHOME	2014	$z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$

SubSPIA	2015	$p = 1 - \sum_{x=0}^{r-1} \frac{\binom{t}{x}\binom{m-t}{n-x}}{\binom{m}{n}}$
MinePath	2015	$p - value_p = \frac{\left(\frac{(a_p - 1) + b_p}{a_p - 1}\right)\left(\frac{c_p + b_p}{c_p}\right)}{\left(\frac{n}{(a_p - 1) + c}\right)}$
HiPathia	2017	$S_n = v_n \cdot \left(1 - \prod_{S_a \epsilon A}(1 - S_a)\right) \cdot \prod_{S_i \epsilon I}(1 - S_i)$

3.

References

1. Miguel A. García-Campos, Jesús Espinal-Enríquez, Enrique Hernández-Lemus (2015). *Pathway Analysis: State of the Art*.
2. Tuan-Minh Nguyen, Adib Shafi, Tin Nguyen, Sorin Draghici (2019). *Identifying significantly impacted pathways: a comprehensive review and assessment*.
3. Xianbin Li, Liangzhong Shen, Xuequn Shang, Wenbin Liu (2015). *Subpathway Analysis based on Signaling Pathway Impact Analysis of Signaling Pathway*.
4. Lefteris Koumakis (2014). *Computational methods for knowledge discovery from heterogeneous data sources: methodology and implementation on biological and molecular sources*.
5. Patricia Sebastia´n-Leo´n, Jose´ Carbonell, Francisco Salavert, Rube´ n Sanchez, Ignacio Medina, Joaquı´n Dopazo (2013). *Inferring the functional effect of gene expression changes in signaling pathways*.
6. Ivana Ihnatova,Vlad Popovici,Eva Budinska (2018). *A critical comparison of topology-based pathway analysis methods*.
7. '*KEGG*' (2021) Wikipedia. Available at: https://en.wikipedia.org/wiki/KEGG.
8. '*KEGG PATHWAY Database*' (2021) GenomeNet. Available at: https://www.genome.jp/kegg/pathway.html.

9. ‘*BioCarta Pathways*’ SciScrunch. Available at: https://scicrunch.org/ADC/resolver/RRID:SCR_006917.
10. Shouguo Gao, Xujing Wang (2007). *TAPPA: topological analysis of pathway phenotype association*.
11. Adi Laurentiu Tarca, Sorin Draghici, Purvesh Khatri, Sonia S. Hassan, Pooja Mittal, Jung-sun Kim, Chong Jai Kim, Juan Pedro Kusanovic, Roberto Romero (2008). *A novel signaling pathway impact analysis*.
12. Maria Sofia Massa, Monica Chiogna, Chiara Romualdi (2010). *Gene set analysis exploiting the topology of a pathway*.
13. Charles J. Vaske, Stephen C. Benz, J. Zachary Sanborn, Dent Earl, Christopher Szeto, Jingchun Zhu, David Haussler, and Joshua M. Stuart (2010). *Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM*.
14. Ludwig Geistlinger, Gergely Csaba, Robert Küffner, Nicola Mulder and Ralf Zimmer (2011). *From sets to graphs: towards a realistic enrichment analysis of transcriptomic systems*.
15. ‘*HotNet*’ Raphael Lab. Available at: http://compbio.cs.brown.edu/projects/hotnet/.
16. Fabio Vandin, Eli Upfal, and Benjamin J. Raphael (2011). *Algorithms for Detecting Significantly Mutated Pathways in Cancer*.
17. Maysson Al-Haj Ibrahim, Sabah Jassim, Michael Anthony Cawthorne, and Kenneth Langlands (2012). *A Topology-Based Score for Pathway Enrichment*.
18. Laurent Jacob, Pierre Neuvial and Sandrine Dudoit (2012). *More Power via Graph-Structured Tests for Differential Expression of Gene Networks*.
19. Laurent Jacob, Pierre Neuvial, Sandrine Dudoit (2013). *DEGraph: differential expression testing for gene networks*.
20. Thair Judeh, Cole Johnson, Anuj Kumar, and Dongxiao Zhu (2012). *TEAK: Topology Enrichment Analysis frameworK for detecting activated biological subpathways*.
21. Patricia Sebastián-León, José Carbonell, Francisco Salavert, Rubén Sanchez, Ignacio Medina and Joaquín Dopazo (2013). *Inferring the functional effect of gene expression changes in signaling pathways*.
22. Winston A. Haynes, Roger Higdon, Larissa Stanberry, Dwayne Collins, Eugene Kolker (2013). *Differential Expression Analysis for Pathways*.

23. Gabriele Sales, Enrica Calura, Paolo Martini and Chiara Romualdi (2013). *Graphite Web: web tool for gene set analysis exploiting pathway topology*.
24. S Nam, HR Chang, K-T Kim, M-C Kook, D Hong, CH Kwon, HR Jung, HS Park, G Powis, H Liang, T Park and YH Kim (2014). *PATHOME: an algorithm for accurately detecting differentially expressed subpathways*.
25. '*Z Test: Definition & Two Proportion Z-Test*' Statistics How To. Available at: https://www.statisticshowto.com/probability-and-statistics/hypothesis-testing/z-test/.
26. Xianbin Li, Liangzhong Shen, Xuequn Shang, Wenbin Liu (2015). *Subpathway Analysis based on Signaling-Pathway Impact Analysis of Signaling Pathway*.
27. Lefteris Koumakis, Alexandros Kanterakis, Evgenia Kartsaki, Maria Chatzimina, Michalis Zervakis, Manolis Tsiknakis, Despoina Vassou, Dimitris Kafetzopoulos, Kostas Marias, Vassilis Moustakis, George Potamias (2016). *MinePath: Mining for Phenotype Differential Sub-paths in Molecular Pathways*.
28. Marta R. Hidalgo, Cankut Cubuk, Alicia Amadoz, Francisco Salavert, José Carbonell-Caballero, Joaquin Dopazo (2017). *High throughput estimation of functional cell activities reveals disease mechanisms and predicts relevant clinical outcomes*.
29. Kinza Rian, Marta R. Hidalgo, Cankut Çubuk, Matias M. Falco, Carlos Loucera, Marina Esteban-Medina, Inmaculada Alamo-Alvarez, María Peña-Chilet, Joaquín Dopazo (2021). *Genome-scale mechanistic modeling of signaling pathways made easy: A bioconductor/cytoscape/web server framework for the analysis of omic data*.