Biomarker Identification For Critical Diseases Using Machine Learning Techniques

A thesis submitted in partial fulfillment of the requirement for the degree of

Doctor of Philosophy

Manaswita Saikia

Registration No. TZ155525 of 2015

Department Of Computer Science and Engineering

School of Engineering

Tezpur University
Napaam,Tezpur, Assam-784028

August, 2023

Chapter 7

Conclusion

7.1 Concluding Remarks

In this dissertation we have presented the results of our attempt to develop schemes and frameworks for identifying genetic biomarkers for critical diseases. The following provides a summary of our contributions in the research presented in this dissertation as well as the concluding remarks.

- For microarray data, pre-processing and subsequent analysis required for BicGenesis, Integrative DEA , and CBDCEM such as statistical tests, bicluster generation, identification of DEGs, and CEN construction can be implemented with ease and with limited additional processing This may be in part due to the fact that most methods / approaches take microarray data into consideration when developed.
- The bulk RNA-Seq data requires relatively more preparation and pre-processing so as to enable meaningful subsequent analysis. For many of these methods, the data should be prepared in a certain manner for good results and such methods are easier to implement on microarray data when compared to bulk RNA-Seq data. In terms of computational intensity, all subsequent methods employed by our frameworks are more or less the same. scRNA-Seq requires extensive pre-processing and data preparation so as to facilitate most of the methods we have employed across all frameworks, such as creation of CENs, and identification of DEGs. Creation of condition-type objects and cell-type objects in scDiffCoAM facilitates the implementation of these methods.
- Implementation of the eight chosen bicluster generation algorithms in BicGenesis requires extensive experiments to make a choice of the parameters for optimal results. For most algorithms, each iteration with the same parameters leads to varying results. The only exception to this is the BiMax [556] algorithm which creates the same biclusters regardless of the number of iterations. Implementation of the FLOC [809] algorithm requires the most number of iterations for the choice of parameter

set to obtain the desired outcome. Furthermore, even with implementation of multi-threading, the FLOC algorithm is observed to be extremely computationally intensive when compared to the other seven algorithms.

- In the Integrative DEA framework, observing multiple iterations of execution, we observe that consideration of only the genes common to all three methods in the proposed consensus function, namely Limma [637, 638], SAM [697], and EBAM [148] in microarray, and limma+voom [332], edgeR [584], and DESeq2 [449, 450] in bulk RNA-Seq, leads to information loss. To overcome the information loss, *lFDR* is found effective for microarray data analysis and *q-value* for bulk RNA-Seq data analysis in the consensus function.
- The Integrative DEA framework is found useful in observing the changes in expression of genes individually under two varying conditions (normal and disease). As such, we also take into consideration the biological significance of each relevant DEG detected by our framework and introduce the concept of Top Enriched DEG (TED).
- Our hub-gene finding method employed by CBDCEM outperforms the other competing methods in most cases and similarly in certain cases. However, in very sparse modules the WGCNA [327] intra-modular connectivity (IMC) outperforms our hub-gene finding method.
- The observation that IMC performs better than our hub-gene finding method in sparser modules as our method does not take into consideration the connectivity among the genes is also highlighted by the incorporation of the method into scDiffCoAM. As scRNA-Seq data have large numbers of missing values compared to the construction of CENs on the cell-type object, we detect blocks of genes, and consider them as modules. This, in turn, leads to the construction of very well-connected dense modules. In case of scDiffCoAM, however, we replace four measures that are ineffective in modules of this nature with more network appropriate measures. Thus, with these changes and detection of denser highly connected modules, our methods outperforms IMC in most scenarios.

The following gives a summary of the genes identified and validated as potential biomarkers for ESCC by all four frameworks.

- For each proposed framework, we identify a set of genes that are candidates for potential biomarkers . We call these genes as Biomarker Candidate Genes (BCGs). We biologically analyze these BCGs through a) GO enrichment analysis, b) KEGG path-

way enrichment analysis, c) gene regulatory network (GRN) analysis , and d) tracing literature evidence that associate the BCG with the disease of primary interest and other associated diseases.

