Journal of Genetic Engineering and Biotechnology 16 (2018) 447–457

HOSTED BY

Contents lists available at ScienceDirect

Journal of Genetic Engineering and Biotechnology

journal homepage: www.elsevier.com/locate/jgeb

PNME – A gene-gene parallel network module extraction method

Bikash Jaiswal, Kumar Utkarsh, D.K. Bhattacharyya *

Dept. of Computer Science and Engineering, Tezpur University, Napaam, Tezpur 784028, Assam, India

ARTICLE INFO

Article history:
Received 19 April 2018
Received in revised form 6 August 2018
Accepted 29 August 2018
Available online 11 December 2018

Keywords:
Coexpression network
Graphical processing unit
Module extraction
Generalized topological overlap measure

ABSTRACT

In the domain of gene-gene network analysis, construction of co-expression networks and extraction of network modules have opened up enormous possibilities for exploring the role of genes in biological processes. Through such analysis, one can extract interesting behaviour of genes and would help in the discovery of genes participating in a common biological process. However, such network analysis methods in sequential processing mode often have been found time-consuming even for a moderately sized dataset.

It is observed that most existing network construction techniques are capable of handling only positive correlations in gene-expression data whereas biologically-significant genes exhibit both positive and negative correlations. To address these problems, we propose a faster method for construction and analysis of gene-gene network and extraction of modules using a similarity measure which can identify both negatively and positively correlated co-expressed patterns. Our method utilizes General-purpose computing on graphics processing units (GPGPU) to provide fast, efficient and parallel extraction of biologically relevant network modules to support biomarker identification for breast cancer. The modules extracted are validated using p-value and q-value for both metastasis and non-metastasis stages of breast cancer. PNME has been found capable of identifying interesting biomarkers for this critical disease. We identified six genes with the interesting behaviours which have been found to cause breast cancer in homo-sapiens.

1. Introduction

Genetic analysis is a part of molecular biology specifically concerned with the understanding of the information which is encrypted in genes that are necessary for growth, reproduction and evolution of living organisms. Genes are some regions of the DNA (Deoxyribose Nucleic Acid) or RNA (Ribose Nucleic Acid) and act as a collection of biological information which is necessary to build and maintain an organism's cells [1]. Genetic analysis includes molecular technologies such as DNA sequencing, DNA microarrays, cytogenetic and Polymerase Chain Reaction (PCR).

Due to the proliferation of DNA microarray technology, it is now possible to generate gene expressions and analyse expression patterns of several genes in a systematic and comprehensive manner [2]. Microarrays help in detecting messenger RNA (mRNA) (which convey genetic information from DNA). Since it has tens of thousands of probes hence it can accomplish many genetic tests in parallel [3]. Using microarray technology, relative expression levels for genes are computed. The result of the computation forms the gene expression dataset [4,5].

We need to find gene-gene network modules which are a collection of genes that are functionally similar. Traditional machine learning [6] and statistical methods rely on disease-identification markers to support appropriate analysis of a disease. However, it has been shown that genes are usually involved in more than one function, and it is the interplay among genes that lead to diseases like cancer [7].

In order to find semantic similarities between a pair of genes from a dataset, a gene-gene network, referred here as the Gene Co-Expression Network (GCEN) has been constructed. It is a graph where each node represents a gene and an edge between two nodes represents either the interaction or some other relationship [8–11].

Next, a module extraction technique has been introduced, which enables to extract 'modules' from a GCEN. A module is a set of genes forming a dense region in the co-expression network. In other words, the modules are network components containing highly-correlated genes. Genes with high correlation correspond to some common biological phenomenon. By choosing the appropriate gene expression dataset and constructing efficient GCEN, module extraction can help to determine what genes are responsible for a biological process, like the progression a disease [12,13] or a common phenomenon like metabolism.

The major contributions of this paper are as follows.

Peer review under responsibility of Academy of Scientific Research & Technology.
* Corresponding author.
E-mail address: dkb@tezu.ernet.in (D.K. Bhattacharyya).

https://doi.org/10.1016/j.jgeb.2018.08.003
1687-157X/

- A fast method for construction and analysis of GCEN implemented on GPU.
- An effective network module extraction technique to support biomarker identification for a given disease. Here, we consider breast cancer, as an example disease.
- Six interesting genes have been identified w.r.t the ten causal genes of breast cancer for homo-sapiens.
- Some interesting topological associations have also been identified among these six biomarkers across the non-metastasis and metastasis stages.

The rest of the paper is organized as follows: Section 2 provides a discussion on some relevant literature. Section 3 reports the proposed method, i.e., PNME, whereas in Section 4.1 the implementation of PNME and some interesting results are reported. The concluding remarks are given in Section 4.2.1.

2. Related work

In literature, a number of techniques have been proposed for gene co-expression network construction. When inferring co-expression networks from gene expression data, a gene expression dataset is taken as primary input and then by using correlation-based proximity measures, the corresponding co-expression network is constructed. Frequently used correlation-based measures are Pearson Correlation Coefficient (PCC) and Spearman Correlation Coefficient (SCC). Approaches reported in [8] use PCC to extract the association among genes in a co-expression network.

The use of Generalized Topological Overlap Measure (GTOM) as a (dis) similarity measure has been reported in [14]. It has also been established that the original TOM mentioned in [15] discovers smaller modules, and higher order GTOM mentioned in [14] can help discovers larger modules. The implementation of GTOM in GPU makes use of the Strassen's Matrix Multiplication algorithm, implementation of which is described in [16]. The algorithm was implemented using the CUBLAS library of the CUDA platform, documentation of which can be found in [17].

Multiple ways of defining gene modules have been proposed in literature [10,9,14]. For detection of biologically meaningful modules, the generalized TOM measure (GTOM) [14] has been found very effective [18]. The study of the significance of module detection and extraction in revealing genes that are essential for biological phenomena has been explored in [19]. Several methods of choosing a correlation threshold to arrive at a network exist. A strategy based on statistical significance has been reported in [14,10] and also the authors highlight the limitations.

Implementation of GTOM requires massive matrix operations, especially when large datasets [20] are used. The use of GPU in such applications has been justified in [21]. Programming the GPU requires the use of the CUDA platform. The features, advantages, usage of CUDA have been discussed in [22] and in [23]. Further applications of CUDA and GPU, in general, are discussed in [23]. For parallel vector/matrix operations on the GPU, CUDA provides the CUBLAS library containing data structures, algorithms and functions for handling and processing large vectors and matrices. The implementation of Strassen Multiplication Algorithm in CUBLAS library has been analyzed and discussed in [24]. Finally, the documentation for CUDA [23,25] and CUBLAS [17] describes the data structures and operations in detail, including operations for transferring data to and from the GPU. Finally, the biological significance of module extraction and scope for further applications is reviewed in [15,10,18].

Although in the past decade a good number of network module extraction techniques have been evolved to support gene-gene network analysis, still it demands an effective technique which can fulfil the following requirements.

