Big data omics for precise and personalized medicine

Hossein Sharifi-Noghabi
Database and Data Mining Laboratory, School of Computing Science, Simon Fraser University,
Burnaby, Canada.
Laboratory for Advanced Genome Analysis, Vancouver Prostate Centre,
Vancouver, Canada.
Senior supervisor: Prof. Martin Ester.
Secondary supervisor: Prof. Colin Collins.
Email: hsharifi@{sfu.ca/prostatecentre.com}.

Qingyuan Feng
Database and Data Mining Laboratory, School of Computing Science, Simon Fraser University,
Burnaby, Canada.
In-silico Drug Design Laboratory, Vancouver Prostate Centre,
Vancouver, Canada.
Senior supervisor: Prof. Martin Ester.
Secondary supervisor: Dr. Artem Cherkasov.
Email: qfa4@sfu.ca

Abstract

In 2015, US President Barack Obama announced the Precision Medicine Initiative and since then, this term has been enjoying rising popularity in the scientific community. The ultimate goal of Precision Medicine is to provide each group of similar patients with precise and personalized treatment based on their genetic landscapes and environmental factors. Therefore, it is of utmost importance and significance to investigate computational means for this goal in deeper levels. In this review, we study relevant papers published in the top computational biology and bioinformatics conferences, including RECOMB, ISMB and PSB. This review tries to cover most of current edges of research in the field, in order to pave the ground for readers to acquire a bigger picture of Precision Medicine from a computational perspective.

I. Introduction

Advancement and development of diverse omics technologies ushered in the new paradigm called Precision Medicine, which is often being used interchangeably with Personalized Medicine. According to the National Institutes of Health (NIH), the definition of the term Precision Medicine is "an emerging approach for disease treatment and prevention that takes into account individual variability in genes, environment, and lifestyle for each person." [1].

There is a lot of overlap between the terms Precision Medicine and Personalized Medicine. According to the National Research Council, the latter is an older term with a meaning similar to the former. However, usually, there was a misinterpretation of the word "personalized" because people thought it implies that treatments and preventions are being proposed for each individual uniquely. In Precision Medicine, the focus is on identifying which approaches will be more determinant and effective for which group of patients based on genetic and environmental (e.g. lifestyle) factors. The Council, therefore, preferred the term Precision Medicine to Personalized Medicine [2]. However, they are still used by some people (including the authors of this review) interchangeably. In this review, six highly important aspects to realize Precision Medicine are discussed. These covered areas are 1) biological networks, 2) deep learning, 3) patient stratification, 4) clonality inference, 5) drug design/discovery and finally 6) health records.

In biological networks, group of genes or proteins interact with each other to accomplish a certain goal or act as a functional group for a specific task. A group of genes interacting with each other to perform a specific task in the cell is called a pathway. Pathways are extremely important for studying diseases with genetic cause, for instance, in the case of cancer, it is more effective to study driver genes in their pathways rather than a single target gene because of heterogeneous nature of mutations in cancer [3, 4]. For proteins, the idea is quite similar, because group of proteins need to interact with each other in order to perform complex molecular functions [5].

On the quest for accurate and precise medicine, one needs appropriate computational tools and currently, deep neural networks are among the most effective ones. Deep learning has the capability to learn different representations and extract salient features from them. For example, the input space can be raw sequence data, bunch of images, a signal or something as complicated as SMILES [6] representation of chemical compounds. More interested readers can refer to [7] for more information regarding deep neural networks.

Once groups of signature genes or other biomarkers such as proteins are known, they can be used to stratify patients. Patient stratification is an approach by which groups of similar patients are subdivided into different categories based on different criteria such as the underlying mechanism of the disease, their probable response to a therapy or their genetic profiles, e.g., somatic mutations. Stratifying patients into biologically meaningful subgroups with diverse clinical conditions can be helpful in finding new ways for the development of more effective personalized treatment strategies [8, 9].

So far, we realized that somatic mutations can be both challenging due to their heterogeneity and at the same time, informative for stratifying patients. Somatic mutations can also be helpful to determine how pure a tumor sample is and find its sub-clonal compositions. In clonality inference, we are interested in knowing which mutations happen first in the evolutionary process of the disease, especially cancer, and specify subclones for each family of mutations [10].

Computational drug design and discovery is another crucial step in Precision Medicine. Determining interactions between a new drug and a target experimentally can be extremely expensive and time consuming, therefore, reliable computational techniques are required to evaluate interactions between targets and newly proposed drugs [11].

Genome-wide Association Studies (GWAS) are tremendously contributing to pinpoint rare and common variations in human genome for different diseases and provide all of the stated aspects of Computational Precision Medicine with more proper inputs. Linked Electronic Health Records (EHR) with biorepositories can also provide us with a rapid and cost-effective data collection for genetic association studies. EHR contains huge amount of patients' data, and since it is linked to biorepositories, it can be utilized in genetic studies [12].

In this review, different computational methods and examples of the stated areas are discussed in detail. The rest of the paper is as follows: section II presents a comprehensive review of the 10 selected papers about the aforementioned areas and section III concludes the papers.

II. Related works

In this section, the selected papers are discussed. For each of them, motivation, summary of the method, pros and cons and possible extensions are stated.

