 doi:10.20944/preprints201804.0352.v1

Integrative approaches to reconstruct regulatory networks from multi-omics data: A review of state-of-the-art methods

Nisar Wani, Khalid Raza

Abstract—Data generation using high throughput technologies has led to the accumulation of diverse types of molecular data.These data have different types (discrete,real,string etc.) and occur in various formats and sizes. Datasets including gene expression, miRNA expression, protein-DNA binding data (ChIP-Seq/ChIP-ChIP), mutation data(copy number variation, single nucleotide polymorphisms), GO annotations, protein-protein interaction and disease-gene association data are some of the commonly used genomic datasets to study biological processes. Each of them provides a unique, complementary and partly independent view of the genome and hence embed essential information about their regulatory mechanisms. In order to understand the functions of genes, proteins and analyze mechanisms arising out of their interactions, information provided by each of these datasets individually may not be sufficient. Therefore integrating these multi-omic data and inferring regulatory interactions from the integrated dataset provides a system level biological insights in predicting gene functions and their phenotypic outcomes. To study genome functionality through interaction networks, different methods have been proposed for collective mining of information from an integrated dataset. We survey here data integration approaches using state-of-the-art techniques such as network integration, Bayesian networks, regularized regression (LASSO) and multiple kernel learning methods.

Index Terms—network inference, data integration, regulatory networks,transcription factor, gene expression

1 Introduction

LIVING cells belong to a class of highly studied complex systems, whose functioning continues to allure the research community. The cell houses a diverse molecular structure, comprised of a large number of molecular entities viz. genes, proteins, metabolites and mRNA, forming a complex and dynamic molecular machinery. Earlier, experimental approaches used to focus on the effect of individual molecular entities or environmental factors to study a particular cell function, without taking into consideration the effect of other important factors. However, these methods are limited by the scale and time of the study, hence making it difficult to obtain a system-wide cellular response of a stimulus e.g., a drug dose, an environmental perturbation or a gene knockdown experiment.

- *Nisar Wani ,Govt. Degree College Baramulla,J& K,India.*
- *Dr. Khalid Raza, Department of Computer Science, jamia Milia Islamia,New Delhi India. E-mail: (kraza@jmi.ac.in).*

For a holistic measurement of cellular responses and identification of functions, novel, high throughput technologies that generate genome-scale data sets open global perspectives of living organisms at the molecular level. Such a collective quantification and characterization of pool of diverse bio-molecules that give rise to structure, function and dynamics of organism is what is commonly nowadays referred to as Omics layers (Gligorijević & Pržulj, 2015). For example, a whole set of proteins and their interactions form the proteome , mRNA abundance together with factors responsible for transcription give rise to a transcriptome, genes within the DNA form the genome, metabolites within the cell form metabolome and the diverse phenotypes as phenome. The mechanisms which translate the functional codes within the genome depend on various intermediate omics layers and their inter-relationships (interactome) giving rise to complex phenotypic traits in the phenome (Gligorijević & Pržulj, 2015).

High throughput technologies such as , protein mass spectrometry, yeast two-hybrid assays, microarrays and a range of protocols under the banner of next generation sequencing (NGS) viz; RNA-seq, ChIP-seq, DNase-seq and miRNA-seq etc. produce vast amounts of disparate biological data (Omics), these data sets together with a knowledge-base of experimental literature on biochemical processes are disrupting earlier notion of isolated functionality and linear causality pathways and challenging us with multiple functions. For example, a protein produced from the information contained in a genetic sequence may function as a transcription factor (TF) binding to regulatory sites on promoter region of a gene or multiple genes, an enzyme responsible for catalyzing a metabolic reaction, a cellular component or as a component of signal transduction pathway, thereby implying networks of interacting biological components or regulatory systems. Therefore, it is a widely accepted notion in the modern era of high throughput biology that understanding any biological system in a comprehensive manner can only be achieved through an integrative analysis of relevant omics datasets. Such an analysis approach reflects the need for building a joint model that simultaneously captures the information content of all the data involved in the unified model (Ritchie et al., 2015).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 April 2018 doi:10.20944/preprints201804.0352.v1

2 Need for Multi-omics Data Integration

The task of understanding the mechanisms by which living systems carry out their functions can be divided into two sub-tasks, i) identifying the components that make up these systems and ii) interaction between them that result in either function or dysfunction in such systems. As for the first sub-task, there has been an extensive research effort carried over a period of several decades, to characterize the structure and function of individual cellular components. Research endeavors that lead to completion of Human Genome Project , the growth of Protein Data Bank (PDB) (Bank, 1971), Signaling pathway databases (KEGG) , Gene Ontology Consortium (GO) (G. O. Consortium, 2014), ENCODE project(E. P. Consortium et al., 2004), Roadmap Epigenomics (Bernstein et al., 2010) and 1000 genomes (Siva, 2008) projects have generated huge databases and knowledge-bases that catalog both structural and functional aspects of elements of life. But in order to understand the underlying mechanisms and the overall functioning of the whole system, biological datasets need to be integrated under a mathematical or relational model that can describe the relationships between these components contrary to single data-type study designs.

Data integration therefore refers to the fusion of multiple Omics datasets, in order to build an informative model that can yield comprehensive results to our queries in terms of cellular functions/dysfunctions and phenotypic traits which are likely to be an outcome of the interplay between various cellular components at various levels of regulation. Our prime motivation in fusing multiple genomic data types into an integrative analysis framework is to identify key biological factors, that can explain a biological mechanism and help researchers in predicting a disease risk or a certain phenotypic outcome.Data integration may provide the platform for discovery of biologically important factors and their inter-relationships with improved accuracy.Besides modeling the complexity, interaction between single nucleotide polymorphisms (SNP) data, copy number variations (CNVs), gene expression profiles, methylation data, metabolomes proteomes may help us in improving our understanding of biological mechanisms underlying a complex-trait architecture.

With data integration, the scope for exploring more robust and new research questions becomes more open despite the challenge of assembling all the heterogeneous datasets in a biologically relevant manner.These challenges may be due to missing and noisy data, different sizes and data types across multiple datasets and the measurement errors that may lead to different correspondences and correlations from different technologies.Over the years a number of tools for integrative analysis of multiple genomic datasets have been developed , but no single approach is reported to perform optimally for all the studies, as observed by (Marbach et al., 2012), the performance and robustness of inferences greatly improves by combining predicted results from multiple inference methods together with integrating diverse datasets.Therefore, comprehensive and more inclusive toolbox is the need of the hour to discover,interpret and understand the intricacies of biological systems.

3 Data Integration Approaches for Omics Data

Large and heterogeneous omics datasets are being generated across all branches of life sciences. These datasets hold great promise not only in terms of scalable and unbiased investigation of biological systems, but also leading to conceptual development and new discoveries. Because most of such studies are publicly funded, therefore a large number of omics datasets find their way into publicly available online database resources. With such an unprecedented rate of biological data generation and its accumulation, there is a growing concern about much of this data not being used or being fully analyzed, thus creating a greater disparity between data generation and data utilization.Although these datasets are publicly available, they cannot simply be put into a mathematical framework or fitted straightaway in a statistical model. There are peculiar challenges to integrate these datasets. These challenges arise because of the differences these datasets manifest in their size, format and dimensionality, noisiness, information content and their mutual concordance. Also before integrating multiple datasets, it is imperative to evaluate each of the data types for quality. For genomic datasets such as, DNA-sequencing, RNA-sequencing, ChIP-sequencing and genome-wide methylation profiling methods etc., it is essential that the quality control steps be implemented (Patel & Jain, 2012; Wani & Raza, 2017).

Another very important factor in data integration is to handle curse of dimensionality which is most common in biological datasets. This problem can be solved by using data reduction technique that can select limited number of variables with high predictive power for single data type studies, besides, data reduction as a pre-processing step can also be applied to multiple datasets for performing integration analysis. Datasets with dimensionality issues offer a limited statistical power compared to data where the number of independent variables and samples is even to some extent. Since biological datasets occur with inherent skewness in size, investigators apply some sort of data reduction before performing association, correlation or modeling analysis. Reducing the amount of data through filtering strategies such as, principal component analysis (PCA), matrix factorization (MF) and singular value decomposition (SVD) ensures data integration , the selection of more robust features and analysis of a small and refined dataset.Data reduction can improve the computational efficiency of the inference algorithms and can potentially reduce the burden of testing multiple-hypothesis.It is imperative to perform some level of data reduction when working with complex models that can explore millions of measurements from a single dataset.Therefore applying data reduction techniques is a requirement for analyzing single datasets or performing integrative analysis across a diverse set of data.Various approaches have been developed for data integration, these approaches are primarily categorized on the basis of (a) type of data and (b) integration strategy.

3.1 Approaches based on data type

Although a clear cut framework and definition of data integration is hardly available in literature, various studies

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 April 2018 doi:10.20944/preprints201804.0352.v1

have proposed certain methods in this regard. For instance, (Lu et al., 2005) in a functional genomics context defines data integration as a process whereby data from different sources are combined in a statistical manner to make large-scale inferences for obtaining a holistic view of entire genome. Studies such as (Gligorijević & Pržulj, 2015; Hamid et al., 2009), categorize data integration into homogeneous and heterogeneous methods based on the type of data being combined. These datasets are produced by different experimental protocols, occur in different formats and are composed of different data types. For example, microarray gene expression data, gene expression profiles from RNA-seq experiments, protein data, SNPs, mutations, DNA sequences, CNVs and interaction data differ in their source, structures, dimensions and formats.

Taking these differences into account, datasets are treated as homogeneous, where the data is drawn from similar sources, having same data type and possessing a uniform format across different experiments. For example, all gene expression data from multiple samples or conditions or time points. Similarly all proteins, SNPs, CNVs or clinical data from multiple studies can be combined in a homogeneous integration. On the other hand, heterogeneous data sets include datasets generated from different sources, different experimental protocols, having different types (sequences, graphs, real numbers, integers, categorical) and possessing varying formats. For example, DNA-sequencing, gene expression, proteins, Chip-seq, CNV data, SNP data and clinical data.

A recent review on Network based integration methods by (Gligorijević & Pržulj, 2015) treat datasets with similar nodes (e.g. proteins) and different edges (e.g. gene interaction networks, protein-protein interaction networks, etc.) as homogeneous datasets and while on the other hand datasets that characterize various biological entities and embed different types of relationships are grouped as heterogeneous. They are represented as inter-relation collection of networks with different nodes and edges. Representations such as gene-disease association (GDA) networks, drug-chemical similarity (DCS) networks and protein-protein interaction (PPI) networks form a heterogeneous network with different types of nodes and edges.

3.2 Approaches based on integration strategies

In (Ritchie et al., 2015), authors categorize data integration methods based on how individual datasets are combined in the integration process into multi-staged integration analysis and meta-dimensional analysis.

A multi-staged analysis approach operates in a linear or hierarchical step wise manner. The model construction using this approach accepts data either as numerical or categorical features, for example, gene expression values denoting level of expression are continuous and numerical , whereas the indication of over/under expression can be represented using values from the categorical domain.

Meta-dimensional analysis approach on the other hand is based on the fusion of multiple features or scales. The aim is to find a complex meta-dimensional model in which multiple variables from different data types are combined simultaneously. It is broadly divided into three strategies, early integration (concatenation based) , late integration (model based) and intermediate integration (transformation based) as depicted in Fig. 1. These integration strategies were initially reported in (Pavlidis et al., 2001) and have been adopted by a number of meta-dimensional analysis studies, for example, protein function prediction (Lanckriet et al., 2004), protein classification (Lanckriet et al., 2003a), gene function prediction (Mostafavi et al., 2008) and protein-protein network inference (Yamanishi et al., 2004).