- ability to handle voluminous gene expression data instances,
- ability to handle gene expression data with high dimensionality,
- ability to extract module(s) of high biological significances, and
- ability to provide response quicker.

3. PNME: The proposed framework for network construction and module extraction

Biological Networks are useful for understanding gene-gene association and other functional properties of genes [26–28]. The commonly used biological networks include protein-protein interactions, signalling networks, metabolic networks, gene regulatory networks and GCEN. Gene co-expression networks have coverage of nearly all human genes, including cancer-related genes [29].

We define gene expression data as $G = \{G_1, G_2, \ldots, G_m\}$ be a set of m genes and $R = \{T_1, T_2, \ldots, T_n\}$ be the set of n conditions or time points of a gene expression dataset. The gene expression dataset X can be represented as a matrix of order $m \times n$ i.e., X_{mn} where each entry $X(i,j)$ in the matrix corresponds to the logarithm of the relative abundance of mRNA of a gene.

Co-expression network is an undirected graph where genes are nodes of the graph. Two nodes (genes) are connected if their activities have a significant association with a series of gene expression measurements. This association can be computed using a suitable correlation measure like Pearson, Kendall or Spearman. We use the Pearson Correlation Coefficient (PCC) to obtain the correlation as a matrix C. Each element C_{ij} of the matrix corresponds to the similarity between expression values of two genes G_i and G_j. PCC handles both negative and positive correlations, the value computed varies between +1 and −1. The Pearson Correlation between G_1 and G_2 over a Dataset D is represented by $r(G_i, G_j)$ and defined by Eq. (1)

$$r(G_i, G_j) = \frac{n\sum G_i G_j - \sum G_i \sum G_j}{\sqrt{n(\sum G_i^2) - (\sum G_i)^2}\sqrt{n(\sum G_j^2) - (\sum G_j)^2}} \quad (1)$$

For each similarity value between a pair of genes with a given threshold (support count), we compute 0–1 adjacency matrix as a representation of a graph. This graph represents the co-expression network.

Algorithm 1. Algorithm for constructing CEN

Input: Preprocessed Dataset $\mathbf{D}$, Threshold Θ
Result: Adjacency Matrix A representing the CEN
1 Initialization;
2 Generate correlogram matrix from $\mathbf{D}$;
3 **for** *each gene pair* (G_i, G_j) *from* D **do**
4 | Compute PCC for the expression values of gene pairs ;
5 | Compute $A(i,j)$ from the PCC value obtained and regulated via threshold Θ;
6 **end**
7 return $\mathbf{A}$;

Next, PNME extracts highly correlated network modules from G using GTOM. GTOM helps to find highest non-overlapping pairs of genes. It acts as a similarity measure which has been found useful in biological networks [14].

A network module C_i is defined as a set of genes forming a dense region in the co-expression network, with GTOM $\in$ *any top n GTOM values* corresponding to the extracted network modules [14]. GTOM is defined as following:

$$w_{ij} = \frac{l_{ij} + a_{ij}}{min(k_i, k_j) + 1 - a_{ij}} \tag{2}$$

where $l_{ij} = \sum_u a_{iu} a_{ij}$, and $k_i = \sum_u a_{iu}$ is the node connectivity. Basically, w_{ij} is an indicator for the agreement between the sets of neighboring nodes of i and j. The inclusion of the term a_{ij} in the numerator makes w_{ij} explicitly dependent on whether there is a direct link between the two nodes in question. The purpose of the quantity $1 - a_{ij}$ in the denominator is to avoid double-counting i as a neighbour of j and vice versa.

Algorithm 2. Algorithm for Module Extraction

Input: Adjacency Matrix A representing the CEN,Threshold($\mathbf{\Phi}$)
Result: Extracted Network Module
Initialization;
Normalise adjacency matrix $\mathbf{A}$;
for *m:1 to n* **do**
  Compute GTOMm for the Adjacency matrix A ;
  Extract the Module from the GTOMm Matrix obtained and regulated via threshold $\mathbf{\Theta}$;
end
return **Module**;

while the remainder of the code still runs on the CPU. CUDA is a parallel programming platform that allows programmers to use CUDA-enabled GPUs for general purpose processing [25]. CUDA exposes a software model and an API that gives direct access to the GPUs instruction set and computation elements for the execution of programs (compute kernels).

In this work, PCC measure has been implemented using GPGPU platform. Each thread of the GPU processes two rows of the data matrix at a time to compute arithmetic mean towards finding correlation between the pair based on PCC. Each thread accesses a single location in a 2D array to store the computed PCC value represented by the 2D indices of the array corresponding to the gene id's of the rows it is processing.

Before storing the value in the array, a threshold Θ is applied.

$$pcc(i,j) = \begin{cases} 0 & \text{if } r(G_i, G_j) \leqslant \Theta \\ 1 & \text{if } r(G_i, G_j) \geqslant \Theta \end{cases} \tag{3}$$

The final matrix of the order $m \times m$ (m being the No. of genes, or the No. of rows in the dataset) can now be treated as an adjacency matrix representing an undirected graph. Note that the following properties must hold for the graph to be undirected.

- $pcc(i,j) = pcc(j,i)$
- $pcc(i,i) = 0$

If two expressions are the same, they must be fully correlated (positively). However, for representation of the undirected graph, we set it to zero.

4. Implementation and results

A general purpose GPU (GPGPU) pipeline is a form of parallel processing between one or more GPUs and CPUs that analyzes data as if it were two-dimensional or three-dimensional, like an image or texture. GPUs have large cores that operate at lower frequencies, hence suitable for applications having small parallel units. Though, there is a limitation on parallel processing of application as it offloads compute-intensive portions of the application to the GPU,

Table 1
p-Values and q-values for tests on Metastasis for GTOM1.

Θ ↓	p-Value with Φ				q-Value with Φ			
	Φ = 0.4	Φ = 0.5	Φ = 0.6	Φ = 0.7	Φ = 0.4	Φ = 0.5	Φ = 0.6	Φ = 0.7
0.45	1.26E−5	1.04E−4	2.05E−4	2.3E−5	5.0E−2	8.3E−2	1.9E−2	2.2E−2
0.53	2.74E−4	1.73E−4	1.75E−4	2.67E−4	1.1E−1	1.51E−1	1.5E−1	1.6E−1
0.65	2.1E−4	2.9E−4	9.4E−4	2.14E−4	1.85E−1	1.4E−1	1.64E−1	1.18E−1
0.75	6.4E−4	2.5E−4	3.13E−4	7.2E−4	1.5E−1	1.20E−1	8.67E−2	8.6E−2

Table 2
p-Values and q-values for tests on Non-metastasis for GTOM1.