1- Biological networks: genetic pathways and Protein-Protein Interactions

A. Finding Mutated Subnetworks Associated with Survival in Cancer [3]

Motivation and contributions: as stated above, mutations in cancer are known to be highly heterogeneous. Therefore, it is important to deal with this diversity in order to determine association between the mutations and clinical parameters such as survival time. This research focused on identification of subnetworks of a gene-gene interaction network that have mutations associated with survival time. This research has three major contributions: first, the authors formulated the problem of finding a subset of k genes with mutations which have highest association with survival time by using the log-rank statistic, and proved it is an NP-hard problem; second, for the aforementioned objective, the authors proposed NoMAS which is a color-coding-based algorithm that takes advantage of log-rank test as a metric to evaluate subnetworks (level of association between mutations in a group of genes and survival); third, they tested NoMAS on simulated and real data.

Summary: In this paper, absolute value of normalized log-rank statistic was used for the stated objective and method. This measure is based on non-censored/censored survival information. In order to find a connected set of k genes in a gene-gene interaction network with maximum association with survival time, one should look for a subnetwork with maximum score (based on the aforementioned metric), and this problem is called the max connected k-set log-rank problem. The authors proved that the problem (both itself and its relaxed version without gene-gene interaction network) is NP-hard. The proposed iterative algorithm has the goal of finding a colorful subnetwork (all vertices have distinct colors) of the colored graph with the highest score. In each iteration, the subnetworks will be merged using dynamic programming, and the highest-scoring one will be reported. After identifying the best solution, the authors used p-value for the log rank statistic and permutation test to assess its statistical significance. Finally, the Planted Subnetwork Model was introduced, which could fit the situation where mutations are not placed adversarially and the optimal solution does not have many neighbors. The proposed method can perform well under Planted Subnetwork Model assumptions.

Experiments were performed on simulated data and real cancer datasets for brain, ovarian and lung cancers (data is obtained from TCGA). Moreover, they used hotnet2 to generate the graph and its connections. In the case of simulated data, NoMAS was able to find the global optimal solution for sample sizes currently available; however, in the case of larger sample sizes, it may not find it. Although NoMAS was not entirely successful, the sub-optimal solution contained many of the genes in the global optimal solution. Interestingly, in all of the cases, the global optimal was among top ten solutions ever found. Therefore, this method can be applied for datasets with sample sizes currently available. In the case of real datasets, they compared their method with exhaustive algorithm and some greedy methods. Not only the proposed method found the optimal solution in all of the cases but also did it run faster for bigger k than the other methods. Moreover, NoMAS using the additive version of the proposed metric also outperformed the compared methods. Through literature review, they found that NoMAS discovered some gene subnetworks important in cancer, while they are not significant when considered individually.

pros and cons: Applying parallelization to speed up the convergence is an advantage of the algorithm. However, there are also several concerns regarding this paper, for example, the regulatory effects of genes (using a directed graph as an input) is not considered in the model. Also, they used their own metric for comparison, what about other previous ones, can they be used in the proposed algorithm as well? Besides, it is unclear why it would be necessary to use 2 different methods to assess the statistical significance.

possible extensions: It could be useful to find the top-k highest scoring subnetworks instead of only the top one; probably we could consider stratifying patients with respect to mutations in subnetworks, such that patient stratification and subnetwork discovery can be performed simultaneously. More types of genomic variants could be considered, and more clinical parameters other than survival time could be incorporated into a new and more general model.

B. pathTiMEx: Joint Inference of Mutually Exclusive Cancer Pathways and Their Dependencies in Tumor Progression [4]

Motivation and contributions: The main focus of this paper is centered around two crucial challenges: first, which mutations cause progression in tumor and second, determination of which mutations occur sooner and which ones happen in later stages. Finding these mutations and their order of happening is highly important for therapeutic purposes. In this paper, a probabilistic generative model named pathTiMEx was proposed in order to model progression of tumor at the level of pathways (more robust and reliable in comparison with single genes). In fact, pathTiMEx is a method based on a synergism that generalizes two well-known techniques, TiMEx and CBN. The former is a waiting time model for mutually exclusive cancer alterations, and the latter is a waiting time model for cancer progression at the level of genes. The proposed method successfully considered independent and dependent alterations in its model. Therefore, it can determine driver events and their corresponding order of occurrence. PathTiMEX was the first joint probabilistic generative model of cancer progression through multiple paths, at the level of mutually exclusive driver pathways.

Summary: Given two sets of waiting time for events happening in genes and mutually exclusive pathways and a vector corresponding to partitioning of the genes set, a poset (partially ordered set) operator is defined which deals with the order of occurrences of events (each event can happen only when all its parents have happened). A probabilistic graphical model was constructed. In the model, the whole system progresses until the observation time, which follows an exponential distribution with an unknown rate. The mutation in the gene with shortest waiting time determines tumor progression. Once all the parents of a pathway have mutated, waiting times of the genes in the pathway follow exponential distributions as well. The authors used the same framework as the previous methods for mutual exclusivity of pathways. Based on that, the status of a gene depends on its own waiting time (of mutation), waiting time of its pathway and the observation time (larger than waiting time). Since noise and uncertainty is an integral part of biological analysis, deviation from both mutual exclusivity and evolutionary constraints were included in a single error variable.

PathTiMEx maximizes log likelihood for inference and uses TiMEx method with default parameters for generating the initial solution of mutually exclusive groups of alterations. Then CBN estimates the waiting time rates for the pathways and error probability via expectation maximization. Finally, the progression between pathways is determined by SA. Afterwards, a stochastic optimization routine is used, which consists of two steps, one is responsible for optimizing the assignment of genes to pathways which is done by MCMC, and the other is dealing with optimization of evolutionary order in the level of pathways which is done via Simulated Annealing (SA), given the assignment of genes obtained in the previous step. These two steps are iterated until convergence.