(a) *Early integration* or concatenation based approach builds a single model from a single composite dataset formed as a result of combining multiple separate datasets. To achieve this, the individual datasets need to be transformed into a uniform representation. This transformation may also result in a loss of information in certain scenarios (Pavlidis et al., 2001; Žitnik & Zupan, 2015a). An advantage of this integration strategy is that, once we figure out how to fuse variables into a unified matrix, it becomes relatively simple to analyze continuous and categorical data by using any statistical method.

(b) For *late integration* or model based approach, training sets derived from multiple data types are used to build separate models.These models created during the training phase are then combined into a unified final model, preserving properties which are data specific.This type of integration strategy is driven by individual hypothesis and analysis specific to a particular model and a procedure that combines the resulting models in a meaningful and coherent manner.Although we will be able to analyze each dataset individually using model-based integration, building models in isolation ignores their mutual relationships and therefore can drastically reduce the performance of the final model (Gevaert et al., 2006; Žitnik & Zupan, 2015a).

(c) In *intermediate integration* or transformation based approach individual datasets are combined through development of a joint model. The datasets are initially transformed before the integration process. For example, while handling classification problems we might transform the data into a similarity matrix or a graph or a covariance matrix before joining them. The transformation-based integration approaches preserves the data-type-specific properties of each dataset during the process of transformation and generate appropriate intermediate representations of the individual datasets with the advantage of retaining most of the information content of the original data. The approach offers a versatile framework to integrate diverse types of datasets, viz. continuous or real values, categorical data and sequence data, provided the data contain unification features, such as gene identifiers connecting different genomic data types (gene expression, PPI,TF binding etc.). Moreover, this approach offers robust handling of data having different measurement scales. This strategy has been reported to produce high prediction accuracy in (Lanckriet et al., 2004; Gevaert et al., 2006; Žitnik & Zupan, 2015a; Van Vliet et al., 2012), while on the other hand results from (Özen et al., 2009) have reported superior performances of concatenation and model based

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 April 2018 doi:10.20944/preprints201804.0352.v1

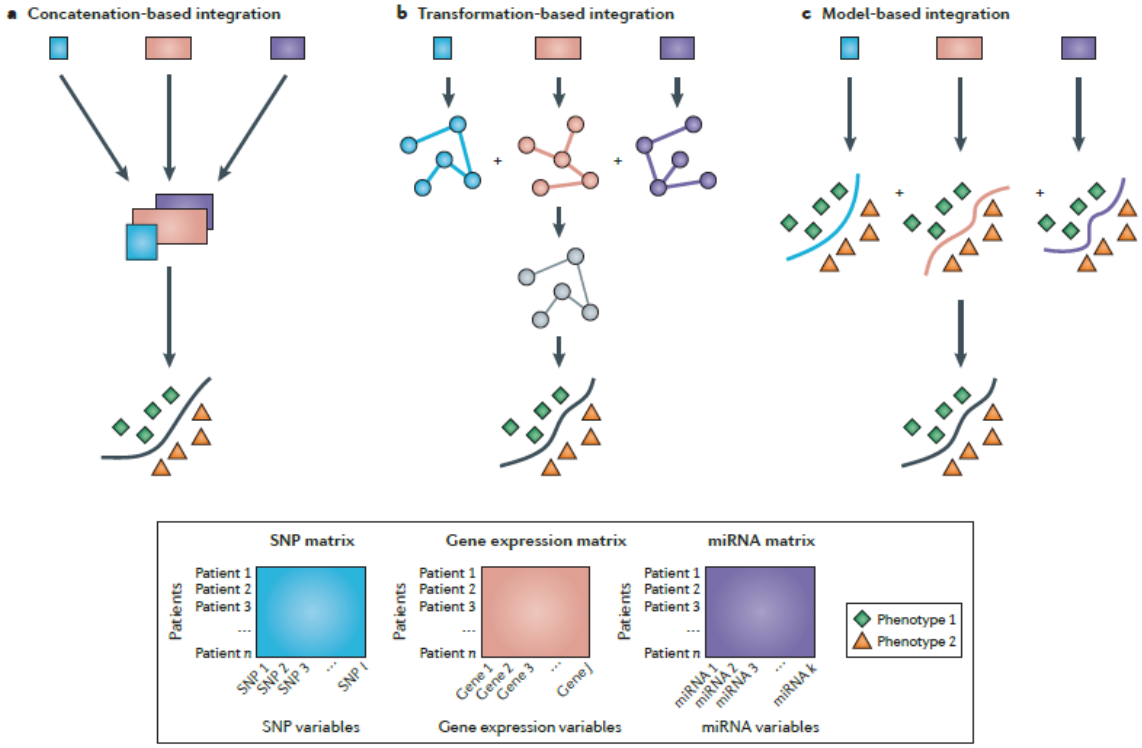

Fig. 1. Integration Strategies of meta-dimensional analysis using different data types adapted from (Ritchie et al., 2015): a) Early (concatenation based) integration b) Intermediate (transformation based) integration. c) Late (model based) integration .

integration strategies , with little or no information loss because of non-transformation of data.

The advantage of building models using a systems genomics approach involves combining multiple data sets of different omics types that can compensate for any missing or unreliable information provided by a single data type. Also genes or pathways that derive their information content from multiple genomic sources are likely to produce less false positives. By considering different levels of genomic, genetic and proteomic regulation, it is highly likely that a near complete biological model can be discovered. A summary of state-of-the art methods used for the inference of regulatory networks is presented in Table 1.

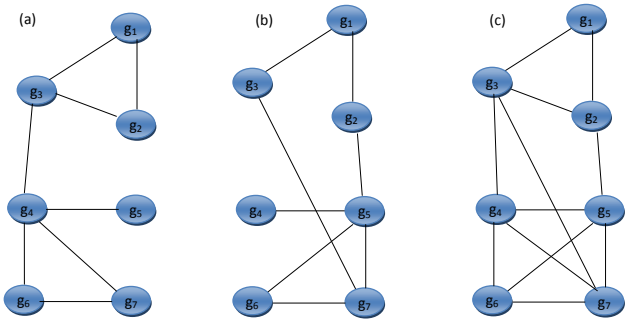

Fig. 2. a) Network1 b) Network2 c) Integrated gene interaction network

4 Network based Integration and Inference

Network based integration approaches provide the easiest and simplest way to integrate different types of network data into a single unified representation. It is flexible and can handle both homogeneous and heterogeneous datasets. In network data integration lexicon although data types are different, the homogeneous and heterogeneous data are characterized by the similarity/dissimilarity of network nodes. In homogeneous data integration, N networks, $G_i = (V, E_i)$, with similar vertices (genes, proteins etc.) across different networks and different edges E_i are combined by merging edges over the same set of vertices across all the individual networks i,e. $G_{int} = (V, U_{i=1}^{N} E_i)$ as shown in Fig. 2.

Similarity Network Fusion (SNF) (Wang et al., 2014) is a typical example of constructing merged networks by employing homogeneous network data integration principal. SNF that integrates expression data (mRNA, miRNA) and DNA methylation data, creates a network for each data type and then builds an integrated network from the fusion of individual homogeneous networks. The initial steps of the SNF build data set specific similarity matrices. These matrices are equivalent to similarity networks where nodes are genes/miRNA and the weighted edges represent pairwise node similarities. This is followed by a network fusion step that iteratively builds a single integrated network by employing message passing theory based techniques. The advantage of this iterative network construction is that low

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 April 2018 doi:10.20944/preprints201804.0352.v1

TABLE 1
An overview of integrative approaches that can be employed to infer Gene Regulatory Networks from multiple Omics datasets.Details regarding these approaches are described in separate sections

Integration Approach	Method Name	Omics Datatypes	Strengths / Limitations
Network Based approaches	SNF (Wang et al., 2014)	Gen expression,microRNA expression, DNA methylation	Simple to implement, less complex but may result in loss of information during the integration process.
	KeyPathwayMiner (Alcaraz et al., 2011)	Gene expression,DNA methylation, PPI data,Gen interaction Network	
	Mashup (Cho et al., 2016)	PPI network and Disease gene interaction network data.	
Regression Based approaches	Fused LASSO (Omranian et al., 2016)	Multiple gene expression datasets	Can handle large and heterogeneous datasets, model complexity is high owing to incorporating regularized constraints for sparsity.
	LASSO Regularized regression (Qin et al., 2014)	Gene expression,ChIP-ChIP/ChIP-seq	
Probability based approaches	Bayesian Networks (Isci et al., 2013)	Gene expression, eQTL (Expression of quantitative trait loci),TBFS ()Transcription factor binding sites) and PPI data.	BN models are ideal choices to work with heterogeneous data sets of modest scale, are computationally expensive and handle high dimensional and large scale data.
	Bayesian Variable Selection (Santra et al., 2013)	TBFS, PPI among TFs, mRNA expression profiles	
	FUSENET (Markov networks) (Žitnik & Zupan, 2015b)	Gene expression (RNA-seq),Mutation data (base substitution and short indels)	Can integrate data from non-identical data distributions, markov random fields exploit local independence property in the neighborhood to model co-regulatory effects while inferring GRNS, but cannot infer causal connections because of undirected edges in the graph. Computational complexity is high.
	GRACE (Markov random fields) (Banf & Rhee, 2017)	Gene expression data, TF-DNA binding motifs, Conserved promoter sequences, Physical PPI data	
Kernel Based approaches	Multiple kernel Learning model (Yamanishi et al., 2004)	Gene expression data, PPI data, Protein localization data, Phylogenetic profiles	Integrate multiple datasets of diverse nature, loss while building kernel matrices,can handle very large scale datasets and can identify informative datasets from the integrated set.
ANN based approaches	Multi-modal deep learning model (Liang et al., 2015)	Gene expression data, DNA methylation and Drug response data.	Offers component-based learning and hence reduced complexity, can learn from heterogeneous data views, are less scalable and face over fitting.

weight (weak) edges exclusive to all component networks are discarded in order to remove noise and high weight (strong) edges are incorporated into the final network.Also, low-weight (weak) edges common to all the networks being

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 April 2018 doi:10.20944/preprints201804.0352.v1

integrated are not discarded depending on how densely connected their network neighborhoods are across the networks. With such non-linearity SNF makes full use of the local network structure, combining common as well as complementary information of the structure across networks. Similarly, using this principle, multi omics data of cohort features can be integrated for the reconstruction of gene regulatory networks (GRNs), for example, given multiple RNA-seq/ micro-array datasets generated under different perturbations/ treatments/condition can be used to learn multiple Gene Regulatory Networks (GRN) from ech data set, we can then use SNF to fuse these individual GRNs into an integrated GRN. Similarly, GRNs inferred using algorithms, such as, CLR (Faith et al., 2007),ARACNE (A. Margolin et al., 2006), SIRENE (Mordelet & Vert, 2008) can also be fused by a method like SNF to achieve robust results, which implements the philosophy - 'the wisdom of crowds' (Marbach et al., 2012). A similar network fusion method has been proposed in KeyPathwayMinerWeb for pathway enrichment analysis using multi omics data (Alcaraz et al., 2011).

Another study by (Gade et al., 2011) builds a transcriptional regulatory network of prostate cancer from mRNA and miRNA expression profiles. Besides these two data sets, their method also incorporates prior knowledge about miRNA targets from MicroCosm (Enright et al., 2003) target database. Here the authors start by building a bipartite graph connecting miRNAs mi_j to their respective mRNA m_i targets. These relationships are established by computing Pearson correlation coefficients between the expression vectors of both mRNA and miRNA $p(m_i, mi_j)$ resulting in p-values form every mRNA-miRNA pair .