Θ ↓	p-value with Φ				q-value with Φ			
	Φ = 0.4	Φ = 0.5	Φ = 0.6	Φ = 0.7	Φ = 0.4	Φ = 0.5	Φ = 0.6	Φ = 0.7
0.45	8.4E−7	2.04E−5	4.3E−5	8.5E−4	8.5E−4	2.37E−2	7.07E−2	7.6E−2
0.53	1.75E−5	8.6E−5	1.31E−4	8.63E−5	2.3E−2	7.7E−2	6.7E−2	7.8E−2
0.65	9.4E−5	1.50E−5	1.7E−4	5.5E−4	1.0E−1	1.37E−2	5.6E−2	3.54E−2
0.75	1.67E−3	1.04E−3	3.7E−4	2.81E−4	1.74E−1	1.53E−1	1.2E−1	7.93E−2

Table 3
p-Values and q-values for tests on Metastasis for GTOM2.

Θ ↓	p-value with Φ				q-value with Φ			
	Φ = 0.4	Φ = 0.5	Φ = 0.6	Φ = 0.7	Φ = 0.4	Φ = 0.5	Φ = 0.6	Φ = 0.7
0.45	9.31E−8	2.48E−6	1.85E−5	2.85E−6	5.0E−5	4.1E−3	3.67E−2	4.52E−3
0.53	2.6E−5	4.6E−5	1.52E−5	5.44E−6	4.4E−2	4.8E−2	2.41E−2	7.1E−3
0.65	2.67E−6	2.72E−5	7.49E−5	3.33E−4	5.02E−3	2.4E−2	1.07E−1	1.38E−1
0.75	1.7E−4	1.55E−3	1.04E−3	3.91E−4	1.10E−1	1.07E−1	8.9E−2	7.63E−2

Table 4
p-Values and q-values for tests on Non-Metastasis for GTOM2.

Θ ↓	p-value with Φ				q-value with Φ			
	Φ = 0.4	Φ = 0.5	Φ = 0.6	Φ = 0.7	Φ = 0.4	Φ = 0.5	Φ = 0.6	Φ = 0.7
0.45	8.7E−14	1.8E−12	2.0E−8	3.9E−6	3.9E−10	3.2E−10	5.3E−5	8.75E−3
0.53	9.3E−9	2.2E−6	1.2E−4	9.9E−5	2.5E−5	5.2E−3	4.2E−2	4.8E−2
0.65	4.9E−5	3.7E−5	9.2E−6	1.6E−4	3.8E−2	5.5E−2	1.19E−2	6.39E−2
0.75	2.2E−3	1.66E−6	7.7E−3	1.82E−2	1.82E−1	6.4E−4	9.8E−2	1.4E−1

Table 5
p-Values and q-values for tests on Metastasis for GTOM3.

Θ ↓	P-value with Φ				Q-value with Φ			
	Φ = 0.4	Φ = 0.5	Φ = 0.6	Φ = 0.7	Φ = 0.4	Φ = 0.5	Φ = 0.6	Φ = 0.7
0.45	1.9E−7	1.1E−6	4.22E−6	5.9E−6	5.2E−4	2.46E−3	7.7E−3	8.58E−3
0.53	6.8E−5	1.6E−5	4.4E−6	9.8E−7	5.1E−2	2.8E−2	6.21E−3	1.08E−3
0.65	3.6E−5	37.4E−5	9.2E−5	1.54E−4	2.2E−2	1.06E−1	6.8E−2	9.8E−2
0.75	1.71E−3	1.7E−4	8.1E−4	3.9E−4	1.1E−1	1.1E−1	8.98E−2	7.63E−2

Table 6
p-Values and q-values for tests on Non-Metastasis for GTOM3.

Θ ↓	P-value with Φ				Q-value with Φ			
	Φ = 0.4	Φ = 0.5	Φ = 0.6	Φ = 0.7	Φ = 0.4	Φ = 0.5	Φ = 0.6	Φ = 0.7
0.45	8.5E−12	3.5E−8	2.2E−6	1.09E−5	2.6E−8	9.4E−5	2.84E−3	2.03E−2
0.53	1.2E−6	8.58E−6	7.8E−5	4.10E−5	3.02E−3	1.42E−2	3.8E−2	3.5E−2
0.65	9.8E−4	3.4E−3	3.4E−3	1.1E−4	1.1E−1	9.6E−2	6.7E−2	4.91E−2
0.75	7.54E−4	65E−3	1.09E−2	2.16E−2	1.1E−1	1.05E−1	9.4E−2	1.44E−1

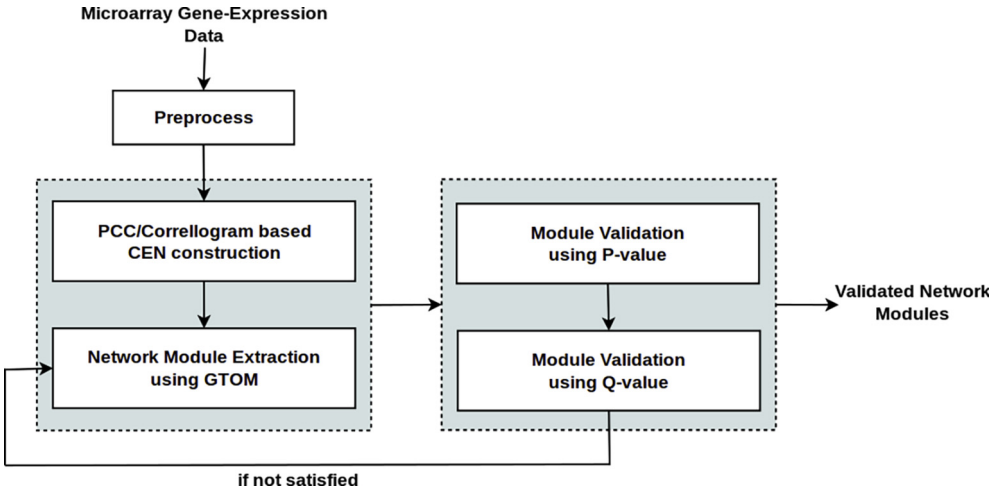


Fig. 1. Conceptual framework for parallel gene-gene network module extraction.

We cannot have a parallel implementation of GTOM from the definition given above. A parallel method of computing GTOM has been proposed by [14]. We use this method to implement GTOM algorithm in GPU.

The co-expression network C obtained in the previous step is stored as a matrix. In order to calculate GTOMm for a given value m, we need to multiply the matrices in the GPU to obtain $A^2, A^3, A^4, \ldots, A^m$. We compute each of the A^i for all $i = \{1, 2, \ldots, m\}$. Our algorithm computes A^i by multiplying A and A^{i-1}, thus requiring m calls to the cublasSgemm() function. The intermediate results are stored in an array of matrices which are added then normalized to obtain the matrix G.