- We consider Esophageal Squamous Cell Carcinoma (ESCC) as the disease of interest and also take into consideration five other SCCs namely, Head and Neck SCC (HNSCC), Laryngeal SCC (LaSCC), Lung SCC (LSCC), Oral SCC (OSCC), and Tongue SCC (TSCC), each of which is associated with ESCC. To validate a BCG identified as a potential biomaker for ESCC, we employ the biomarker criteria as discussed in Section 2.5. All BCGs that fall under Case 1 and Case 2 of the biomarker criteria are identified as potential biomarkers for ESCC.
- Four TFs detected by BicGenesis, E2F1, HMGA2, PCBP1, and PIN1, fall under Case 1 (Table 3.12) of the biomarker criteria and are potential biomarkers for ESCC. Similarly, eighteen BCGs, namely, AGK, BNIP3, COL1A1, CTTN, FGF5, IPO5, NOX4, CAV1, TGFA, FN1, SEMA3F, CA12, SOCS5, DGKA, PPARGC1A, SEMA3B, UBE2C, and KPNA2 fall under Case 2 of the biomarker criteria, and thus are potential biomarkers for ESCC.
- The integrative DEA framework identified two BCGs that are TFs, DNMT3B, and MCM7, as potential biomarkers for ESCC as they fall under Case 1 (Table 4.12). Additionally, two TEDs identified by our framework, STAT1, and HIF1A, are also TFs and fall under Case 1, making them potential biomarkers for ESCC. Twenty genes, namely, HOMER3, PSMD4, PSAT1, TFRC, MCL1, EPHA2, KPNA2, CKS2, PRMT1, HLA-F, HLA-G, CXCL10, ISG15, PSMA3, FCGR2A, C3AR1, CAV1, VEGFC, CDK4, and MSH2, are identified by the Integrative DEA as potential biomarkers for ESCC as they fall under Case 2 of the biomarker criteria. Similarly, GSK3B is identified by the framework as a TED and falls under Case 2 and thus is a potential biomarker of ESCC.
- Three transcription factors (TFs), HSF1, MCM7 and PML, detected by CBDCEM, fall under Case 1 (Table 5.12) of the biomarker criteria discussed in Section 2.5 and are potential biomarkers for ESCC. Similarly, CBDCEM identifies twelve genes, namely, DGKA, MAP4, PFN2, DUSP6, ACVR1B, PRDX6, MAPK9, SEL1L, EHD2, KIFC1, STMN1, BIRC5, as potential biomarkers for ESCC as they fall under Case 2 of the biomarker criteria.
- Four TFs, MCM7, GATA3, NOTCH3, and ETV5, identified by scDiffCoAM fall un-

der case 1 (Table 6.11) of the biomarker criteria deascribed in Section 2.5 and are potential biomarkers for ESCC. Twenty BCGs, namely, CCNB2, CDH1, LAMP3, NLRP3, TUBB3, CAV1, UCHL1, G6PD, FGF5, EPHA2, EGFR, DSC2, IDO1, ITGA3, SEMA3B, MMP7, SGK1, CXCL14, CLDN4, and CDKN2B, fall under case 2 and are identified by scDiffCoAM as potential biomarkers for ESCC.

- Caveolin-1 (CAV1) [26, 301, 283] is identified as a potential biomarker for ESCC by Integrative DEA in GSE130078, by BicGenesis in GSE23400 , and by scDiffCoAM in Myeloid cell type of GSE160269.
- The transcription factor, mini-chromosome maintenance complex component 7 (MCM7) [563, 906, 10], is identified as a potential biomarker for ESCC by three of our proposed frameworks. MCM7 was identified as a potential biomarker by the integrative DEA framework in GSE20347, by CBDCEM in GSE23400, and by scDiffCoAM in Tcell cell type of GSE160269.
- Both fibroblast growth factor 5 (FGF5) [273], and semaphorin 3B (SEMA3B) [673, 142] are identified by scDiffCoAM in Epithelial and Endothelial cell types of GSE160269, respectively as potential biomarkers for ESCC . FGF5, and SEMA3B are also identified by BicGenesis as potential ESCC biomarkers in GSE20347, and GSE23400, respectively.
- Both EPH receptor A2 (EPHA2) [504, 660] and Karyopherin alpha 2 (KPNA2) [475, 596] are detected as potential biomarkers for ESCC by our Integrative DEA framework in GSE20347. EPHA2 is also detected as a potential biomarker for ESCC in Endothelial cell type of GSE160269 while KPNA2 is detected by BicGenesis in GSE23400.
- BicGenesis and CBDCEM detected Diacylglycerol kinase alpha (DGKA) [76] as potential biomarkers for ESCC in GSE23400 and GSE20347, respectively.

7.2 Future Works

The following are some of the possible directions of that can be foreseen in the area.

- In the integrative DEA approach as mentioned earlier taking into account only DEGs that are common to the three corresponding methods leads to information loss. As such we introduced *q-value* and *lFDR* into the consensus function. We observe that although the consensus function does lead to lower information loss, the number of

DEGs detected also increases drastically. So, incorporation of another measure or a combination of the *q-value* and *lFDR* with other measures into the consensus function may be a possible improvement leading to detection of optimal number of DEGs with minimal information loss.

- In CBDCEM, while our method outperforms IMC, which takes into account the intra-modular connectivity of the network, in relatively lager and denser modules, IMC performs better than our method in sparser modules. This may be due the fact that IMC measures how well-connected a gene is to other genes in the module, and as such in sparser modules, it performs better than our method. Our hub-gene finding method can be explored for further improvement by the incorporation of a method that takes into account connectivity among genes.
- It is possible to explore the development of an ensemble-based centrality approach which can ensure to generate unbiased results.
- Deep Learning methods can be explored for selection of an appropriate subset of genes prior to downstream processing.
- For effective DEG identification, control samples can be pre-processed by developing an appropriate measure towards identification of a subset of varied, non-redundant and less correlated control samples.
- In scDiffCoAM, it is possible to explore the development of a substitute algorithm for PCA that takes biological parameters into consideration rather than statistical during dimensionality reduction.
- There are scopes for strengthening the ranking scheme using a multi-objective approach.