Besides the mRNA-miRNA pair, p-values $P_{i,j}^{pred}$ for target predictions from the MicroCosm database (Enright et al., 2003) are included as a secondary source of information. Incorporation of this predicted information strengthens the confidence of relationship between a mRNA and miRNA when some supportive information is provided from the available biological data. MicroCosm database generally stores mRNA-miRNA pairs whose p-values are below 0.05, in case the pair does not exist in database then the value is set to 1.

$$p_{i,j}^{cor} = P(H_0 : (m_i, mi_j) = 0) \quad \forall i \in 1, p_1, j \in 1, p_2 \tag{1}$$

P-values from both the sources (mRNAs & miRNAs) are combined using truncated product method (Zaykin et al., 2002) to yield a weight matrix $W_{i,j}$ viewed as adjacency matrix, describing relationships between mRNA and miRNA using a bipartite graph. The resultant graph W has edges from mRNA to miRNA and is interpreted as a directed graph. Although the study was not primarily designed to infer a regulatory relationship between mRNA and miRNA, but as a pre-requisite to identify prognostic signatures that are able to predict clinical outcome in prostate cancer. The results of the study clearly indicate an improvement in prediction performance with graph incorporated as regulatory information. Also introduction of likelihood-based boosting in (Binder & Schumacher, 2009) as a possibility to integrate gene interaction network into feature selection improves the predictive performance of the model.

The heterogeneous network integration method uses simple projection techniques to combine different types of vertices and edges across a number of networks. This type of network integration is usually done for association studies, for example, network projections of gene-gene interaction (GI) on a disease similarity network (DSN). Also projecting GI network on to a disease association network disturbs the GI network structure resulting in loss of information carried by the GI network. More sophisticated methods based on diffusion (information flow across network connections) overcome the projection based integration. The diffusion process explores the structure of each network and their corresponding mutual relationships in order to infer an integrated network. These approaches fall under the domain of network propagation methods and have been applied to a variety of biological problems, such as drug-target prediction and drug re-purposing (Cheng et al., 2012), drug-disease association (Huang et al., 2013) and gene-disease association (Guo et al., 2011). Another important study in this direction was carried out by (Cho et al., 2016) that led to the development of integrative framework called 'Mashup'.The Mashup framework exploits the topological representation of individual networks, analyzing each network separately by running a localized network diffusion algorithm, random walk with restart (Tong et al., 2006) to obtain the distribution of each node and their relevance to other nodes in the network. These feature vectors are then used as input by the framework to carry out mutiple tasks, such as gene interaction and gene function prediction etc.

More recently heterogeneous network integration approaches are being used for building Drug-target interaction networks in order to identify new indications for the old drugs (drug re-purposing). DTINet, a computational pipeline developed by (Luo et al., 2017)combines diverse information from heterogeneous networks by integrating a dimensionality reduction technique, DCA (diffusion component analysis) with a network diffusion algorithm RWR (random walk with restart) in order to obtain low dimensional feature vector representation of the network nodes, a process known as compact feature learning.This feature vector embeds relational properties (e.g. similarity), association information and topological information of each drug (or protein) node in the heterogeneous network.

5 Regression Based Approaches

Regression based techniques contribute a fair share to the range of methods developed so far to extract one-to-many relationships from gene expression profiles. These methods are employed to analyze high dimensional data sets, involving regularization to learn sparse models. Among the regularized regression methods, ridge regression and LASSO (least absolute shrinking and selection operator) (Tibshirani, 1996) have been effectively used to learn gene regulatory networks (Cai et al., 2013). Besides Computational biology, LASSO has also been successfully used in medical imaging, signal processing and economics (Hesterberg et al., 2008; Dasgupta et al., 2011; Yang et al., 2010). GENEI3 (Irrthum et al., 2010) is another inference algorithm that uses regression to infer regulatory networks. The inference problem for P

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 April 2018 doi:10.20944/preprints201804.0352.v1

genes is decomposed into P regression problems and tree-based ensemble methods using either Random forests or extra-trees are used to predict the expression pattern of target genes from the expression profile of all other input genes. A putative regulatory link between the input gene and its target is predicted based on the role played by the input gene in the prediction of target gene expression pattern. Links from all the genes are ranked and then aggregated to reconstruct a GRN.

In summary, although performance of regularized regression models in GRN inference compared to other approaches yielded better results, an increase in accuracy was observed in (Qin et al., 2014) when the authors integrated ChIP-ChIP/ChIP-seq data in the form of prior knowledge using $L_p(P < 1)$ i,e. $P_0, P_{1/2}$ regularization. In (Chartrand, 2007; Z. Xu et al., 2012), regularization models $L_p(P < 1)$ have been reported to perform optimally even on data with fewer samples and achieved higher sparsity and more accurate solutions.Since L_0 and $L_{1/2}$ regression models suffer from non-convexity, therefore iterative hard and half thresholding algorithms are used to solve these models (Fornasier & Rauhut, 2008). The advantage with these algorithms is that, they have a low computation cost and a fast convergence rate, both essential properties worth consideration while designing algorithms for genomic scale data.

5.1 L_p regularization models

A linear system representation of relationships between target genes and their regulatory proteins (TFs) can be written as:-

$$AX = B + \epsilon \tag{2}$$

Here $A \in R^{m \times r}$ represents a matrix containing gene expression values of candidate TFs , $B \in R^{m \times n}$ is the matrix containing gene expression values of target genes, $\epsilon \in R^{m \times n}$ is the error matrix and $X \in R^{r \times n}$ is the target-TF regulation matrix, m and r represent the matrix dimension and number of TFs respectively while n is the total number of target genes. The objective here is to minimize the difference between AX_j and B_j with few TFs, a sparse optimization problem, which can be written as :-

$$\begin{aligned} &\|AX_j - B_j\|_2 \\ &s.t \|X_j\|_0 \leq K \end{aligned} \tag{3}$$

Where $\|.\|$ denotes the Euclidean norm as $\|X_j\|_2 = \sqrt{\sum_{i=1}^{r} X_{i,j}^2}$ and $\|X_j\|_0$ denotes the number of non-zero elements in X_j and less $\|X_j\|_0$ means higher sparsity of X_j, an indicator of number of TFs found to be responsible for regulating the target gene j.

For solving such inference problems , the practical approach is to transform the sparse optimization problem into a regularization problem. For example , with B_j as the expression profile of a given gene j, the L_0 regularization model imposes a minimization constraint on the difference between AX_j and B_j and a maximization constraint on the sparsity of X_j as given by :

$$\min_{X_{j \in R^r}} \|AX_j - B_j\|_2^2 + \lambda \|X_j\|_0 \tag{4}$$

where regularization parameter $\lambda > 0$ provides a trade-off between sparsity and accuracy. L_0 regularization is very close to our initial probelm, but solving it to achieve a global optimal solution is NP-hard, therefore a relaxation of L_0, L_1 (LASSO) regularization is introduced to solve it:-

$$\min_{X_{j \in R^r}} \|AX_j - B_j\|_2^2 + \lambda \|X_j\|_1 \tag{5}$$

where $\|X_j\|_1 = \sum_{i=1}^{r} \|X_{i,j}\|$. In many instances , the solutions obtained using LASSO (L_1) model are less sparse than the L_0 regularization.

Recently , $L_{1/2}$ regularization has been proposed and has been shown to perform better than LASSO (L_1) regularization (Z. Xu et al., 2012).This model is described as below:-

$$\min_{X_{j \in R^r}} \|AX_j - B_j\|_2^2 + \lambda \|X_j\|_{1/2}^{1/2} \tag{6}$$

where $\|X_j\|_{1/2}^{1/2} = (\sum_{i=1}^{r} \|X_{i,j}\|)^2$

Qin et al. applied iterative thresholding algorithms to solve L_0, $L_{1/2}$ and L_1 regression models. For example, L_1 models are solved using iterative soft thresholding algorithms and L_0 and $L_{1/2}$ regression models use iterative hard and half thresholding algorithms for obtaining optimum solutions. These iterative algorithms are computational very efficient with fast convergence rate and effective tools for handling large scale sparse optimization problems. A detailed discussion on iterative thresholding algorithms can be found in (Fornasier & Rauhut, 2008).

Regularization parameter λ for all the thresholding algorithms is updated iteratively to maintain the sparsity of X_j^k. All the three regularization models were evaluated on mESC (mouse embryonic stem cell) data set. The initial experiments were performed on transcriptomic data alone and compared with TF-target datasets from high throughput (microarray/ChIP-X) and low throughput (literature curated). Both these TF-target datasets serve as gold standard against which the model accuracy is tested. The ROCs of all the models on both evaluation data were close to random with transcriptome only dataset. However, on the integration of the chip-X data into the models , it was observed that the performance of all the models improved drastically with L_0 and $L1/2$ models significantly outperforming L_1 regularization model. ROC of all three regression models $L_p(p = 0, 1/2, 1)$ are high with high throughput gold standard and low with low throughput gold standard. At an FPR of 0.05, the integrated models yield a TPR of 0.63, 0.59 and 0.07 for $L_0, L_{1/2}$ and L_1 models, while transcriptome only models achieve 0.03, 0.03 and 0.04 TPR. They demonstrated that $L_0, L_{1/2}$ models achieve higher sensitivity with integrated chip-X data compared to transcriptome only dataset.

Another regression based study that infers gene regulatory networks used Fused LASSO to integrate multiple transcriptomic datasets subject to different perturbations (Omranian et al., 2016). The approach formulated by authors combines L_2 norm (residual sum of squares) with L_1 (sum of absolute values of regression coefficients), shrinking coefficients and yielding sparsity as the coefficients approach towards zero. Unlike LASSO where the minimization of L_1 norm is imposed only on regression coefficients, the fused LASSO imposes this constraint also on the consecutive

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 April 2018 doi:10.20944/preprints201804.0352.v1

difference of the regression coefficients of corresponding regressors (Tibshirani et al., 2005).Further, the fused LASSO formulation has been extended here by incorporating information about regressors and response genes based on the similarity of their differential behavior. To ensure that the inferred networks are sparse, the fusion approach imposes a similarity constraint on the inferred relationships from each dataset with biologically meaningful evidence. Since this approach relies on similarity of the differential behavior, the model operates on P regressor genes and a single gene as a response gene, multiple gene expression datasets represented by X^i, obtained under K different experimental conditions along corresponding controls $X^{c,i}$ $(1 \leq i \leq k)$, and the expression profiles Y^i for response genes along corresponding response control profiles $Y^{c,i}$ over k conditions $(1 \leq i \leq k)$ are used to obtain condition specific differential behavior of regressors and response genes. The differential behavior thus obtained is used to calculate the weight matrix denoted by $W^i_{P\times P}$ from the probability of differentially expressed regressor $Pr^{i,t}_j$ and response genes $Pr^{i,t}_Y$ which capture the similarities between each of the regressor and response gene at time point t is given by:

$$W^i_{Y,j} = \frac{1}{N}\sum_{i=1}^{N} |Pr^{i,t}_Y - Pr^{i,t}_j| \tag{7}$$

5.2 Regression model for fused LASSO

For efficient inference of GRNs, the regression model formulated in equation (8) is based on three fundamental criteria:-

1) Obtained regression should be sparse to ensure less false positives and high likelihood of detecting direct relationships. Biologically , the expression level of a few TF coding genes should explain the expression level of their presumed targets.
2) For effective explanation of the analyzed condition, the model should infer the regulatory network from all k datasets simultaneously.
3) Genes exhibiting similar differential behavior are proffered while assigning direct regulatory links.