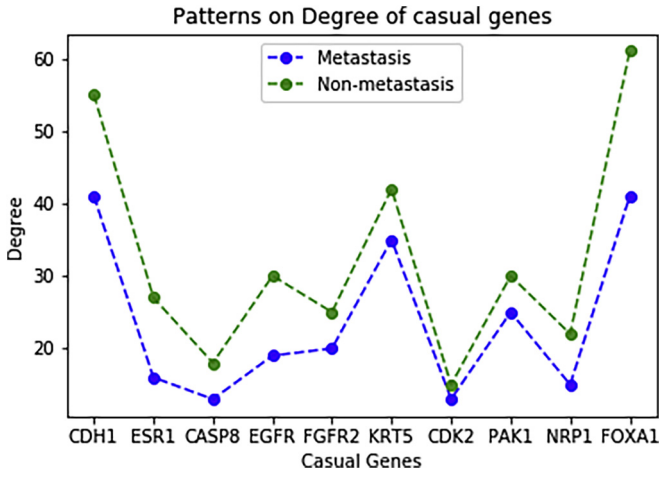


Fig. 2. Variation of degrees of each casual gene in each stage.

$$GTOMm(i,j) = \begin{cases} 0 & \text{if } G(i,j) \leqslant \Phi \\ 1 & \text{if } G(i,j) \geqslant \Phi \end{cases} \quad (4)$$

After normalization, we apply a threshold Φ to convert G into a network. Modules can be extracted by simply identifying the connected components in the network. These modules are sub-graphs where each node again corresponds to a gene. The connected components are extracted from the GTOMm network and stored in a file (as edge-lists) for analysis.

4.1. Dataset used

We performed our experimentation on *GSE2034* dataset, a Breast Cancer Relapse Free Survival Microarray Dataset [30,19,5]:

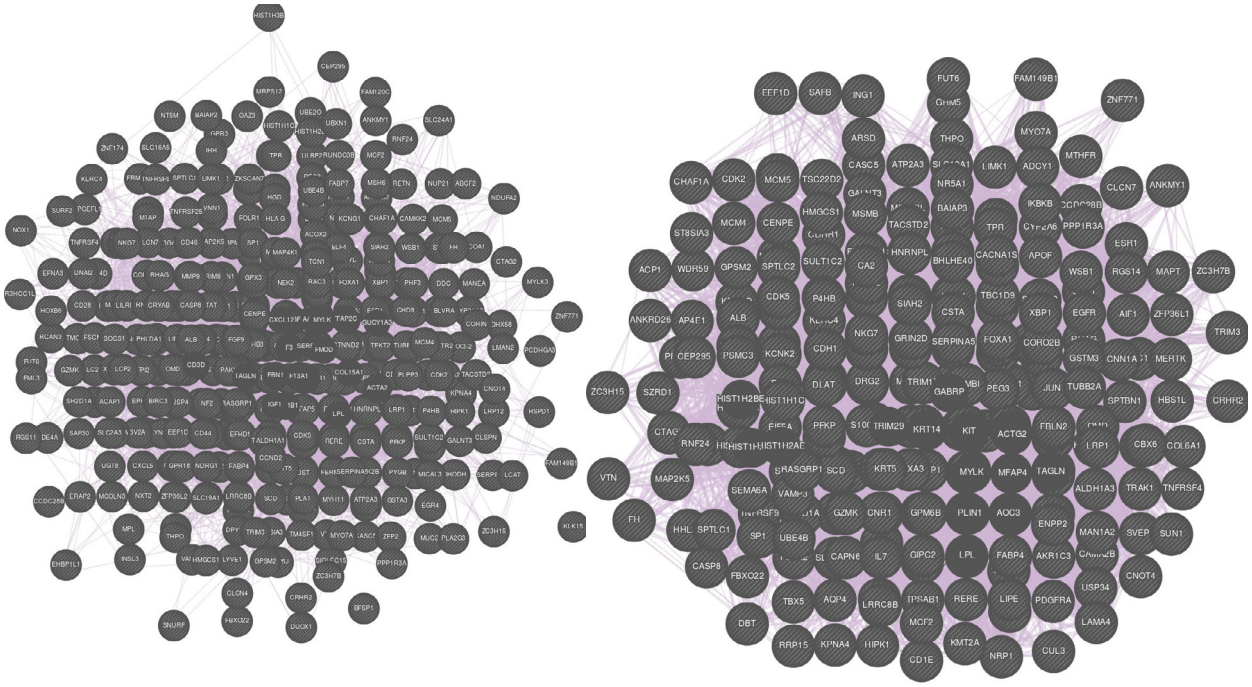
- **Title:** Breast Cancer Relapse Free Survival (GSE2034)
- **Organism:** Homo Sapiens
- **Experiment type:** Expression profiling by array

This dataset represents 180 lymph-node negative relapse-free patients (non-metastasis) and 106 lymph-node negative patients that developed a distant metastasis, for a total of 22,283 genes [30]. We use *log2* transformation in order to scale the values, and variance of 1.2 to select 5292 genes [7].

4.2. Results and observations

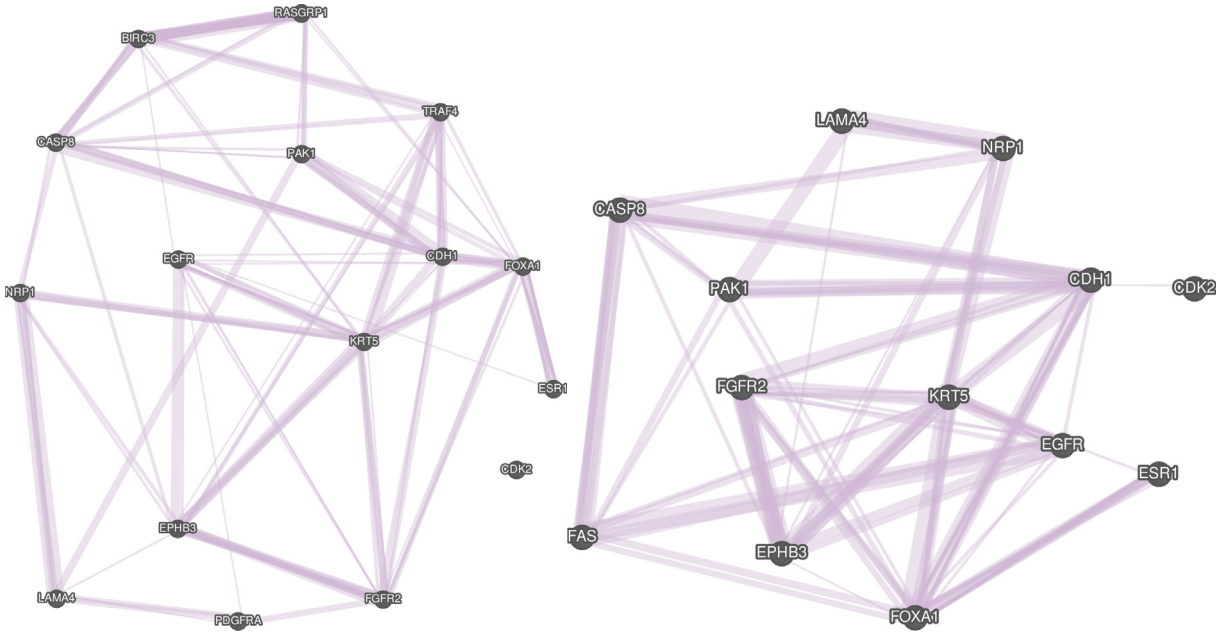
PNME has been implemented in Python 3.6 and C and tested on CentOS 6.5 (Linux Kernel 2.6.32-431) on a Dell T7910 with 2 Intel Xeon-Phi Co-processors, 64 GB RAM, 8 TB storage and 2 NVIDIA Tesla K40 GPUs.