As for the experiments, the authors did extensive experiments on simulated and real cancer datasets. For simulated data, they designed multiple scenarios with alteration frequency ranging from low (<1%) to high (>60%), different levels of noise and different sample sizes. They assigned 12 genes to 5 pathways randomly, with the requirement that each pathway had at least one gene. Further, the evolutionary constraint was generated as a DAG with different settings for edge density (uniformly sampled). The convergence of pathTiMEx increased by increasing sample size or decreasing noise rate, and also number of required iterations decreased in the stated settings. Meanwhile, its performance deteriorated in the presence of large sample sizes and high levels of noise. Regarding real cancer datasets, they studied two small and large datasets for colorectal cancer and another dataset for brain cancer (all available on TCGA). Under fixed linear progression with an undetermined number of stages, they performed optimization of assigning genes to pathways faster than the state-of-the-art method. The results of pathTiMEx were highly compatible with current knowledge and literature for all of the studied cancers. However, they reported slower convergence rate for the brain cancer because it was less known and more complex than the others.

Pros and cons: The advantage of the proposed method is that it has merits of both TiMEx and CBN. The pathTiMEx is both theoretically justified and applicable in practice. However, the simplifying assumption that pathways are mutually exclusive may not be practical in real life. Besides, mutations could be reversible in practice, not the irreversible assumption used in the paper.

C. Separating the Causes and Consequences in disease transcriptome [14]

Motivation and contributions: The inference of the molecular mechanism of a complex human disease can be enhanced by finding causes and consequences of that disease. Although techniques such as GWAS provide us with potential genetic causes, for therapeutic purposes we still need to know the mechanism of pathogenesis of the disease. Transcriptome can prepare the ground for such a purpose but it is challenging because signals from causes and consequences are twisted. Therefore, it seems crucial to find a novel way to make the best use of transcriptome, GWAS and other sources of information to tackle this issue.

The proposed pipeline is called DiseaseExPatho. Its first step extracts gene modules via ICA which is a matrix factorization method to obtain gene modules. In the second step, it detects gene enrichment via a new statistical inference method. The goal of this step is to label obtained modules in step one as differentially expressed and/or putative causal. In the final step, DiseaseExPatho uses a hierarchical ranking scheme based on gene regulation network in order to prioritize putative causal gene modules.

This scheme is novel because it provides an intuitive ranking of nodes which is consistent with structure of the regulation network. Moreover, unlike previous approaches, the proposed method ranks from top to bottom and both transcription factors (TF) and non-TF receive a meaningful rank based on their location in the network. The inputs of DiseaseExPatho are transcriptome profiles of the disease we are studying, gene regulation network and putative causal genes of the disease, which is based on GWAS study and might not be available for some diseases. The output of the proposed method is list of ranked and prioritized modules.

Summary: After some preprocessing steps such as data cleaning and merging datasets, a matrix factorization method called ICA decomposes transcriptome data into gene modules that correspond to pathways that co-express in a sample. Each module is a soft clustering of genes, with dominant genes having the highest positive or negative weights. After determination of gene modules, the authors applied a linear model to measure differential expression of them in healthy and non-healthy samples. Each module is labelled based on the significance of the coefficients corresponding to disease status variables and confounding variables. Further, the authors proposed a ranking mechanism for genes based on their positions in the regulation network, and each module’s ranking is the weighted average ranking of its constituent genes. Finally, the authors proposed an algorithm based on linear regression to find the putative causal modules based on GWAS of the disease. For a specific module and putative causal gene, when the coefficient of this linear model is significantly different from zero, the module is considered putatively causal. A similar inference approach was also adopted to find association of putative causal genes with differential expressions.

Transcriptome datasets for three psychiatric disorders (SZ, BD, MD), diabetes (T2D) and inflammatory bowel diseases (IBD) such as CD and UD were obtained from NCBI website; GWAS studies with specific p-values for genes and diseases were also obtained from sources such as dbGPA, NHGRI and NHLBI. Relation between GWAS and transcriptome was examined in the paper and enrichment of MD, BD, T2D and CD were obtained but not SZ and UC. Regarding the differentially expressed gene modules, after applying ICA and bidirectional linear model, 17, 16 and 8 putative causal gene modules were found for BD, MD and SZ, respectively. They observed that many of the gene modules showed a stronger enrichment of putative disease causal genes compared to the overall differential expression profiles. Furthermore, based on the stated linear regression model and correction of p-value, they observed that majority of putative causal genes were not differentially expressed for the studied diseases. Since GWAS data was not available for all complex diseases, they also proposed a ranking scheme based on regulation network. After ranking individual genes, modules were ranked according to the genes' weights in each module. Results stated that for studied diseases, putative causal modules are ranked much lower than other modules. The authors also studied biological functions of gene modules for psychiatric disorders and found interesting connections to synapses and neural systems.

Pros and cons: Hierarchical ranking scheme works even when the disease causal genes are unknown which is an important contribution of the proposed method in this study. The rank measure designed in this paper is perhaps the most creative part of this research. However, it is unclear how many times they ran matrix factorization in their pipeline. Moreover, they also claimed that for each gene expression matrix, they learnt 50 gene modules, but it is not clear how they came up with this number.

Possible extensions: The gene regulatory network could be a starting point for some new research problems. Prof. Ester thought that when the gene regulatory network is represented as a directed graph, it could be formulated as a set cover problem, such that the we could search for the smallest closest set of mutated nodes (genes) leading to all up or down regulated genes.