While LASSO captures the first criterion, the second criterion of inferring regulatory networks from all datasets can be achieved by imposing the LASSO penalty simultaneously on all the integrated and transformed datasets. k transcriptomic datasets are clubbed together to form a single block matrix X denoting expression profiles for P genes , while as vector Y is response expression profiles over k different conditions The reconstructed networks should be evaluated for proximity in terms of their interaction strengths given by regression coefficients that correspond to links among the response Y and regressor genes X in the k networks.This vector is represented as β.

For reconstructed networks to be in close proximity to each other, the final model includes a fusion term $\|\beta'' - \beta'\|$, where $\beta' = [\beta^1, \beta^2, ..., \beta^{k-1}]^T$ and $\beta'' = [\beta^2, \beta^3, ..., \beta^k]^T$ .This term is included to impose a minimization constraint for the sum of absolute differences among the estimated regressor coefficients over multiple datasets. Within the fused LASSO formulation the inclusion of W^i $(1 \leq i \leq k)$ implies the similarity of the differential behavior between response and the regressor genes. This matrix multiplied with the regression coefficients ensures that the regressors having high explanatory powers and associated with non-zero coefficients and are less penalized over multiple datasets.Therefore the expression $\|W\beta\|_1$ is added as an additional penalty in the proposed regression model. For reconstruction of gene regulatory network interactions an final model over k given datasets is given as:-

$$\hat{\beta} = arg\min_{\beta} \|Y - X\beta\|_2^2 + \lambda_1\|W\beta\|_1 + \lambda_2\|\beta'' - \beta'\|_1 \tag{8}$$

Here regressors whose regression coefficients shrink towards zero are the potent regulators of the response gene Y. A comparative analysis of this approach with other state of the art methods viz, global silencing (Barzel & Barabási, 2013), network deconvolution (Feizi et al., 2013), Gaussian Graphical Models (GGM)(Schäfer & Strimmer, 2004; Krämer et al., 2009; Chun et al., 2013), mutual information (ARACNE and CLR) (A. A. Margolin et al., 2006; Faith et al., 2007; Zoppoli et al., 2010) (Meyer et al., 2007), Bayesian Networks (Friedman et al., 2000), the GENIE3 (Irrthum et al., 2010) approach and other regularization based models is performed using network sparsity and removal of the indirect relationships as dual criteria for model evaluation. The GGM,ARACNE and CLR perform poorly for accurate inference of regulatory networks, on the contrary, fused LASSO approach in addition to retrieving the regulatory links efficiently also predicted the type of regulation (activation/repression), therefore outperforming most of the contending methods on the set criteria.

6 Probability Based Approaches

6.1 Bayesian Networks

Bayesian networks (BNs) or belief networks as they are often called belong to the family of probabilistic graphical models used to represent knowledge about an uncertain domain. These networks are represented using directed acyclic graphs (DAGs), wherein each node represents a random variable and the edges between the nodes represent the causal connections / probabilistic dependencies among the corresponding random variables (Ben-Gal et al., 2007). These conditional dependencies are estimated by using statistical and computational methods .Therefore BNs combine principles from graph theory, probability and statistics.

Bayesian networks are recognized as important tools in data driven biological science research and play a pivotal role in performing several biological tasks such as inferring cellular networks (N. Friedman, 2004), data integration (Troyanskaya et al., 2003; Santra, 2014), classification (Bradford et al., 2006) and genetic data analysis (Beaumont & Rannala, 2004).

The graphical representation and use of probabilistic formalisms make BNs suitable for combining data and domain knowledge, express causal relationships, learn from incomplete datasets and provides a natural setting for modeling stochastic nature of biological systems and its measurements. The graph in Fig 3(a) provides an example a BN describing a gene regulatory network. This network represents a joint probability distribution (JPD) over a set

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 April 2018 doi:10.20944/preprints201804.0352.v1

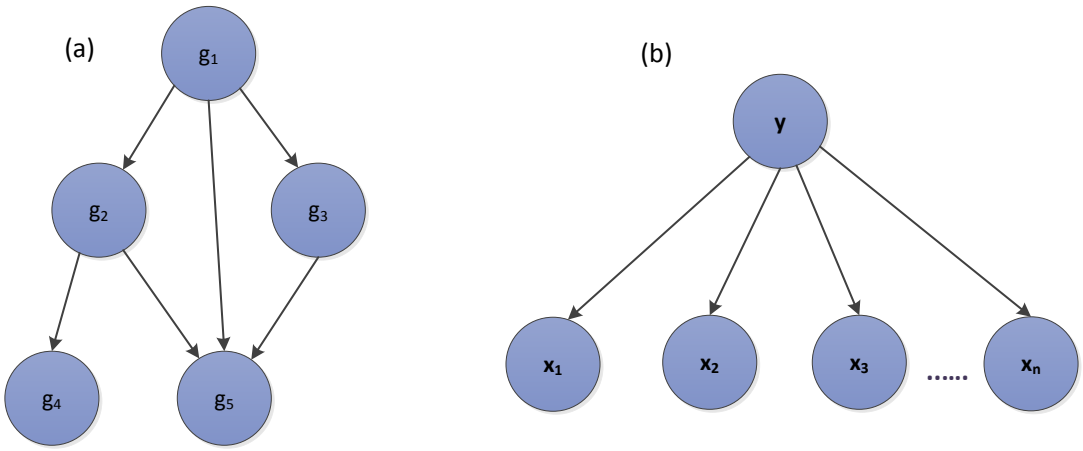

Fig. 3. (a) A graphical illustration of Modeling of Gene Regulatory Network using BN. Vertices denote the genes whereas the directed edges represent the regulatory relations between the genes.Here Gene g_1 regulates the expression of g_2,g_3 and g_5, genes g_2 and g_3 also regulate g_5, whereas g_4 is only regulated by g_2.(b) Graphical depiction of a *naive BN* with class node y acting as parent to independent nodes $x_1, x_2, x_3....x_n$.

of random variables V, depicted as an annotated acyclic graph. The pair B= (G, Θ), where G is the DAG whose nodes represent genes as random variables $g_1, g_2, g3, ..., g_n$ and whose edges represent direct dependencies between the genes. The other component of the network Θ is used to represent a set of parameters. Th elements of this set are the parameter $\theta x_i|\pi_i = P_B(x_i|\pi_i)$ for every realization x_i of g_i conditioned on π_i , set of parents of X_i in G.

The DAG structure embeds conditional independence relationships, i.e. given its parents in G, a node g_i does not depend on its non-descendants The JPD defined by B over V can be written as:-

$$P_B(g_1, g_2, g_3, ...g_n) = \prod_{i=1}^{n} P_B(g_i|\pi_i) \tag{9}$$

Bayesian Networks have been used to build suitable frameworks for modeling and integration of different types of biological data. Network inference from disparate data is one of the biggest challenges in the field of systems biology, sparse network construction containing important gene associations (strength of this relationship within the network is represented by conditional probabilities in a joint probability distribution). Constructing a BN that provides an effective description of our data involves two important steps: structure learning and parameter learning. (Needham et al., 2007; Ben-Gal et al., 2007). Because the possible number of BN structure grows super exponentially with the number of nodes , the problem of finding a BN with best description of data is Np-complete, therefore heuristic methods have been employed for solving such problems (Chickering, 1996).

Studies employing bayesian networks for integrating gene expression data with other genomic datasets in a heterogeneous data integration framework have widely been reported in the literature.For example, (Zhu et al., 2008) built a probabilistic causal yeast regulatory network by combining gene expression data(GExD), transcription factor binding site (TFBS), eQTL(Expression of quantitative trait loci) and PPI data using bayesian learning .While testing the performance of constructed BN, they demonstrated an integrated BN model possesses higher predictive power than a BN model derived from single data type (gene expression data).A similar approach to construct a probabilistic model for GRN inference from brain tissue of late-onset Alzheimer's has been adopted by (Zhang et al., 2013) by integrating patient data .

Earlier studies of data integration using BNs has been reported in (Gevaert et al., 2006) to predict the outcome of breast cancer from heterogeneous genomic datasets. They constructed the BN network in order to classify the patients into different prognosis groups (good/poor).They increased the performance of the BN by subjecting it to all the three integration strategies as described above and observed that intermediate strategy was the one with improved results.Another study from (Van Vliet et al., 2012) using multiple classifiers with varying complexity also concluded in favour of intermediate integration.

Most of the data integration studies for heterogeneous biological data have used Naive Bayes (simple BN) for constructing integrated networks from genomic datasets, such as gene-gene association network (Lee et al., 2004; Linghu et al., 2009; Franceschini et al., 2012; Jansen et al., 2003). Here all independent nodes representing heterogeneous data sources are treated as child nodes of a class node, such simplicity in Bayesian Network structures enable efficient learning and much faster inference approaches. For example, gene-gene association studies use naive BN structures for prediction purposes where class nodes represent a set of proteins/genes (interacting/non-interacting) while independent nodes represent multiple data sources.A graphical depiction of naive baye's is shown in Fig 3(b).

Recently, (Santra, 2014) proposed a BN method based on bayesian variable selection for GRN inference that incorporates TFBS,PPI among TFs and mRNA expression profiles subject to genetic perturbations. The approach uses a linear model based on an idea, that when a network is perturbed , there is a widespread change in the expression level of genes

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 April 2018 doi:10.20944/preprints201804.0352.v1

as the effect of perturbation rapidly propagates through the entire network.This premise has been shown to be true through the works of (Rogers & Girolami, 2005; Lo et al., 2012). They have shown that the response expression x^i of target genes (g_i) are linearly dependent upon the response expression $X^i$of their direct regulators g^i against a set of perturbations (n_p).

$$x^i = {X^i}^T \beta^i \tag{10}$$

Where β^i are the linear coefficients representing the type and strength of interaction between genes (g_i) and its direct regulators (g^i). Since no perturbation experiments are error free, therefore measurements of the expression levels are always accompanied by some sort of noise called '"residuals'". To accommodate these residuals in the linear model, the above equation now becomes:-

$$x^i = {X^i}^T \beta^i + \epsilon^i \tag{11}$$

where ϵ^i is the residual error and a linear combination of these variables (ϵ_i) denotes total noise of individual measurement errors (Kariya & Kurata, 2004) that the perturbation responses of the genes (g_i) and their regulators (g^i) are associated with.

In order to find the direct regulators (g^i) of the gene (g_i), the requirement is to calculate β^i in least square sense. In this formulation , the task of identifying the direct regulators g^i of target genes g_i is determined by the values of β^i. Elements ($\beta_{ik}, k = 1, ..., np$) of β^i whose values are > 0 are chosen as significant direct interactions and therefore the direct regulators of genes g_i are corresponding genes g_k.Performing perturbation experiments on a large scale is neither feasible nor possible under most of the circumstances. Hence in such situations reconstructing a full network is not possible and the problem can be resolved by Bayesian variable selection (BVS) algorithm (Santra et al., 2013).

In BVS formulation an adjacency matrix A is used to represent a GRN topology, elements of the matrix with non-zero values A_{ik} denote a direct regulatory relationship between genes g_i and $g_k, (k \neq i)$, on the other hand elements with zero values represent no direct regulation. There is a close relationship between elements in matrix A and the elements of matrix interaction strength β.Given two genes (g_i, g_k), absence of direct interactions ($A_{ik} = 0, k \neq i$) between them implies a zero interaction strength ($\beta_{ik} = 0, k \neq i$), in β and vice versa. Hence, to locate direct regulators of a gene g_i within the binary adjacency matrix is equivalent to locating elements of A in the ith row of the adjacency matrix whose corresponding strength of interaction in β is not zero. Since it is computationally inefficient to iterate across all the combinations of A^i, BVS algorithm adopts a Bayesian approach to avoid the iteration process. The Bayesian algorithms learns in a more natural way, the knowledge updates for some events is carried out only as the new information becomes available.In bayesian lexicon, prior distributions are used to encode the information about certain events and hence assigning a prior probability to each possible outcome of the event (Heckerman, 1998).