In our experimentation, two thresholds have been used at two different points of analysis. One threshold is the PCC threshold θ used for computing correlation between two genes' expression series. The other threshold is the GTOMm threshold Φ used during



(a) Non-metastasis (b) Metastasis

Fig. 3. Figure (a): highly connected genes in Non-Metastasis Stage, figure (b): highly connected genes in Metastasis Stage.



(a) Non-metastasis (b) Metastasis

Fig. 4. Network between 10 primary and 6 secondary genes on (a) non-metastasis stage and (b) metastasis stage.

extraction of modules from the GTOMm graphs. Also, before computing the GTOMm matrix, the parameter m (m is number of iteration of GTOM) has to be set.

We derive Θ, Φ and m from the following sets. These values (set elements) have been decided through exhaustive experimentation.

- $\Theta = \{0.45, 0.53, 0.65, 0.75\}$
- $\Phi = \{0.4, 0.5, 0.6, 0.7, \}$
- $m = \{1, 2, 3, 4\}$

For given values of Θ, Φ and m, 128 network modules have been generated for analysis. To choose the best value among them we compute p-value and q-value of each network module. A p-value is a probability for a set of genes to be improved with the same functional group [7]. A q-value is an adjusted p-value for False Discovery Rate(FDR). The p-value for a module M enriched with functional group F is given as:

$$p\text{-}value = 1 - \sum_{i=0}^{q-1} \frac{\binom{|F|}{i}\binom{|V|-|F|}{|M|-i}}{\binom{|V|}{|F|}} \tag{5}$$

In order to choose the best values for parameters Φ, Θ and m, the p-value for each module is computed for the values of parameters from the sets defined above. The results are given in Tables 1–4. Our observation is that the best p-value is obtained for $\Theta = 0.45$, $\Phi = 0.4$ and $m = 3$ (GTOM3). The lowest p-value signifies modules that are most biologically-similar amongst themselves.

Tables 1, 3 and 5 reports p-values and q-values for metastasis corresponding to GTOM1, GTOM2 and GTOM3, respectively and Tables 2, 4 and 6 report p-values and q-values of non-metastasis corresponding to GTOM1, GTOM2 and GTOM3, respectively and with different parameters (see Figs. 1 and 2).

4.2.1. Observations

We obtain a correlated and highly connected co-expression network with $\Theta = 0.45$ and $\Phi = 0.4$ in both metastasis and non-metastasis stage as shown in Fig. 3, Tables 3 and 4. From the network, 180 genes have been found to participate both in metastasis and non-metastasis phase of cancer. Out of these 180 genes, we identify 10 causal genes which participate in breast cancer as per Malacard's [31] database. Using each of these 10 causal genes as primary genes, we identify some secondary set of genes, not present in Malacards, but are highly associated with the primary gene.

Next, each secondary gene is verified whether it follows the common genetic pathway with the primary genes or not. In order to do this, the online tool DAVID [32–34] has been used. The results are shown in Figs. 5, 6 and Tables 7, 8).

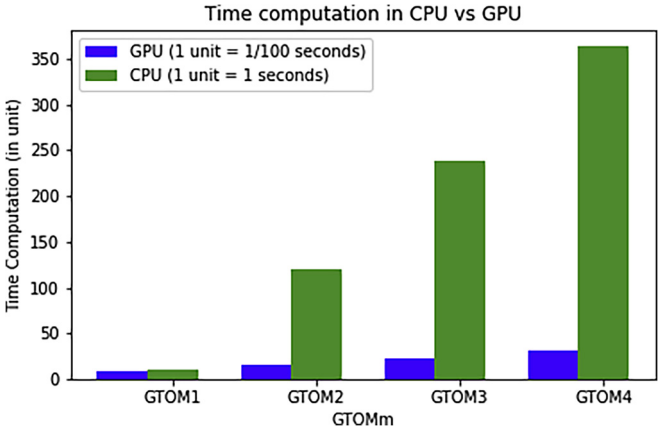


Fig. 5. Computation time for different *GTOM*m (m from 1 to 4) in GPU and CPU.