D. Complexes Detection in Biological Networks via Diversified Dense Subgraphs Mining [5]

Motivation and contributions: Understanding the organization and structure of biological processes and cellular components requires finding dense subgraphs within protein-protein interactions networks, because proteins interact with each other as a group to perform a complex biological function. In order to find diversified dense subgraphs, the authors first defined density, dense subgraph, coverage and diversity. This is the first method which brings diversification into dense subgraph identification problem. The key part of the proposed method is a set of efficient search trees that traverse all dense subgraphs via depth-first search method. Further, the authors proposed a potential score in order to guide the search tree and development of a pruning scheme based on density and diversity of subgraphs.

Summary: The inputs of the proposed method are a weighted graph, a density threshold, a diversity threshold, number of diversified maximal dense subgraphs and maximum marginal gain of an initial solution (maximum dense subgraph).

Dense subgraphs can be constructed one node at a time with appropriate ordering of the vertices to ensure the monotone decreasing (Pseudo- anti-monotonicity) property of density along with the growth in order to prune with density. Knowing the maximum dense subgraph at the beginning will help us to make online decision about the effectiveness of marginal gains for further subgraphs. Moreover, we need a proper ordering of nodes to build the search tree and construct the maximal dense subgraph iteratively. To find such an ordering, the authors defined the notion of potential specific to nodes.

Therefore, the proposed method first sorts (by GRASP method) nodes based on their degrees and selects the one with the highest degree. The algorithm expands this node to the dense subgraphs from it via a backtracking-based growth procedure (pruning and diversifying are also embedded in it). Whenever the expanding process is finished, the proposed method selects the next node (with highest degree) among all the nodes that have not been used in any of the dense subgraphs identified so far. This procedure repeats until there is no node to be considered. In the process of expansion, first we find all candidates and sort them based on their potentials. Later, for each candidate we add it to the current set if the density of the updated set is greater than or equal to the density threshold and the predicted marginal gain is not less than diversity threshold. Otherwise, declare the new set as a maximal dense subgraph. Pruning by both density and diversity is performed when the process is starting a new search subtree which is for the internal nodes, while diversifying is done when a maximal dense subgraph is found which is for the leaves of the tree. We prune a subtree whenever its predicted marginal value is lower than the threshold based on a fixed vertex. The output is the list of maximal dense subgraphs.

For experiments, the authors used five PPI (Protein-Protein Interaction) datasets of yeast and another one for human and also utilized different reference complex sets such as MIPS and String-GT for validation. For evaluation and comparison, the authors applied 3 different scores including fraction, accuracy and maximum matching ratio and compared their method to state-of-the-art techniques. In order to study the effectiveness of the proposed method, they studied different values for density threshold ranging from 0.5 to 0.9 and reported no significant difference in density or coverage; however, the proposed method did find more complexes. In cases where a predicted complex did not match a reference complex, the authors investigated functional homogeneity because the golden standard sets of protein complexes are possibly incomplete and such a result is not always undesired. The authors also studied the scalability of their method and showed that it is in fact linear, which is desirable. Regarding the human genetic dataset, they first partitioned the data and identified the maximal subgraphs in each partition, and then merged them into one set and did the diversifying task.

Pros and cons: For the pros one can state the following: speeding up the proposed algorithm by a partitioning method; smart ordering of candidates in the proposed method makes it more successful in enumeration; pruning in the proposed method can boost efficiency and diversifying can remove redundancy in finding candidates. Regarding cons, the paper is extremely unclear and poorly written, for example, some terms like “beta” are not defined in it. The optimal way for determination of density threshold is not well studied. Finally, it is not clear that how one can know weights of the edges in the graph.

Possible extensions: The method could probably be applied in other graph problems, like social networks. It is possible that the method could be combined with the previous papers in this section about cancer driver subnetworks.

2- Deep learning to the rescue of landmark genes

A. Gene expression inference with deep learning [13]

Motivation and contributions: NIH did an interesting project to selection ~1000 genes out of the human genome (~22000), which were later known as landmark genes, that can characterize the cellular states, to reduce experimental costs. These landmark genes are special because one can obtain around ~80% of the information of the other genes via them. However, the problem is that the original computational part of this project is based on linear regression. Obviously, such a simple model cannot capture non-linear relations between different genes. Therefore, providing users with new computational tools with such a capability is crucial and at center of this paper’s objective.

In order to predict the expression level of target genes based on landmark ones, the authors utilized a deep learning method which is a multi-task multilayer feed forward neural network. The input of this network is the training data of 943 landmark genes and the output is the predicted value for 9520 target genes.

Summary: The network as stated, has an input layer with 943 nodes, multiple hidden layers all with same number of nodes and hyperbolic tangent activation function in order to capture nonlinear relations. Linear activation function was used in output units. Sum of mean squared error for each output node was used. Due to memory constraints, 9520 output nodes were separated randomly into two sets, each with 4760 output nodes. With this modification, the authors tested three different architectures with 3000, 6000 and 9000 nodes in each hidden layer. The training algorithm was back propagation with mini-batch gradient descent. They also used dropout with 0%, 10% and 25% rates, momentum coefficient and a uniform distribution-based initialization method for weight

They applied the model on two datasets, including GEO microarray data and GTEx1000 RNA-Seq data and compared the results with linear regression and KNN. Based on the obtained results, the proposed method with 10% dropout rate and three hidden layers each with 9000 nodes reached the best performance. Based on visualization of major weights, particularly those of the output layer, the authors claimed that there could be a strong local correlation between the landmark genes and target ones. Moreover, they also analyzed captured non-linearity via coefficient of determination and showed that intermediate hidden nodes captured some non-linearity that would be ignored by other compared method.