For GRN inference the prior distribution ($P(A^i)$) encodes the prior knowledge about gene g_i and its direct regulators g^i.In (Santra et al., 2013), the model does not use any prior knowledge and the approach favors sparsity, penalizing models with too many regulators. Here, a more informative prior distribution A^i is injected into the model using integrated TFBS data and PPI data among TFs.

The application of BVS algorithm was demonstrated by reconstructing a liver specific GRN by using perturbation data. Different prior distributions were tested for model evaluation:-

1) Prior distribution of A^i was formulated without using any prior knowledge from an external data source.
2) Although no prior knowledge is incorporated , the prior distribution was tailored to favor sparse regulatory programs.
3) A prior distribution of A^i was formulated based on the predicted information about direct regulatory interactions from publicly available TFBS data.
4) Both direct and indirect regulatory interactions were used to construct a prior distribution of A^i from TFBS data and PPI data between TFs.

To estimate the posterior probability of interactions, the BVS algorithm uses Markov Chain Monte Carlo (MCMC) sampling strategy. The probabilities thus obtained represent the confidences on each interaction based on the perturbation responses. An interaction is considered to be true if the probability is above a certain threshold (P_{th}) and absent in case the probability is below or equal to the threshold (P_{th}).GRNs reconstructed from the true interactions are then compared with the gold standard.

Results of BVS formulation were compared with LASSO based algorithms using different prior settings.The first case does not use any prior and the regularization parameter λ_1 is set to 0.2. In case of second and third setting a prior distribution of TFBS and TFBS+PPI with regularization parameters set to λ_1 =0.2 and λ_2=0.8 respectively. Evaluation with ROC and PRE curves suggests that with the inclusion prior information, the performance of BVS algorithm improved significantly. Prior knowledge about direct and indirect regulatory interaction from TFBS integrated with PPI data between TFs were observed to improve predictiveness of regulatory relationships than the TFBS data alone. More over, the proposed BVS formulation outperformed all LASSO models in all circumstances.

Modeling using BNs provides an efficient and suitable framework for integration of multiple datasets (homogeneous/heterogeneous). They are able to capture noisy conditional dependencies between data variables and the construction of CPDs makes them suitable for a variety network inference problems in bioinformatics.However, there are certain limitations of BNs as well:

i Although some important associations are captured using the sparse network representation, some of the associations may be missing in the final network.
ii Network loops which in many biological networks represent important control mechanisms can not be inferred using BNs because of their acyclic nature.
iii As the number of nodes in the model increase, BN structure grows exponentially with a huge computational cost, therefore making BNs less attractive for large scale studies (e.g. eukaryotic genomes).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 April 2018 doi:10.20944/preprints201804.0352.v1

6.2 Network inference through Markov networks

Graphical probabilistic models with undirected edges are called Markov networks.These networks provide a simple definition of independence between any two nodes and are used to represent conditional dependence relationships between genes. Mechanisms within the biological cells are governed by complex interactions among various genes and it would be of great interest for biologists to understand the conditional dependencies between these genes. It is possible to uncover these dependencies from the experimental data characterizing genes and gene products and represent them in the form of a gene network. Markov networks present a popular statistical approach to model such conditional independence relationships in high-dimensional genomics data. In particular the absence of a link between genes s and t is indicative of independence relationship between s and t, considering the immediate neighborhood of s. From (Murphy, 2012) with such properties we can conclude that, two nodes (genes) without any direct links between them are conditionally independent of each other given the rest of the genes in their vicinity. Such a property of conditional independence (Markov) allow existence of a rich set of dependence relationships between nodes and therefore, we can uncover complex relations among the nodes of a Markov network (Allen & Liu, 2013).

The problem of learning network structures using Markov networks have witnessed wide acceptance from the fields of computer vision (S. Z. Li, 1994), social networks, image and signal processing (C. Wang et al., 2013; Metzler & Croft, 2005) and genomics (Wei & Li, 2007). The applications of undirected graphical models (Markov networks) in bioinformatics include construction signaling network for cancer from proteomic data (Mukherjee & Speed, 2008), genetic interaction network reconstruction from integration of multiple datasets (Isci et al., 2013) and in recent times various inference algorithms based on Markov networks using NGS data have been developed (Allen & Liu, 2013; Gallopin et al., 2013).

Žitnik and Zupan developed *FUSENET* (Žitnik & Zupan, 2015b) as a GRN inference algorithm based on Markov network formulation that integrates multiple non-identically distributed heterogeneous datasets. The input to FUSENET is a dataset colletion with each dataset containing a set of gene expression profiles. Gene expression measurements are taken from RNA-seq experiments (discrete, non-negative) and are combined with mutation data including CNV, SNP, short indels and multiple substitutions. While the RNA-seq data follows Poisson or negative binomial distribution, the data from mutation datasets are modeled either using categorical or multinomial distributions . A prominent feature in FUSENET is the latent factor representation of model parameters. Employing latent factors and the flexibility to share these factors across different datasets makes it easy for the algorithm to perform network inference tasks from multiple datasets considered simaltaneously and arising from different probability distributions. The network model offered by FUSENET offers a collective approach to inference, whereby the network estimation is performed jointly from non identical data distributions.The FUSENET model is illustrated as below:- Given a collection

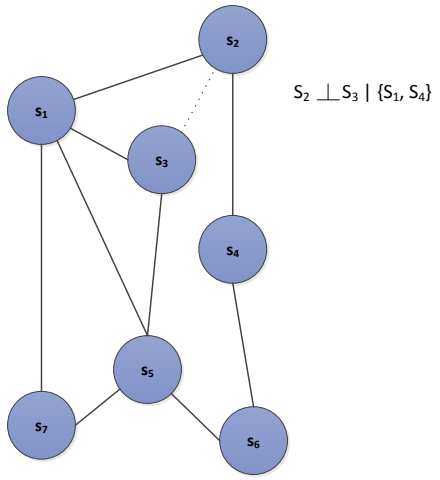

Fig. 4. A Markov network modeled using FUSENET, that infers dependencies among any two genes conditioned on their neighbors. Here dotted line denotes an absence of edge between s_2 and s_3 implying that s_2 acts indepent of s_3 given its neighbours s_1 and s_4. The $\perp$ symbol stand for conditional independence.

$\mathcal{D}$ of n observations, $\mathcal{D} = \{x^1, x^2, ..., x^n\}$ where x^i is a p-dimensional vector drawn from a specific probability distribution. This distribution is associated with a graph $G = (V, E*)$ with parameters$\{\theta_c^*, c \in \mathcal{C}\}$. Because Graph G embeds the Markov independence properties among the nodes, therefore the task of reconstructing the topology of G is to infer an edge E^* corresponding to the probability distribution where from the initial observations in D were drawn.Therefore set of edges as a function of parameters can be written as:-

$$E^* = \{(s,t) \in V \mathrm{x} V : \exists \quad clique \quad c \in \mathcal{C} : \{s,t\} \subset c \wedge \{\theta_c^* \neq 0\}\} \tag{12}$$

Learning the structure of network using FUSENET therefore aims at estimating the weights $\{\hat{\theta}, c \in \mathcal{C}\}$ whose values should possibly close to the true but otherwise unknown parameters $\{\theta_c^*, c \in \mathcal{C}\}$.For the final model to take shape, authors chose a pairwise Markov network with joint probability distribution model having cliques of size at most two:

$$P(X) \propto exp\Big(\underbrace{\sum_{s \in V} \theta_s^* B(X_s)}_{setofnodes} + \underbrace{\sum_{(s,t) \in VxV} \theta_{st}^* B(X_s) B(X_t)}_{setofedges} + \sum_{s \in V} C(X_s) \Big) \tag{13}$$

with entries $\theta_{st}^* \neq 0$ if $t \in N(s)$ and $\theta_{st}^* = 0$ if $t \notin N(s)$.The overall Markov network structure is then estimated by adopting neighborhood estimation approaches following the works of (Ravikumar et al., 2010; Jalali et al., 2011; Allen & Liu, 2013), the estimate for the entire network $\hat{E}$ is then obtained using :

$$\hat{E} = \bigcup_{s \in V, t \in \hat{N}(s)} \Big\{(s,t)\Big\} \tag{14}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 April 2018 doi:10.20944/preprints201804.0352.v1

where (s,t) represent an edge between nodes s and t and $\hat{N}(s) = \{t \in V\backslash\{s\} : \hat{\theta}_{st} \neq 0\}$ is the estimated neighborhood of node s. In order to assess the performance of FUSENET, several competing inference algorithms were considered. The FUSENET model was compared with both Gaussian and non-Gaussian models including Graphical LASSO (J. Friedman et al., 2008), Local Poisson Graphical Model (LPGM) (Allen & Liu, 2013) and Multinomial Morkov Graphical Model (Mult-GM) (Jalali et al., 2011) for performance evaluation. The considered inference methods along with FUSENET were applied to Poison-distributed simulated data.Four network types viz. random, hub, small world and scale free were generated , the last three being characteristic of many real biological networks. Datasets having $P = 100$ (nodes) variables and with $n = 200$ observations were generated. ROC curves for all the simulated networks using comparative and proposed methods were reported. It was concluded that FUSENET outperforms Gaussian based competitors (GLASSO, Log-GLASSO) as well as methods modeled using Poisson data (LPGM,Mult.GM).This shows an overall and consistent performance of FUSENET across all considered networks and more than one data distributions.For in depth understanding of the model refer(Žitnik & Zupan, 2015b)

Since GRN inference for most of the eukaryotic organisms remains a challenging task, another important study by (Banf & Rhee, 2017) using Markov random fields exploits the prior biological knowledge and heterogeneous data integration to build high confidence network prediction models.The study uses two datasets to evaluate the prediction accuracy of the proposed algorithm. One of the findings on A.Thaliana dataset is covered here. For reconstruction of a GRN, data from three different sources were integrated.(i) A conserved non-coding sequence 2000bp promoter region of 17610 genes,(ii) DNA binding predictions within these sequences for 120 TFs and (iii) gene expression dataset of A.Thaliana development, comprising RNA samples from 83 tissues.This data was used to derive a condition specific co-expression network.A variance based filtering was applied to remove genes that exhibit little variance across all tissues and development stages.

For network inference, a highly robust and scalable tree based regression which decomposes the network inference into separate regression problems for each target gene, this model calculates an important measure for each predictor, which is used as an indicator for a link to be present between the regulators and target gene. Given a large number of regulators in A.Thaliana , Banf & Rhee (2017) computed ran tree based regressions with more than 5000 decision trees for each target gene to ensure that regulators are selected multiple times during bootstrap aggregation so as to provide stable prediction for each target gene. For all the evaluations, they retained all the predictions beyond $95th$ percentile of the distribution.

Given a set of regulatory links l, there is a concept of meta gene regulatory networks that describes connections between two links l and l' i.e., $l \leftrightarrow l'$. A connection between two links l, l' is based on the co-regulation principle i.e., two different target genes g and g' are controlled by the same regulator r.

$$l \leftrightarrow l' = r \rightarrow \{g, g'\} \tag{15}$$

The weight of such a connection is based on a distance metric that combines two measures which are assumed to reflect co-regulation.Given this meta gene regulatory network, modules of connected regulatory links can be extracted from the meta network.Each module represents a group of large genes g that are pair-wise co-regulated by a specific TF r. Individual links that are not connected to any other link are also retained as individual (single link) modules.The co-regulatory effects within each module are modeled using Markov random fields $G = (V, E)$, which implements a local independence assumption referred to as Markov property (Schwaller, 2015). This property imposes a node to be independent of any other nodes given all its in direct neighbors, i.e.:

$$\forall i \in V, X_i \perp X_V -_i | X_{N_i} \tag{16}$$

Where $N_i, j|\{i,j\} \in E$ denotes the set of immediate neighbors of node V_i.An important notion in the model is a clique. It is defined as fully connected subset of nodes within the graph, which is considered as maximal if it is not connected within any other larger clique.In order to retain the links in the final network , the model favors links with high weights as well as strongly connected pairs of regulatory links to be in the same state.