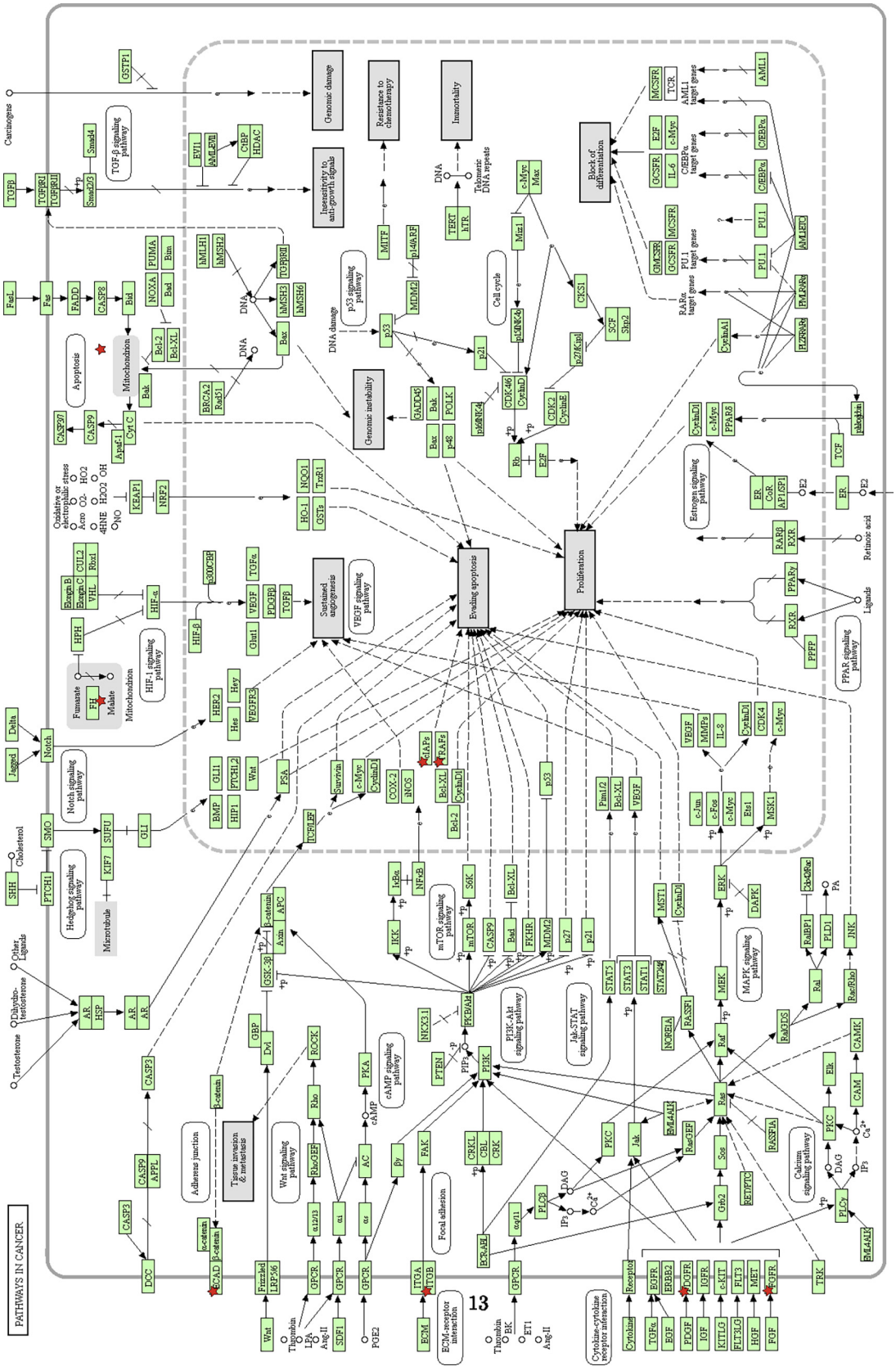


Fig. 6. Pathways in cancer of secondary genes in Non-metastasis. The above figure is generated using genemania [https://genemania.org/] [44].

Table 7
Secondary genes following same pathways as primary genes in Non-metastasis stage.

Non-Metastasis stage				
Primary Genes(P)	Degree(Dp)	Seconday Genes(S)	Degree(Ds)	KEGG Pathways
CDH1	55	TRAF4	18	Pathways in cancer
CDH1	55	FH	11	Pathways in cancer
PAK1	30	RASGRP1	52	T cell receptor signaling pathway
PAK1	30	LAMA4	39	Focal adhesion
FGFR2	25	PDGFRA	50	Pathways in cancer
NRP1	22	EPHB3	48	Axon guidance
CASP8	18	BIRC3	39	Pathways in cancer

Table 8
Secondary genes following same pathways as primary genes in metastasis stage.

Metastasis stage				
Primary Genes(P)	Degree(Dp)	Secondary Genes(S)	Degree(Ds)	KEGG Pathways
CDH1	41	FH	7	Pathways in cancer
PAK1	25	RASGRP1	35	T-cell Receptor Signalling Pathway
PAK1	25	LAMA4	25	Focal adhesion
EGFR	19	FAS	18	Pathways in cancer
NRP1	15	EPHB3	40	Axon guidance

4.2.2. Significance of secondary genes in Breast Cancer

In this section, six interesting genes(referred here as secondary genes) are reported which have been found to have close associations with those 10 causal genes(as available in Gene malacards [31]) both topologically as well as behaviorally. From the selected literature survey, it has been observed that all these six genes have significant roles in causing or driving breast cancer metastasis in homo sapiens (see Fig. 6, 7).

(a) TRAF4
Tumor Necrosis Factor Receptor-associated Factor 4 (TRAF4) plays an important role in tumorigenesis of breast cancer [35]. It drives breast cancer metastasis [36].
(b) LAMA4
Laminin Subunit Alpha 4 (LAMA4) expressions are high in breast cancer patients with worse relapse-free survival and low LAMA4 in patients with improved relapse-free survival. Also, malignant breast cancer cells express higher levels of LAMA4 relative to pre-malignant cells [37].
(c) FAS
Fas cell surface death receptor (FAS) shows significant associations with an increasing risk of breast cancer [38]. It has been observed that the risk of breast cancer may be elevated among women with polymorphisms in the FAS gene [39].
(d) PDGFRA
Platelet-Derived Growth Factor Receptor Alpha (PDGFRA) has been found as being uniquely expressed and active in inflammatory breast cancer (IBC) patient tumor cells and may be a promising target for therapy in IBC [40]. PDGFRA activation signature is also associated with small metastasis-free survival.
(e) BIRC3
BIRC3 is involved in chemo-resistance to doxorubicin in breast cancer cells [41]. As per the human protein atlas [42], BIRC3 is favourable for breast cancer.
(f) EPHB3
The largest family member of receptor tyrosine kinases, EPH receptors regulates cancer initiation and metastatic progression. Eph's expression is often elevated in breast cancer [43].

4.2.3. Association among primary and secondary genes

Another experimental topological study has been carried out to understand the associations among the primary and secondary genes across the two conditions (i.e., from non-metastasis to metastasis). It is aimed to understand the existence of the secondary genes (isolated or co-occurred) w.r.t. the primary genes across these two conditions. Fig. 4 and Table 9 report the topological behaviour. Some interesting observations are enumerated below.

(a) Two highly correlated secondary genes namely, TRAF4 and BIRC3 in the non-metastasis stage, have been found no association among other genes, during metastasis stage.
(b) FAS has been found to be highly associated with other genes in metastasis, while it was found missing in the non-metastasis stage.
(c) LAMA4 and PDGFRA are weakly associated with both the stage and their degrees of an association have become weaker in metastasis stage.
(d) EPHB3 shows the highest degree of association among all other secondary genes in the non-metastasis stage, have also been found to be associated with others strongly in metastasis stage.

4.2.4. Time computation

The algorithm for construction of the co-expression network has time complexity $O(n^2m)$, where n is the number of genes and m is the number of expression values for a single gene. However, for the parallel implementation, the computation time is $O(wm)$, where w is the *warp factor*. The warp factor is the ratio of the total number of blocks scheduled for the CUDA kernel to the warp size, where warp size is the number of threads that can run concurrently in the GPU. From Fig. 4, it can be stated that execution time doubles approximately as the value of m increases in GTOMm.