Pros and cons: The authors' efforts to interpret weights and structure of the learned network was an interesting way to obtain more knowledge from the black box of deep learning. Examining the coefficient of determination between the output of the last hidden layer and the final targets was a useful idea to analyze the nonlinearity captured in the model. However, as for the cons, it is not clear why they used another metric for test samples; the resource limitations that forced them to separate the network has changed the network structure and possibly have negatively affected the performance.

3- Patient stratification

A. Bayesian biclustering for patient stratification [9]

Motivation and contributions: This paper tried to stratify patients via a novel probabilistic approach. Since one of the main challenges in this task is dealing with noisy and uncertain observations, the authors looked for a solution to tackle this issue as well. Moreover, the problem of fair comparison metrics was also investigated. Finally, the authors were also curious about taking advantage of multiple input datasets to see how influential it would be on patient stratification.

This is the first probabilistic integrative model for biclustering and, because of being probabilistic, it can successfully deal with noisy data while stratifying patients. To better serve this purpose, the authors included prior knowledge in the model. The model is based on Bayesian networks, in which diverse types of datasets can be utilized. Another interesting contribution of this paper is that the detection of natural number of clusters is done by the proposed method itself; furthermore, there is no assumption on the number of clusters for samples and features, unlike deterministic methods. Finally, the authors also proposed that the log-rank p-value be used for comparing clustering algorithms.

Summary: In this paper, the authors tried to stratify patients into molecular subtypes via probabilistic graphical models. This probabilistic method uses multiple datasets with different numbers of features but the same sample size. There are several parameters and variables related to the proposed probabilistic method, such as observed variables and hyper-parameters which are provided as inputs to the model, and model parameters and latent variables, which are learned during the learning process. These parameters are learned by Gibbs sampling method in an iterative way based on conditional probability. Because of the stochastic nature of the proposed method, several independent executions were performed and a consensus matrix was generated to determine the final clusters.

Datasets were obtained from TCGA for breast and brain cancers which includes somatic point mutations, gene expression and CNV. Based on 50 independent executions, strong prior improves sample clustering but mildly disrupts gene clustering. Regarding the integrative study, gene expression alone has the best performance and including point mutations did not cause any improvement. Because these mutations may be different but cause the same phenotypes among individuals. As for the CNV data, as they are highly corrupted with noise, they were not helpful for integration modelling. Finally, although NMF was slightly more robust, the proposed method B2PS had more meaningful results in terms of patient stratification and clustering. Interestingly, with an upper bound for the number of clusters based on NMF for B2PS, it performs better than NMF.

Pros and cons: Being probabilistic is more informative because it can capture similarities between different subtypes (in probabilistic approaches, it is possible to have a shared feature between two clusters). As stated above, applying prior knowledge is an important advantage of the B2PS. Not only is there no assumptions in row and column clusters, but also the probabilistic approach gave more informative clusters in both.

As for the cons, the computational complexities of the proposed method and NMF were not discussed. In general, if we are working on overall survival time and know it as an output, one might claim, why not applying a conventional machine learning method like neural networks to predict it? Why are we doing patient stratification instead?

The authors claimed that this probabilistic approach is beneficial because one can capture similarities between clusters, however, consensus matrix procedure blocked this advantage by making it a hard clustering.

Possible extensions: More data types could be incorporated into the model. Some discrete data types used in the model could possibly be turned into continuous ones.

B. Patient-specific data fusion for cancer stratification and personalized treatment [8]

Motivation and contributions: Similar to the previous paper, this paper is also about patient stratification via multiple datasets (an integrative approach). This paper attempts to stratify patients to subgroups, identify driver genes for ovarian cancer and find new drugs as a treatment for specific subgroups of this type of cancer. In fact, the proposed framework is the first method that addresses all these three issues in a unified approach.

Summary: This paper focused on fusion (integration) of three datasets including patients, somatic point mutations and drugs. For this purpose, the authors used Non-negative Matrix Tri-Factorization (NMTF) with regularization terms. The aim of applying these regularization terms is to include information of molecular networks (between genes) and drug compounds (chemical similarity) to the objective function of NMTF. Therefore, similar genes and drugs are forced to be in the same clusters corresponding to genes and drugs.

This method factorizes matrices related to these three datasets to other low-dimensional matrices and bicluster them. For example, in the case of patients, an n*k matrix means n patients are assigned to k clusters, and the highest value for each row indicates membership to that cluster. Unlike the previous paper in this subsection, the number of clusters is required to be determined before the factorization process (number of clusters is rank of the matrices for genes, drugs and patients). Similar to the previous paper, the authors generated consensus matrix over multiple runs, and used the dispersion coefficient for the analysis of cluster stability. In fact, they used these consensus matrices to perform patient stratification and cancer driver gene prediction with hierarchical clustering and hypergeometric test, respectively. Moreover, they used matrix completion property of NMTF to repurpose new drugs for specific genes.

Similar to the previous paper, the authors obtained their datasets from TCGA as well. The molecular network stated above was obtained by a combination of protein-protein interactions, genetics interactions and pathway analysis. Further, they used an iterative network propagation approach as a preprocessing phase to deal with sparsity of somatic mutations dataset. As for the drug-target interactions, they used DrugBank database and calculated the similarities between compounds by Tanimoto similarity coefficient.