7 Kernel Based Approaches

Kernel methods represent a mathematical framework which embeds data points (genes, proteins, miRNA etc) from input space I to feature space $\mathcal{F}$ by employing a kernel function. Genomic datasets viz., mRNA expression levels from RNA-seq, miRNA expression profile from miRNA-seq and TF-gene regulation matrix obtained from different databases such as ENCODE (E. P. Consortium et al., 2004), TRED (Jiang et al., 2007), TRRUST (Han et al., 2015) etc. comprise heterogeneous datasets that serve as the building blocks of gene regulatory networks which can be fused together using kernel methods.

Each dataset is transformed into a symmetric positive semi definite kernel matrix by means of a kernel function, that is a real valued $k(x_1, x_2)$ satisfying $k(x_1, x_2)$= $k(x_2, x_1)$ for any two objects x_1 and x_2 and positive semi-definite i.e., to say $\sum_{i=1}^{n} a_i a_j \; k(x_i, x_j) \geq 0$ for any integer n, set of objects $(n = x_1.....x_n)$ and any set of real numbers $(a_1.....a_n)$ (Charpiat, 2015). Kernel functions provide a coherent representation and a mathematical framework for the input data (genes, TFs, miRNA etc.) and represent the object features via their pairwise similarity values comprising the n x n kernel matrix, defined as.

$$\mathbf{K} = \begin{pmatrix} k(x_1,x_1) & k(x_1,x_2) & .. & k(x_1,x_n) \\ k(x_2,x_1) & k(x_2,x_2) & .. & k(x_2,x_n) \\ . & . & .. & . \\ . & . & .. & . \\ k(x_n,x_1) & k(x_n,x_2) & .. & k(x_n,x_n) \end{pmatrix} \tag{17}$$

Kernel methods offer a modular approach to pattern analysis (Shawe-Taylor & Cristianini, 2004). An algorithmic procedure is devised together with a kernel function

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 April 2018 doi:10.20944/preprints201804.0352.v1

that performs an inner product on the inputs in a feature space. This algorithm is more generic and can work for any kernel and hence for any data domain. The kernel part is data specific that offers an elegant and flexible approach to design learning systems, that can easily operate in very high dimensional space. It is a modular framework, where modules are combined together to obtain complex learning systems. Some examples of commonly used kernels are (Shawe-Taylor & Cristianini, 2004) :

Linear kernel

$$\Phi(x_1)^T \Phi(x_2) = x_1^T x_2 = k(x_1, x_2) \tag{18}$$

Polynomial kernel

$$k(x_1, x_2) = \langle x_1, x_2 \rangle^d \tag{19}$$

Gaussian kernel

$$k(x_1, x_2) = exp\left(-\frac{\|x_1 - x_2\|}{2 - \sigma^2}\right) \tag{20}$$

For suitable values of d (degree of the polynimial kernel) and σ (spread parameterof gaussian kernel), the similarity measure $k(x_1, x_2)$ between x_1 and x_2 is always positive with is maximum at $x_1 = x_2$. All points x have same unit norm (since $k(x_1, x_2) = 1 \; \forall \; x$) suggesting that images of all points in x lie on the unit sphere in $\mathcal{H}$ (Schölkopf & Smola, 2002).

In order to handle genomic datasets like DNA sequences , protein sequences and network data such as, protein-protein interactions, molecular and cellular networks, customized kernel functions have been proposed for building similarity matrices. For example,kernel functions for computing the similarity between two genes/proteins are spectrum kernel (C. Leslie et al., 2001; C. S. Leslie et al., 2004) , pairwise kernel (Ben-Hur & Noble, 2005) , motif kernels (Ben-Hur & Brutlag, 2003) for defining similarity between TF-DNA binding motifs (Ben-Hur & Brutlag, 2003), and pfam kernel (Gomez et al., 2003). Network data represented using adjacency matrices is transformed into a similarity matrix using diffusion kernel.For more in depth coverage on range of kernels devised for handling biological data refer to (Schölkopf et al., 2004).

Besides the sequence data that normally occurs in string format, network data from molecular networks and interaction networks, pathways and Gene Ontology (GO) associations are frequently used in kernel based integration studies. These types of datasets can be handled using graph with different scenarios. In the first case,the input varibales can be treated as nodes within a graph,e.g. proteins in PPI networks or genes in case of GRNs and in the second scenario proteins can be represented using graphs, e.g. protein representation using phylogenetic trees. Computational biologists have developed kernels for both scenarios. For example, kernels built from phylogenetic profiles modeled as trees (Vert, 2002), graphical representation of small molecules (Kashima et al., 2004) and graphs modeling protein structures (Borgwardt et al., 2005).

7.1 Multiple Kernel Learning Model

Multiple kernel learning (MKL) is a paradigm shift from traditional single feature based learning and offers an advantage of combining multiple features of objects such as genes, proteins, metabolites etc., as different kernels (Sonnenburg et al., 2006). This information can be fed as an ensemble into an MKL learning algorithm as a combined kernel matrix for classification or regression tasks on unknown data. The basic algebraic operations of addition, multiplication and exponentiation when performed in combining multiple kernel matrices preserves the positive semi-definite property and enable the use of powerful kernel algebra. A new kernel can be defined using k_1 and k_2 with their corresponding embeddings $\Phi_1(x)$ and $\Phi_2(x)$. This resultant kernel is

$$K = k_1 + k_2 \tag{21}$$

with the new induced embedding

$$\Phi_x = [\Phi_1(x), \Phi_2(x)] \tag{22}$$

Given a kernel set $K = \{k_1, k_2, ..., k_m\}$, an affine combination of m parametrized kernels can be formed as given below: -

$$K = \sum_{i=1}^{m} \mu_i k_i \tag{23}$$

subject to the constraint that μ_i (weights) are positive i,e. $\mu_i \geq 0, i = 1.....m$ A kernel based statistical classifier such as SVM induces a margin in feature space, separating the two classes using a linear discriminant.In order to find this linear discriminant, an optimization problem needs to be solved, known as a quadratic program (QP). A Quadratic program is a form of convex optimization problem, which are easily solvable.

Kernel based integration methods were first proposed in (Lanckriet et al., 2004), wherein a 1-norm soft margin SVM is trained for a classification problem separating membrane proteins from ribosomal proteins. They combined heterogeneous biological datasets viz. PPI,amino acid sequences and gene expression data, characterizing different proteins by transforming them into multiple positive semidefinite kernel matrices using different kernel functions.Their findings reveal an improved classifier performance when all datasets are integrated as a unit compared to testing the classifier on individual datasets. In an earlier study on function prediction for baker's yeast proteins (Lanckriet et al., 2003b) trained an SVM classifier with multiple datasets and achieved an improved performance over the a classifier trained using single data type.

In another study for network inference using kernel data integration (Yamanishi et al., 2004) four different datasets viz Gene expression data , protein interaction data , protein localization data and data from phylogenetic profiles were transformed into different kernel matrices. Datasets that assign a vector to protein/gene viz. gene expression, protein localization and data from phylogenetic profiles can use Gaussian, ploynomial or linear kernels as tranformation functions. Graph datasets are kernalized using diffusion kernel (Kondor & Lafferty, 2002) defined as $k = exp(\beta H)$. Here H is the graph laplacian obtained by subtracting $(H = A - D)$ diagonal matrix D of the graph from its adjacency matrix A and $\beta > 0$ is the parameter. All genomic datasets were transformed into different types of kernels.The gold standard protein network and the noisy protein interaction datasets were represented by a diffusion kernel with parameter = 1. Gene expression data were

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 April 2018 doi:10.20944/preprints201804.0352.v1

kernalized using Gaussian RBF kernel with σ = 5. A linear kernel function was used to transform both localization data and data from phylogenetic profiles.

The study (Yamanishi et al., 2004) compared both unsupervised and supervised inference methods on single and integrated datasets.For unsupervised inference both spectral as well as direct approaches tested either on a kernel derived from single genomic dataset or on a combined kernel from multiple genomic datasets. Using the spectral approach, a feature space is defined by selection initial L=50 principal components. To assess the accuracy of both the methods in terms of their capacity to recover existing interactions, the inferred network is compared with a gold standard protein network. The overall accuracy reported for both the methods seems to capture little information from the gold standard. Although results generated using spectral approach with integrated datasets displayed improved accuracy, nevertheless an increased rate of false positives was observed for any true positive rate.

Again with number of principal components L = 50 , that define the features space, various combinations of kernels from multiple datasets were tested against a gold standard kernel. To evaluate the performance of the supervised approach, a $k = 10$ fold cross-validation procedure was performed. A graph is gradually built by selecting true positives and plotting them as a function of false positives. Contrary to the direct and spectral approaches, the supervised approach seems to capture most of the information about the gold standard protein network and make interesting predictions. Datasets such as gene expression and phylogenetic profiles seem to make contribution with equal quantum of information, followed by noisy PPI and localization datasets. Applying supervised approach to integrated datasets seems to produce overall best results, therefore highlighting the importance of guided network inference from integrated prior biological knowledge. In another study (Ben-Hur & Noble, 2005) applied kernel methods to PPI studies and proposed a pair-wise kernel between two pairs of proteins in order to construct a similarity matrix. They represented a protein sequence pair (X_1, X_2) in terms of the domain or motif pairs that appear in it, the pair-wise kernel is described as:

$$K((X_1, X_2), (X_1', X_2')) = K'(x_{12}, x_{12}') \tag{24}$$

where x_{12} is the pairwise representation of the sequence pair (X_1, X_2), and $K'(,)$ is kernel representation that operates on vector data. This pair-wise kernel is based on three sequence kernel , spectrum kernel , motif and pfam kernels. The ROC scores reported for these kernels are $0.76, 0.78$ and 0.81 respectively. This experiment was further extended to explore the effect of adding kernels from non-sequence data, such as gene ontology annotations, homology scores and Mutual clustering coefficient (MCC) derived from protein interactions computed in each cross validation fold.Integrating these non-sequence features with the pairwise kernel resulted in better performance than any method by itself as revealed by their ROC scores. Furthermore, tuning the the soft-margin parameter C of the SVM according to the reliability of the interactions provided another significant boost to the performance, yielding an ROC score of 0.98, at an FPR (false positive rate) of 1%. From the comparative analysis it was observed that gain in performance of the model is due to the contribution coming from the prior biological knowledge incorporated in the form of GO-process kernel and other features of biological relevance.

8 ANN BASED APPROACHES

Artificial neural network (ANN) is an information processing computing paradigm inspired by structural and functional organization of biological neural systems. ANNs are capable of learning from data, can approximate variety of nonliear functions and their robust handling of noisy data make them suitable candidates for inference of gene regulatory interactions from genomics data. Several variants of ANNs have been successfully applied, for modeling gene regulatory interactions, including multilayer perceptrons (Kim et al., 2000; J. Huang et al., 2003; Zhou et al., 2004), self-organizing maps (SOM) (Weaver et al., 1999) and recurrent neural networks (RNNs) (VOHRADSKÝ, 2001; Keedwell et al., 2002; Tian & Burrage, 2003; R. Xu et al., 2004, 2007a, 2007b; Ghazikhani et al., 2011; Noman et al., 2013; Raza & Alam, 2016). However, none of these approaches integrate heterogeneous data to infer GRNs, with majority of them using only gene expression data. There are very few studies reported in the literature where an integration framework is used for genomics studies based on ANN model, such as deep learning (LeCun et al., 2015) based multi-modal framework as illustrated in Fig. 5.