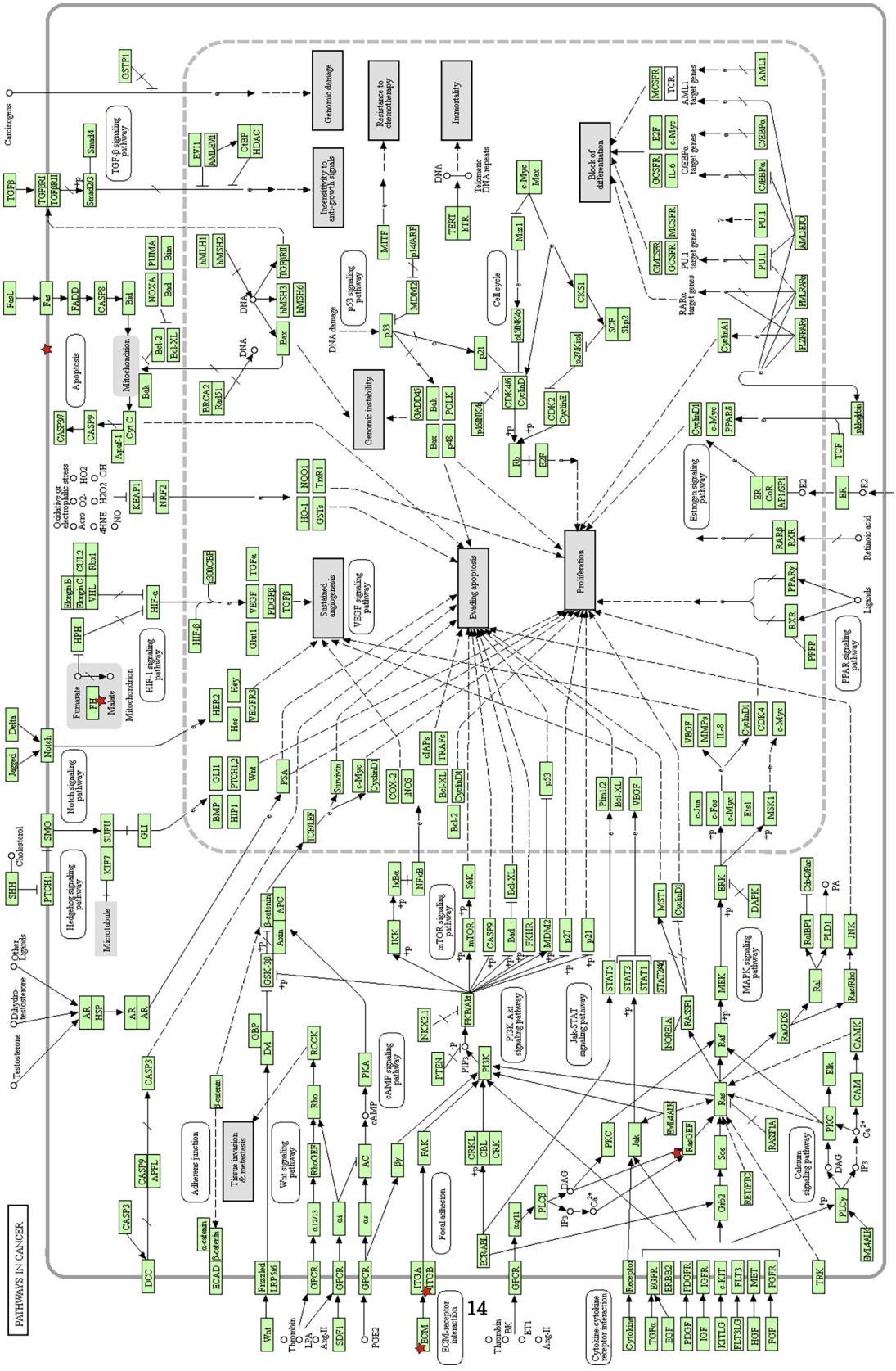


Fig. 7. Pathways in cancer of secondary genes in metastasis.The above figure is generated using genemania [https://genemania.org/] [44].

Table 9
Degree of each secondary genes on network made from 10 primary and 6 secondary genes.

Non-metastasis		Metastasis	
Gene names	Degree	Gene names	Degee
TRAF4	6	TRAF4	–
LAMA4	4	LAMA4	3
FAS	–	FAS	5
PDGFRA	1	PDGFRA	–
BIRC3	5	BIRC3	–
EPHB3	7	EPHB3	8

5. Conclusion

The proposed PNME enables a faster way of constructing and analyzing gene-gene coexpression networks and extraction of highly-correlated modules through the use of a GPU. The extracted modules corresponding to metastasis and non-metastasis stages of breast cancer are validated using P-value.

Also, using the results obtained, the secondary genes which follow a common genetic pathway with causal genes in breast cancer have been identified and reported.

Acknowledgement

This work was supported based on the funding received by Centre Of Excellence under FAST (MHRD, Govt. of India) (Dy. No. 394/14 IF.I dated 26.08.2014) and SAP DRS II of UGC (Dy. No. 28796 dated 15.04.2015).