In order to validate the results, some ground truth information regarding patients such as overall survival, tumor size, etc. as well as driver genes was obtained by the authors. Moreover, they used Kaplan-Meier survival curves and log-rank p-value to analyze the significance of findings. By applying the proposed method, Graph-regularized NMTF (GNMTF) and computing consensus matrices, the most stable values for the numbers of clusters were obtained. There is a good agreement between calculated clusters for somatic mutations and clinical data such as survival rate, tumor size and age; moreover, two other studies also reported the same number of clusters. Compared to Network-based Stratification (NBS), only the proposed approach could achieve statistically significant clusters with different survival rates. In order to find driver genes, mutated genes that are strongly related to known driver genes are extracted based on their association scores (score⩾0.9). Three different databases including COSMIC, CCGD and IntOGen validated 40% of the predicted driver genes. Furthermore, analysis of top two predicted driver genes indicated a strong relation between two known driver genes. Regarding the analysis of the drugs, 37% of the predicted DTIs were confirmed by either MATADOR or CTD. Finally, based on AUC of ROC and PR, the best result was obtained when all datasets were considered in the integration (fusion).

Pros and cons: It considers somatic mutation data along with other datasets in a deterministic way. The incorporation of graph-regularization terms perhaps made this method more successful regarding somatic mutations. Including inter-relations between genes and drugs is also another advantage of this method. As for the cons, there is no discussion regarding computational complexity of the proposed method or the compared ones. Besides, the cancer driver gene prediction used a pretty heuristic method and lacked theoretical justification.

4- Clonality inference

A. Clonality Inference from Single Tumor Samples Using Low Coverage Sequence Data [10]

Motivation and contributions: Somatic mutations can be helpful to determine how pure a tumor sample is and find its sub-clonal compositions. However, the challenge is that this type of mutation is heterogeneous and current state-of-the-art methods for inferring sub-clonal information require multiple samples from a single tumor, which is hardly applicable for the majority of cancers. Therefore, it is of utmost importance to infer these sub-clones and their compositions via a method based on single sample information. In order to determine the number of sub-clones and the purity of a tumor sample, the authors proposed CTPsingle which has a robust clustering framework based on beta-binomial mixture model and moreover, performs phylogeny analysis via a fast-mixed integer linear programming formulation. As stated above, the goal is to find purity and sub-clonal compositions, thus, these two are the outputs of this problem. As for the input side, the input of the proposed method is the sequence data of a single tumor sample and its detected mutations. Therefore, we know the number of variant read counts and total read counts for a specific position in the sequence.

Summary: CTPsingle has some assumptions on the input side, for example, copy number variation is not considered here. After some processing in the input side, it uses a beta-binomial clustering mechanism with probability of success drawn from Dirichlet Process which is itself based on Gamma and Beta distributions. Inference is performed via standard Markov Chain Monte Carlo (MCMC). After clustering phase, we have number of clusters, a.k.a, sub-clones, and mean allelic frequency for each of them. Thus, cellular frequency for sub-clones can be calculated easily. When we know cellular frequency, we can estimate the purity of the tumor sample and cancer cell fractions. Further, RMSE of cancer cell fraction will be used as one of the evaluation metrics.

They did experiments on simulated scenarios with two types of data, including low coverage and ultra-high coverage, and obtained three metrics for evaluation:1) estimated purity, 2) predicted number of clusters and 3) RMSE of cancer cell fractions. They compared CTPsingle with three state-of-the-art methods. According to the obtained results, CTPsingle had superior performance to the other methods in most of the experiments, even when the coverage was low. For future work, the authors are also extending their experiments with clinical data related to prostate cancer. Interestingly, there was a good agreement between predicted result of CTPsingle and the actual one.

Pros and cons: There are several advantages with the proposed method such as algorithm finds number of clusters itself automatically. Working with single sample and low coverage data is also a huge advantage for CTPsingle. Using only freely available libraries can also be another influential factor for this method. As for the cons, this method ignores CNV which is an important aspect in many cancers especially prostate cancer and it can be a good direction for future works. The description is too brief and the clustering process is not introduced.

Possible extensions: CTPsingle cannot handle multi-focal tumors, which could be a future research direction.

5- Computational drug discovery

A. DrugE-Rank: improving drug-target interaction prediction of new candidate drugs or targets by ensemble learning to rank [11]

Motivation and contributions: identifying drug-target interactions is highly important for therapeutic purposes because experimentally exploring possible drugs and compounds is expensive and time consuming. Therefore, developing reliable computational tools is highly important for finding the best candidates for biochemical experiments. Computational approaches for this goal can be categorized into two families, namely feature-based methods and similarity-based methods. Both approaches have their own merits, thus, taking advantage of both of them can be helpful for solving the drug-target interactions problem, particularly for new candidate drugs or targets. In this paper, the authors proposed DrugE-rank which makes the best use of both of the mentioned approaches in an ensemble learning way in four steps. In step 1, it computes the feature vector of the target and an arbitrary drug, further, it estimates the similarity of all targets and all drugs. In step 2, the proposed method runs six component methods by utilizing the computed similarity in step 1. The output of this step are pairs of features corresponding to scores. In step 3, a feature vector for the arbitrary selected drug in step 1 is generated from drug and target features and pair features of step 1 and 2, respectively. Finally, in step 4, all of the features will be the input of Learning To Rank (LTR) and the output will be the list of ranked drugs for that target.

Summary: In the first step for a new given target, first they generated 147 features via CTD method, and they computed genomic similarity between that target and other targets by Smith-Waterman score. For each drug, they computed 36 features (physicochemical descriptors) by RDKit and similarity between two drugs is calculated by Tanimoto coefficient. In the second step, they ran six component methods including KNN, BLM-svc, LapRLS, NetLapRLS and WNN-GIP to predict the score of drug target interaction between all drugs and the given new target and normalize this score to be a value between 0 and 1. In the third step, the final feature vector for all of the drugs and the new target is calculated via concatenating target feature vector (147 features), drug feature vector (36 features) and pair feature vector based on the previous step (6-dimensions vector showing the strength between drugs and the target). Therefore, each drug target pair is represented via 189 features which are the input of the last step, i.e., LTR. In the last step, feature vector of a drug according to step 3 will be the input of LTR method and the ranked list of drugs is the final prediction result or the output.