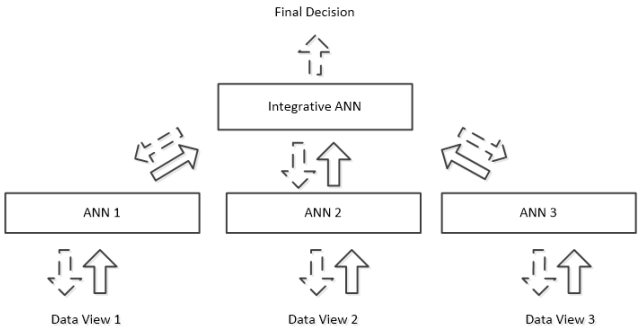

Fig. 5. Additive multi-modal deep learning for data integration. Different deep learning models can be applied, as sub-networks, to individual data views.An integrative network combines information from the sub-networks. The model can be either directed or undirected; either supervised or unsupervised.Bottom-up arrows indicate a discriminative model. Downward or undirected connections indicate a generative model.

The fundamental idea here is to build a subnetwork for each data view/dataset and then integrate the output of the individual subnetworks into the higher layers. Subnetworks at lower layers offer the flexibility for selecting appropriate deep learning models for different data types, such as deep belief net (DBN) (Hinton et al., 2006) or deep Boltzmann machine (DBM) (Salakhutdinov & Larochelle, 2010) for binary/Gaussian data, convolutional networks (LeCun et al., 1995) for image data, RNNs (Graves & Jaitly, 2014) for sequential signal selection and deep feature selection (Y. Li et al., 2015) for discriminative features. The deep learning model can be either supervised or unsupervised with

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 April 2018 doi:10.20944/preprints201804.0352.v1

subnetworks being either directed or undirected.Besides their application to learn representations from image data (Srivastava & Salakhutdinov, 2012), multi modal DBN models have also been applied to genomics data to integrate gene expression, DNA methylation and drug response data for subtyping tumours (Liang et al., 2015) on the basis of survival, showing superior performance to k-means clustering. Deep learning models have also been applied to study gene regulatory mechanisms, algorithms such as DeepBind (Alipanahi et al., 2015) has been shown to be broadly applicable and results in increased predictive power compared to traditional single-domain methods and offers a scalable, flexible and unified computational approach for pattern discovery in regulatory genomics outperforming most of the state-of-art methods when combining multiple genomic datasets (protein binding microarrays, ChIP-seq and RNA-compete assays). Although we can use this multi modal deep learning architecture as a template to build scalable regulatory networks, however, we could not trace any single study based on deep learning integrative approaches that are employed to learn GRNs from multi modal data.

9 Discussion & Conclusion

Molecular information produced by high throughput technologies (NGS, mass spectrometry, microarray etc.) accruing at gigantic scales has opened floodgates to the knowledge of biological world. Large scale information about genomic factors (genes, proteins, mRNA, miRNA, metabolites, SNPs etc.), that characterize diverse biological processes from both prokaryotes and eukaryotes can be obtained using modern experimental technologies. Besides, huge volumes of annotation and functional information about genes, proteins ,DNA, RNA, diseases and signaling pathways is housed in publicly accessible databases. The volume and diversity of this information has led to the emergence of new statistical techniques and development of computational tools with the sole purpose of identifying genomic factors and their interactions that play an essential role in the architecture of complex phenotypes and continue to unravel novel biological insights. The realization that inference studies using single data type possess limited predictive power has led to the development of data integration methods in order to widen the scope of exploring novel interactions, disease risks , drug targets or any other biological outcome. Although not a single gold standard method has emerged out of these system genomics approaches, various strategies can be adopted to perform a powerful integrative analysis without relying on a single best approach when applied to heterogeneous datasets. Still these data integration approaches provide a means of analysis multiple datasets simultaneously and quite comprehensively for a comprehensive understanding of biological systems.

Here we present a detailed review of some sophisticated statistical and machine learning techniques used in the inference of regulatory network from multiple (homogeneous / heterogeneous) datasets. The aim is to bring together diverse set of tools and techniques and highlight their potential and limitations in solving biological inference problems when presented with multiple sets of data.

Given the complexity and heterogeneity of data, approaches included here do not overcome all the data integration challenges as outlined in section II , therefore our emphasis will be how a particular method solves a particular biological problem or handles a specific data type. In that context, the network integration methods offer a simple and straightforward solution whereby similar nodes (genes/proteins) across multiple networks are integrated by merging different types of edges. Similarly Similarity Network Fusion build similarity networks for multiple datasets and then merges these individual networks using a novel fusion method. Although simple, but they are less efficient when it comes to preserving the relationships across multiple networks, particularly when multiple networks are projected on each other. Probabilistic Graphical models such as, bayesian networks and Markov networks are often considered as popular approaches to model regulatory networks as they provide suitable frameworks for biological problems. Both BNs and Markov networks model conditional independence relationships and can handle noisiness inherent in biological data. Although BNs do not scale well for large scale datasets but using a variable selection approach, they can outperform some state-of-the-art methods proposed so far in the literature. Markov networks are the other variants of Probabilistic graphical models employed to infer gene interactions from multiple genomics data sets, for example FUSENET uses Markov networks for uncovering conditional dependence relationship between genes whereas regressions are employed to reconstruct a network from the identified interactions in an iterative manner, thereby circumventing the computational cost of network inference.

Similarly regularized regression using LASSO based formulations provide effective statistical tools to model and infer the regulatory relationships embedded in genomic datasets. Regularization parameters are tuned to control the sparsity of the inferred network, but also harnesses the information contained in multiple datasets. These methods sometimes lose some important independent variables (hence, interactions) because of the feature selection performed by shrinkage parameters. Kernel based methods besides integrating multiple and diverse datasets can also handle large-scale data, provided a kernel function for a specific data type is available. In case there is limited scope for the function to be defined, data integration for heterogeneous data types in those cases becomes less effective. As for the integration strategy adopted by various methods, intermediate (transformation based) have been reported to achieve improved accuracies in inference studies from multiple datasets. Moreover, kernel based methods provide automatic means for weighing data in an integration framework and can be extended to other methods for selecting informative datasets. ANN based deep learning models possess several strengths for data integration, they offer component-wise learning paradigm that can significantly reduce the computational cost. For example, for learning model parameters, the sub-networks can be pre-trained using different data views, separately, and then the parameter of the entire network can be globally fine tuned. Also using multi-modal deep networks, the heterogeneous information from different source can be jointly considered well in the integrative layers for inference, classification and clustering

(LeCun et al., 2015). Multi-modal networks can even learn from incomplete data views which enable the maximal use of available data instead of using samples with complete views (Srivastava & Salakhutdinov, 2012). Because of their complex structure, models based on deep neural networks suffer from over fitting, are not easily scalable and need large amount of data for training.

The methods for data integration and inference included here do not provide perfect solutions, but nonetheless have yielded good results. Although there are no set standards to validate the performance of these methods in terms of quality of the integration, neither there is any proper criteria by which we can compare two methods for performance. Nonetheless, research studies using systems genomics approaches are being carried out at an unprecedented scale to handle this data explosion in the field of biology. As the generation of biological data continues across different omics layers, development of more comprehensive systems genomic strategies and tools will lead to an enhanced comprehension of complex-trait architecture and produce new knowledge about human physiology and diseases.

Acknowledgment

The author Nisar Wani acknowledges Teacher Fellowship received from University Grants Commission, Ministry of Human Resources Development, Govt. of India vide letter No. F.BNo. 27-(TF-45)/2015 under Faculty Development Programme.

References

Alcaraz, N., Kücük, H., Weile, J., Wipat, A., & Baumbach, J. (2011). Keypathwayminer: detecting case-specific biological pathways using expression data. *Internet Mathematics*, *7*(4), 299–313.

Alipanahi, B., Delong, A., Weirauch, M. T., & Frey, B. J. (2015). Predicting the sequence specificities of dna-and rna-binding proteins by deep learning. *Nature biotechnology*, *33*(8), 831.

Allen, G. I., & Liu, Z. (2013). A local poisson graphical model for inferring networks from sequencing data. *IEEE transactions on nanobioscience*, *12*(3), 189–198.

Banf, M., & Rhee, S. Y. (2017). Enhancing gene regulatory network inference through data integration with markov random fields. *Scientific Reports*, *7*.

Bank, P. D. (1971). Protein data bank. *Nature New Biol*, *233*, 223.

Barzel, B., & Barabási, A.-L. (2013). Network link prediction by global silencing of indirect correlations. *Nature biotechnology*, *31*(8), 720.

Beaumont, M. A., & Rannala, B. (2004). The bayesian revolution in genetics. *Nature Reviews Genetics*, *5*(4), 251–261.

Ben-Gal, I., Ruggeri, F., Faltin, F., & Kenett, R. (2007). *Bayesian networks, encyclopedia of statistics in quality and reliability.* John Wiley and Sons.

Ben-Hur, A., & Brutlag, D. (2003). Remote homology detection: a motif based approach. *Bioinformatics*, *19*(suppl_1), i26–i33.

Ben-Hur, A., & Noble, W. S. (2005). Kernel methods for predicting protein–protein interactions. *Bioinformatics*, *21*(suppl_1), i38–i46.

Bernstein, B. E., Stamatoyannopoulos, J. A., Costello, J. F., Ren, B., Milosavljevic, A., Meissner, A., ... others (2010). The nih roadmap epigenomics mapping consortium. *Nature biotechnology*, *28*(10), 1045.

Binder, H., & Schumacher, M. (2009). Incorporating pathway information into boosting estimation of high-dimensional risk prediction models. *BMC bioinformatics*, *10*(1), 18.

Borgwardt, K. M., Ong, C. S., Schönauer, S., Vishwanathan, S., Smola, A. J., & Kriegel, H.-P. (2005). Protein function prediction via graph kernels. *Bioinformatics*, *21*(suppl_1), i47–i56.

Bradford, J. R., Needham, C. J., Bulpitt, A. J., & Westhead, D. R. (2006). Insights into protein–protein interfaces using a bayesian network prediction method. *Journal of molecular biology*, *362*(2), 365–386.

Cai, X., Bazerque, J. A., & Giannakis, G. B. (2013). Inference of gene regulatory networks with sparse structural equation models exploiting genetic perturbations. *PLoS computational biology*, *9*(5), e1003068.

Chartrand, R. (2007). Exact reconstruction of sparse signals via nonconvex minimization. *IEEE Signal Processing Letters*, *14*(10), 707–710.

Cheng, F., Liu, C., Jiang, J., Lu, W., Li, W., Liu, G., ... Tang, Y. (2012). Prediction of drug-target interactions and drug repositioning via network-based inference. *PLoS computational biology*, *8*(5), e1002503.

Chickering, D. M. (1996). Learning bayesian networks is np-complete. *Learning from data: Artificial intelligence and statistics V*, *112*, 121–130.

Cho, H., Berger, B., & Peng, J. (2016). Compact integration of multi-network topology for functional analysis of genes. *Cell systems*, *3*(6), 540–548.

Chun, H., Chen, M., Li, B., & Zhao, H. (2013). Joint conditional gaussian graphical models with multiple sources of genomic data. *Frontiers in genetics*, *4*.

Consortium, E. P., et al. (2004). The encode (encyclopedia of dna elements) project. *Science*, *306*(5696), 636–640.