References

[1] Scitable - Nature Education. Gene Expression. <https://www.nature.com/scitable/topicpage/gene-expression-14121669>.
[2] Gibson G. Microarray analysis. PLoS Biol 2003;1(1):e15. doi: https://doi.org/10.1371/journal.pbio.0000015.
[3] Wheelan SJ, Murillo FM, Boeke JD. The incredible shrinking world of dna microarrays. Mol Biosyst 2008;4(7):726–32. doi: https://doi.org/10.1039/b706237k. 18563246[pmid] <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2535915/>.
[4] Barrett T, Edgar R. Mining microarray data at ncbi's gene expression omnibus (geo). Meth Mol Biol 2006;338:175–90. doi: https://doi.org/10.1385/1-59745-097-9:175. 16888359[pmid] <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1619899/>.
[5] Roslin Institute. Microarray datasets - The macrophage community website; 2014. <http://www.macrophages.com/microarray-datasets>.
[6] Gogoi P, Borah B, Bhattacharyya DK, Kalita J. Outlier identification using symmetric neighborhoods. Proc Technol 2012;6:239–46.
[7] Sharma P, Bhattacharyya DK, Kalita J. Disease biomarker identification from gene network modules for metastasized breast cancer. Sci Rep 2017;7:1072. doi: https://doi.org/10.1038/s41598-017-00996-x. 996[PII] <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5430701/>.
[8] Ruan J, Dean AK, Zhang W. A general co-expression network-based approach to gene expression analysis: comparison and applications. BMC Syst Biol 2010;4(1):8.
[9] Mahanta P, Ahmed HA, Bhattacharyya DK, Ghosh A. Fumet: a fuzzy network module extraction technique for gene expression data. J Biosci 2014;39(3):351–64.
[10] Mahanta P. Analysis of gene co-expression and protein protein interaction data using unsupervised and semisupervised data mining techniques; 2016.
[11] Mahanta P, Ahmed HA, Bhattacharyya DK, Kalita J. Triclustering in gene expression data analysis: a selected survey. In: Emerging trends and applications in computer science (NCETACS), Shillong, IEEE; April–2011. doi: https://doi.org/10.1109/NCETACS.2011.5751409.
[12] Beer DG, Kardia SLR, Huang C-C, Giordano TJ, Levin AM, Misek DE, et al. Gene-expression profiles predict survival of patients with lung adenocarcinoma. Nat Med.
[13] van t Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AAM, Mao M. Gene expression profiling predicts clinical outcome of breast cancer. Nature 2002;415(6871):530–6. doi: https://doi.org/10.1038/415530a.
[14] Yip AM, Horvath S. Gene network interconnectedness and the generalized topological overlap measure. BMC Bioinform 2007;8(1):22.
[15] Barabási A-L, Ravasz E, Oltvai Z. Hierarchical organization of modularity in complex networks. Berlin, Heidelberg: Springer; 2003.
[16] Li J, Ranka S, Sahni S. Strassen's matrix multiplication on gpus. In: Proceedings of the 2011 IEEE 17th international conference on parallel and distributed systems, ICPADS '11; 2011. p. 157–64.
[17] NVIDIA. CUBLAS User Guide; 2017. <http://docs.nvidia.com/cuda/cublas/index.html>.
[18] Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabási A-L. Hierarchical organization of modularity in metabolic networks. Science 2002;297(5586):1551–5. doi: https://doi.org/10.1126/science.1073374.
[19] Klijn J, Berns E, Martens J, Jansen M, Atkins D, Foekens J, et al. Gene expression profiles and molecular classification to predict distant metastasis and tamoxifen-resistant breast cancer. Breast Cancer Res 2005;7(1):S2.
[20] Sarmah S, Das R, Bhattacharyya DK. A distributed algorithm for intrinsic cluster detection over large spatial data. Int J Comput Inform Eng 2012;2(9):246–56.
[21] NVIDIA. GPU accelerating computing in biosciences; 2010. <http://www.nvidia.com/object/bio_info_life_sciences.html>.
[22] NVIDIA. CUDA C programming guide; 2010. <http://docs.nvidia.com/cuda/cuda-c-programming-guide/>.
[23] NVIDIA. CUDA toolkit documentation; 2010. <http://docs.nvidia.com/cuda/>.
[24] Chrzeszczyk A, Chrzeszczyk J. Matrix computations on the GPU. CUBLAS and MAGMA by example; August 2013.
[25] NVIDIA. Parallel programming and computing platform—CUDA; 2010. <http://www.nvidia.com/object/cuda_home_new.html>.
[26] Furlong LI. Human diseases through the lens of network biology 29.
[27] Barabási A-L, Oltvai ZN. Network biology: understanding the cell's functional organization. Nat Rev Gene 2004;5:101. doi: https://doi.org/10.1038/nrg1272. Review article.
[28] Cai J, Borenstein E, Petrov D. Broker genes in human disease. Genom Biol Evol 2010;2:815–25.
[29] Zhao W, Langfelder P, Fuller T, Dong J, Li A, Hovarth S. Weighted gene coexpression network analysis: state of the art 2010; 20: 281–300.
[30] Wang Y, Klijn J, Zhang Y, Sieuwerts A, Look M, Yang F, et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet 2005;365(9460):671–9. doi: https://doi.org/10.1016/s0140-6736(05)70933-8.
[31] Weizmann Institute of Science. MalaCards: the human disease database. <http://www.malacards.org/>.
[32] Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucl Acids Res 2008;37(1):1–13. doi: https://doi.org/10.1093/nar/gkn923.
[33] Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Proto 2009;4(1):44–57. doi: https://doi.org/10.1038/nprot.2008.211.
[34] Laboratory of Human Retrovirology and Immunoinformatics (LHRI). Leidos Biomedical Research, Inc. David bioinformatics resources 6.8, niaid/nih. <https://david.ncifcrf.gov/tools.jsp>.
[35] Zhang J, Li X, Yang W, Jiang X, Li N. TRAF4 promotes tumorigenesis of breast cancer through activation of akt. Oncol Rep 2014;32(3):1312–8. doi: https://doi.org/10.3892/or.2014.3304.
[36] Zhang L, Zhou F, de Vinuesa AG, de Kruijf EM, Mesker WE, Hui L, et al. TRAF4 promotes TGF-β receptor signaling and drives breast cancer metastasis. Mol Cell 2013;51(5):559–72. doi: https://doi.org/10.1016/j.molcel.2013.07.014.
[37] Ross JB, Huh D, Noble LB, Tavazoie SF. Identification of molecular determinants of primary and metastatic tumour re-initiation in breast cancer. Nat Cell Biol 2015;17(5):651–64. doi: https://doi.org/10.1038/ncb3148.
[38] Hashemi M, Fazaeli A, Ghavami S, Eskandari-Nasab E, Arbabi F, Mashhadi MA, et al. Functional polymorphisms of FAS and FASL gene and risk of breast cancer – pilot study of 134 cases. PLoS ONE 2013;8(1):e53075. doi: https://doi.org/10.1371/journal.pone.0053075.
[39] Crew KD, Gammon MD, Terry MB, Zhang FF, Agrawal M, Eng SM, et al. Genetic polymorphisms in the apoptosis-associated genes FAS and FASL and breast cancer risk. Carcinogenesis 2007;28(12):2548–51. doi: https://doi.org/10.1093/carcin/bgm211.
[40] Joglekar-Javadekar M, Laere SV, Bourne M, Moalwi M, Finetti P, Vermeulen PB, et al. Characterization and targeting of platelet-derived growth factor receptor alpha (PDGFRA) in inflammatory breast cancer (IBC). Neoplasia 2017;19(7):564–73. doi: https://doi.org/10.1016/j.neo.2017.03.002.
[41] Mendoza-Rodríguez M, Romero HA, Fuentes-Pananá EM, Ayala-Sumuano J-T, Meza I. IL-1β induces up-regulation of BIRC3, a gene involved in chemoresistance to doxorubicin in breast cancer cells. Cancer Lett 2017;390:39–44. doi: https://doi.org/10.1016/j.canlet.2017.01.005.
[42] The Human Protein Atlas. Expression of birc3 in cancer; 2017. <https://www.proteinatlas.org/ENSG00000023445-BIRC3/pathology>.
[43] Vaught D, Brantley-Sieders DM, Chen J. Eph receptors in breast cancer: roles in tumor promotion and tumor suppression. Breast Cancer Res 10(6). doi: https://doi.org/10.1186/bcr2207. <https://doi.org/10.1186/bcr2207>.
[44] Donaldson DW-FSL, Comes O, Zuberi K, Badrawi R, Chao P, Franz M, et al. The genemania prediction server: biological network integration for gene prioritization and predicting gene function. Nucl Acids Res 2010;38:214–20.

Bikash Jaiswal received his B.Tech in Computer Science and Engineering, from Department of Computer Science and Engineering, Tezpur University in 2017. Currently, he is working as Project Fellow in the project entitled "Machine Learning Research and Big Data Analytics (MLRBDA)", in the same university.

Kumar Utkarsh received his B.Tech in Computer Science and Engineering, from Department of Computer Science and Engineering, Tezpur University in 2017. Currently, he is pursuing M.Tech degree at Department of Computer Science and Engineering, NIT Rourkela.

D.K. Bhattacharyya received his Ph.D. in Computer Science from Tezpur University in 1999. He is a Professor in the Computer Science & Engineering Department at Tezpur University. His research areas include Data Mining, Network Security and Content-based Image Retrieval. Prof. Bhattacharyya has published 280 + research papers in the leading international journals and conference proceedings. In addition, Dr Bhattacharyya has written/edited 11 books. He is a Programme Committee/Advisory Body member of several international conferences/workshops.