They obtained their dataset from DrugBank which contains 1242 drugs, 1324 targets and 5701 interactions. They divided the dataset into five smaller datasets which contains four sets of FDA-approved drugs and 1 set of experimental drugs. They compared the proposed method with all of the similarity-based methods which they used in their ensemble and two other feature-based methods, with AUPR as the evaluation metric. The authors used Data-1 with 5-fold cross validation (1 for test, 1 for training LTR and 3 for training component methods) and paired t-test for statistical analysis. Moreover, they used Data-2 and 3 for testing when Data-1 was used for training (four for training component methods and 1 for LTR). Finally, they used Data-1 to 4 for training and Data-5 for testing. All experiments were performed 10 times. Based on the obtained results, for the experiment over Data-1, BLM-svr had the best performance among the compared methods and the proposed method was able to overcome it and they observed similar result of the second experiment based on Data-2 and 3 and even Data-5, which means DrugE-rank can outperform all the other methods when it has all of the features.

Pros and cons: Applying multiple methods and finding a way to use them in an ensemble way provided the proposed method with the competitive edge to perform better than the other compared methods. The step 4 description was not very clear.

Possible extensions: It could be formulated as a deep learning problem. The input could be SMILES representation; the feature vector produced in the autoencoder could possibly be inputs to feature-based methods mentioned in the paper.

6- Role of EHR in Precision Medicine

A. The challenges in using electronic health records for pharmacogenomics and precision medicine research [12]

Motivation and contributions: The goal is to use patients' genetic information to estimate risk of disease, prevent illness and find the best treatment. Study of pharmacogenomics, i.e., genetic variants that affect drug response and efficacy, can be used in clinical purposes. However, GWAS studies need to be diverse in term of population and currently, most of the studies are based on European populations. Recent emergence of EHR linked to biorepositories provides us with fast, reliable and cheap data collection for GWAS studies. Further, one can select a wide range of phenotypes. In this paper, the authors used EHR linked to a biorepository in order to analyze drug response in an African-American cohort of 12,000 patients for two clinical treatments: 1) the use of antihypertensive medication to lower blood pressure and 2) use of lipid lowering medication to lower blood level of LDL-C. Particularly, they tested SNPs for association with change and percent change in blood pressure or level of LDL-C in blood.

Summary: Dataset was obtained from BioVU and DNA were collected from blood samples of mostly non-Europeans which contain 200,000 SNPs with certain allelic frequency (116,000 were available after quality control). Regarding phenotyping, they extracted electronically systolic blood pressure, diastolic blood pressure and LDL-C from EHR records. The authors considered post-medication if a prescription for an antihypertensive or lipid lowering drug was found prior to the experiment and pre-medication otherwise. This paper only used patients' data with both pre- and post-medication information available, along with BMI (2653 and 1244 patients who had systolic blood pressure and diastolic blood pressure for both pre- and post-medication, and both of them for LDL-C respectively). For statistical analysis, they used LR with difference between the median of pre- and post-medication and percent change between the median of pre- and post-medication as dependent variables. For each dependent variable, three models were run including: 1) unadjusted, 2) adjusted for age and sex and 3) adjusted for age, sex and the first three principal components of ancestry. They tested the association of SNPs on the Metabochip with the change in blood pressure and LDL-C measurement with the use of antihypertensive and lipid lowering medications respectively. After correcting for multiple testing, they did not find significant novel association, nor did they replicate previous studies, which is possibly because of sample size and statistical power. Another possibility is that the studied populations was African-American in this paper, while in the previous studies, they were mostly European. Another explanation is that although they had variants in the gene regions of significant SNPs which were identified in the previous studies on the investigated medications, they did not test most of the specific previously identified SNPs.

Pros and cons: They used a large population and they claimed that it is not common in the field to use a large cohort for the association studies.

As for the cons, a limited coverage of GWAS genotyping platform is a challenge in pharmacogenomics. They ignored some factors, such as dosage of medication in their phenotyping process, it would have been better to mention previous studies in this regard as well. The main problems are that the design of the study and cohort selection were only vaguely described and more importantly, they did not test most of the previously identified SNPs. All of the differences between this study and previous ones can be related to totally different populations (European and African-American).

III. Conclusion

In this review, 10 recent computational papers related to Precision Medicine were presented. All of these papers were published in top bioinformatics conferences in 2016 and covered most aspects of applications of computational methods in Precision Medicine.

The first problem in solving a biological problem by computational methods is availability of data, especially when is comes to machine learning and data mining. Based on the studied papers, it is fairly rational to deduce that, in order to study cancer, one of the best data source is TCGA. In most of the papers related to cancer, data were obtained from this website which indicates applicability, ease of use and diversity in its datasets. For the case of PPI studies, yeast can be a good choice (available online on MIPS and SGD) because it is not too large to incur huge computational cost and also not too small to make the process trivial for the algorithm. For the case of human studies, String-GT can be used. In the studies related to drug design, DrugBank and the list FDA approved drugs were the most popular sources among the studied papers. Finally, for studying gene expression data, GEO, GTEx and 1000 Genomes are highly recommended, especially the last two, which are based on RNA-Seq platform.

When the data is inputted and output is generated, we should choose a metric to analyze and evaluate the output.

In the studies related to PPI, three metrics are recommended as follows: fraction, accuracy, and the maximum matching ratio. Fraction is defined to be the fraction of pairs between predicted and reference complexes with an overlap score no less than 0.25; accuracy is the geometric mean of two other measures, namely the cluster-wise sensitivity (Sn) and the cluster-wise positive predictive value (PPV). Finally, Maximum matching ratio is based on a maximal one-to-one mapping between predicted and reference complexes.

For patient stratification, metrics such as Kaplan-Meier survival curves, as well as the log-rank p-value, can be applied to measure the significance of the difference in survival between different patient clusters. Moreover, in order to study stability, Cophenetic Correlation Coefficient can be used.

In drug design studies, in order to calculate similarity between two drugs, all of the papers applied Tanimoto coefficient. However, in general, sum of mean squared error (also RMSE), AUCPR and AUC ROC are the most popular choices similar to other machine learning studies.

When we are studying a biological problem with many aspects, we may see one that has a good p-value by log-rank test or any other statistical metrics just by chance, and this error is assessed by the permutational p-value. Therefore, sometimes it is essential to apply more than one statistical significance test.

In general, for designing an experiment it is extremely important to form case and control groups carefully. These two groups should only be different in a variable that we are enthusiastic to know more about and assigning randomly is a reliable approach to eliminate the confounding effect.

In the future, one of the main challenges regarding application of deep learning in biological problems is interpretation of weights in the learned network, this task is straightforward in Computer Vision and image processing, however, in bioinformatics and computational biology it requires further and deeper investigations. It is possible that applying unsupervised deep networks such as autoencoders which were successful in chemical design could also be informative in computational genomics.

For patient stratification, both papers unanimously suggested to integrate other types of data and study their effect as well, these datasets can range from methylation to fusion genes. Another good direction regarding the computational aspect itself can be working on combination of deterministic and probabilistic methods to get probabilistic non-negative matrix tri-factorization or investigate matrix factorization by powerful deep learning models wherever we have enough data.

For clonality inference or intra-tumor heterogeneity, obviously in the presented paper the important factor of copy numbers variation was disregarded. Therefore, the most crucial future work for this aspect of Precision Medicine is to consider copy number variations (aberrations) which is an important contributing factor in cancer, especially prostate cancer.

For the case of finding the optimal set of causal genes (rank genes), one can consider it similar to a set cover problem, i.e., finding a set of mutated genes which cover all of up-regulated and down-regulated genes with minimum distance to the effect one. This can be used as a criterion for ranking genes and finding the desired set.

As for the pathways, we reviewed pathTIMEx in this paper. The authors provided readers with some interesting ideas for future works. This method is so simplified and does not consider the reversibility of mutations or cross-talk between pathways. Adding these constraints will make the method more accurate and more consistent with the true nature of the problem. Further, this method for large datasets with a high level of noise requires modifications and the authors proposed to model temporal dependencies between waiting times, specifically accounting for false positive and false negative dependencies.

For PPI dense subgraph mining problem, the authors proposed to extend the method for larger and denser networks and also investigate its applicability for other networks such as social ones. Moreover, they suggested to identify homology relationships between sequences and orthology inference across multiple species.

Finally, for finding the mutated subnetwork, the authors recommended to extend the proposed method with utilizing other types of data such as copy number aberrations which are more complex than single nucleotide variants/indels. Furthermore, in addition to survival time, other clinical parameters can also be studied for their association with mutated subnetworks. Regarding imbalanced datasets and number of iterations, Dr. Fabio Vandin, the co-author of [4] said: "*Since we use Monte Carlo estimates, unbalanced data is not a problem. Once you fix the error probability that you want, you can derive an upper bound of the number of iterations to get the desired error probability. This is a standard approach for (graph) color-coding algorithms.*" However, these remarks are still not clear to the authors of this review and require further investigations.

IV. References

[1] https://ghr.nlm.nih.gov/primer/precisionmedicine/definition
[2] https://ghr.nlm.nih.gov/primer/precisionmedicine/precisionvspersonalized

[3] Tommy Hansen, et al. Mutated Subnetworks Associated with Survival Time in Cancer, RECOMB 2016.

[4] Simona Cristea, et al. pathTiMEx: Joint Inference of Mutually Exclusive Cancer Pathways and their Dependencies in Tumor Progression, RECOMB 2016.

[5] Xiuli Ma, al. Complexes Detection in Biological Networks via Diversified Dense Subgraphs Mining, RECOMB 2016.

[6] Gomez-Bombarelli, et al. Automatic chemical design using a data-driven continuous representation of molecules, arXiv 1610.02415v2 2017.

[7] Goodfellow et al. Deep learning, MIT press, 2016.

[8] Natasa Przulj, et al. Patient-Specific Network Data Fusion for Stratification, Biomarker Discovery and Personalizing Treatment, ISMB 2016.

[9] Sahand Khakabimamaghani and Martin Ester. Bayesian Biclustering for Patient Stratification, PSB 2016.

[10] Nilgun Donmez, et al. Clonality inference from single tumor samples using low coverage sequence data, RECOMB 2016.

[11] Qingjun Yuan ,et al. DrugE-Rank: improving drug–target interaction prediction of new candidate drugs or targets by ensemble learning to rank, ISMB 2016.

[12] Sarah M. Laper, et al. The Challenges in Using Electronic Health Records for Pharmacogenomics and Precision Medicine Research, PSB 2016.

[13] Yi fei Chen, et al. Gene expression inference with deep learning, ISMB 2016.

[14] Yong Fuga Li, et al. Separating the Causes and Consequences in Disease Transcriptome PSB 2016.