Consortium, G. O. (2014). Gene ontology consortium: going forward. *Nucleic acids research*, *43*(D1), D1049–D1056.

Dasgupta, A., Sun, Y. V., König, I. R., Bailey-Wilson, J. E., & Malley, J. D. (2011). Brief review of regression-based and machine learning methods in genetic epidemiology: the genetic analysis workshop 17 experience. *Genetic epidemiology*, *35*(S1).

Enright, A. J., John, B., Sander, C., Marks, D. S., Tuschl, T., & Gaul, U. (2003). Microrna targets in drosophila. *Genome biology*, *5*(1), R1.

Faith, J. J., Hayete, B., Thaden, J. T., Mogno, I., Wierzbowski, J., Cottarel, G., ... Gardner, T. S. (2007). Large-scale mapping and validation of escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS biology*, *5*(1), e8.

Feizi, S., Marbach, D., Médard, M., & Kellis, M. (2013). Network deconvolution as a general method to distinguish direct dependencies in networks. *Nature biotechnology*, *31*(8), 726.

Fornasier, M., & Rauhut, H. (2008). Iterative thresholding algorithms. *Applied and Computational Harmonic Analysis*, *25*(2), 187–208.

Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., ... others (2012). String v9. 1: protein-protein interaction networks, with increased coverage and integration. *Nucleic acids research*, *41*(D1), D808–D815.

Friedman, Linial, M., Nachman, I., & Pe'er, D. (2000). Using bayesian networks to analyze expression data. *Journal of computational biology*, *7*(3-4), 601–620.

Friedman, J., Hastie, T., & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, *9*(3), 432–441.

Friedman, N. (2004). Inferring cellular networks using probabilistic graphical models. *Science*, *303*(5659), 799–805.

Gade, S., Porzelius, C., Fälth, M., Brase, J. C., Wuttig, D., Kuner, R., ... Beißbarth, T. (2011). Graph based fusion of mirna and mrna expression data improves clinical outcome prediction in prostate cancer. *BMC bioinformatics*, *12*(1), 488.

Gallopin, M., Rau, A., & Jaffrézic, F. (2013). A hierarchical poisson log-normal model for network inference from rna sequencing data. *PloS one*, *8*(10), e77503.

Gevaert, O., Smet, F. D., Timmerman, D., Moreau, Y., & Moor, B. D. (2006). Predicting the prognosis of breast cancer by integrating clinical and microarray data with bayesian networks. *Bioinformatics*, *22*(14), e184–e190.

Ghazikhani, A., Akbarzadeh, T. M. R., & Monsefi, R. (2011). Genetic regulatory network inference using recurrent neural networks trained by a multi agent system. In *Computer and knowledge engineering (iccke), 2011 1st international econference on* (pp. 95–99).

Gligorijević, V., & Pržulj, N. (2015). Methods for biological data integration: perspectives and challenges. *Journal of the Royal Society Interface*, *12*(112), 20150571.

Gomez, S. M., Noble, W. S., & Rzhetsky, A. (2003). Learning to predict protein–protein interactions from protein sequences. *Bioinformatics*, *19*(15), 1875–1881.

Graves, A., & Jaitly, N. (2014). Towards end-to-end speech recognition with recurrent neural networks. In *International conference on machine learning* (pp. 1764–1772).

Guo, X., Gao, L., Wei, C., Yang, X., Zhao, Y., & Dong, A. (2011). A computational method based on the integration of heterogeneous networks for predicting disease-gene associations. *PloS one*, *6*(9), e24171.

Hamid, J. S., Hu, P., Roslin, N. M., Ling, V., Greenwood, C. M., & Beyene, J. (2009). Data integration in genetics and genomics: methods and challenges. *Human genomics and proteomics: HGP*, *2009*.

Han, H., Shim, H., Shin, D., Shim, J. E., Ko, Y., Shin, J., ... others (2015). Trrust: a reference database of human transcriptional regulatory interactions. *Scientific reports*, *5*, srep11432.

Heckerman, D. (1998). A tutorial on learning with bayesian networks. In *Learning in graphical models* (pp. 301–354). Springer.

Hesterberg, T., Choi, N. H., Meier, L., Fraley, C., et al. (2008). Least angle and 1 penalized regression: A review. *Statistics Surveys*, *2*, 61–93.

Hinton, G. E., Osindero, S., & Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural computation*, *18*(7), 1527–1554.

Huang, Yeh, H.-Y., & Soo, V.-W. (2013). Inferring drug-disease associations from integration of chemical, genomic and phenotype data using network propagation. *BMC medical genomics*, *6*(3), S4.

Huang, J., Shimizu, H., & Shioya, S. (2003). Clustering gene expression pattern and extracting relationship in gene network based on artificial neural networks. *Journal of bioscience and bioengineering*, *96*(5), 421–428.

Irrthum, A., Wehenkel, L., Geurts, P., et al. (2010). Inferring regulatory networks from expression data using tree-based methods. *PloS one*, *5*(9), e12776.

Isci, S., Dogan, H., Ozturk, C., & Otu, H. H. (2013). Bayesian network prior: network analysis of biological data using external knowledge. *Bioinformatics*, *30*(6), 860–867.

Jalali, A., Ravikumar, P., Vasuki, V., & Sanghavi, S. (2011). On learning discrete graphical models using group-sparse regularization. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics* (pp. 378–387).

Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N. J., Chung, S., ... Gerstein, M. (2003). A bayesian networks approach for predicting protein-protein interactions from genomic data. *science*, *302*(5644), 449–453.

Jiang, C., Xuan, Z., Zhao, F., & Zhang, M. Q. (2007). Tred: a transcriptional regulatory element database, new entries and other development. *Nucleic acids research*, *35*(suppl_1), D137–D140.

Kariya, T., & Kurata, H. (2004). Generalized least squares estimators. *Generalized Least Squares*, 25–66.

Kashima, H., Tsuda, K., & Inokuchi, A. (2004). Kernels for graphs. *Kernel methods in computational biology*, *39*(1), 101–113.

Keedwell, E., Narayanan, A., & Savic, D. (2002). Modelling gene regulatory data using artificial neural networks. In *Neural networks, 2002. ijcnn'02. proceedings of the 2002 international joint conference on* (Vol. 1, pp. 183–188).

Kim, S., Dougherty, E. R., Chen, Y., Sivakumar, K., Meltzer, P., Trent, J. M., & Bittner, M. (2000). Multivariate measurement of gene expression relationships. *Genomics*, *67*(2), 201–209.

Kondor, R. I., & Lafferty, J. (2002). Diffusion kernels on graphs and other discrete input spaces. In *Icml* (Vol. 2, pp. 315–322).

Krämer, N., Schäfer, J., & Boulesteix, A.-L. (2009). Regularized estimation of large-scale gene association networks using graphical gaussian models. *BMC bioinformatics*.

Lanckriet, G. R., De Bie, T., Cristianini, N., Jordan, M. I., & Noble, W. S. (2003a). A framework for genomic data fusion and its application to membrane protein prediction. *Computer Science*.

Lanckriet, G. R., De Bie, T., Cristianini, N., Jordan, M. I., & Noble, W. S. (2003b). Kernel-based data fusion and its application to protein function prediction in yeast. In *Biocomputing 2004* (pp. 300–311). World Scientific.

Lanckriet, G. R., De Bie, T., Cristianini, N., Jordan, M. I., & Noble, W. S. (2004). A statistical framework for

genomic data fusion. *Bioinformatics*, *20*(16), 2626–2635.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436.

LeCun, Y., Bengio, Y., et al. (1995). Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, *3361*(10), 1995.

Lee, I., Date, S. V., Adai, A. T., & Marcotte, E. M. (2004). A probabilistic functional network of yeast genes. *science*, *306*(5701), 1555–1558.

Leslie, C., Eskin, E., & Noble, W. S. (2001). The spectrum kernel: A string kernel for svm protein classification. In *Biocomputing 2002* (pp. 564–575). World Scientific.

Leslie, C. S., Eskin, E., Cohen, A., Weston, J., & Noble, W. S. (2004). Mismatch string kernels for discriminative protein classification. *Bioinformatics*, *20*(4), 467–476.

Li, S. Z. (1994). Markov random field models in computer vision. In *European conference on computer vision* (pp. 361–370).

Li, Y., Chen, C.-Y., & Wasserman, W. W. (2015). Deep feature selection: Theory and application to identify enhancers and promoters. In *International conference on research in computational molecular biology* (pp. 205–217).

Liang, M., Li, Z., Chen, T., & Zeng, J. (2015). Integrative data analysis of multi-platform cancer data with a multimodal deep learning approach. *IEEE/ACM transactions on computational biology and bioinformatics*, *12*(4), 928–937.

Linghu, B., DeLisi, C., Snitkin, E. S., Xia, Y., & Hu, Z. (2009). Genome-wide prioritization of disease genes and identification of disease-disease associations from an integrated human functional linkage network. *Genome biology*, *10*(9), R91.

Lo, K., Raftery, A. E., Dombek, K. M., Zhu, J., Schadt, E. E., Bumgarner, R. E., & Yeung, K. Y. (2012). Integrating external biological knowledge in the construction of regulatory networks from time-series expression data. *BMC systems biology*, *6*(1), 101.

Lu, L. J., Xia, Y., Paccanaro, A., Yu, H., & Gerstein, M. (2005). Assessing the limits of genomic data integration for predicting protein networks. *Genome research*, *15*(7), 945–953.

Luo, Y., Zhao, X., Zhou, J., Yang, J., Zhang, Y., Kuang, W., ... Zeng, J. (2017). A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. *Nature communications*, *8*(1), 573.

Marbach, D., Costello, J. C., Küffner, R., Vega, N. M., Prill, R. J., Camacho, D. M., ... others (2012). Wisdom of crowds for robust gene network inference. *Nature methods*, *9*(8), 796–804.

Margolin, A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R., & Califano, A. (2006). Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. In *Bmc bioinformatics* (Vol. 7, p. S7).

Margolin, A. A., Wang, K., Lim, W. K., Kustagi, M., Nemenman, I., & Califano, A. (2006). Reverse engineering cellular networks. *Nature protocols*, *1*(2), 662–671.

Metzler, D., & Croft, W. B. (2005). A markov random field model for term dependencies. In *Proceedings of the 28th annual international acm sigir conference on research and development in information retrieval* (pp. 472–479).

Meyer, P. E., Kontos, K., Lafitte, F., & Bontempi, G. (2007). Information-theoretic inference of large transcriptional regulatory networks. *EURASIP journal on bioinformatics and systems biology*, *2007*(1), 79879.

Mordelet, F., & Vert, J.-P. (2008). Sirene: supervised inference of regulatory networks. *Bioinformatics*, *24*(16), i76–i82.

Mostafavi, S., Ray, D., Warde-Farley, D., Grouios, C., & Morris, Q. (2008). Genemania: a real-time multiple association network integration algorithm for predicting gene function. *Genome biology*, *9*(1), S4.

Mukherjee, S., & Speed, T. P. (2008). Network inference using informative priors. *Proceedings of the National Academy of Sciences*, *105*(38), 14313–14318.

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

Needham, C. J., Bradford, J. R., Bulpitt, A. J., & Westhead, D. R. (2007). A primer on learning in bayesian networks for computational biology. *PLoS computational biology*, *3*(8), e129.

Noman, N., Palafox, L., & Iba, H. (2013). Reconstruction of gene regulatory networks from gene expression data using decoupled recurrent neural network model. In *Natural computing and beyond* (pp. 93–103). Springer.

Omranian, N., Eloundou-Mbebi, J. M., Mueller-Roeber, B., & Nikoloski, Z. (2016). Gene regulatory network inference using fused lasso on multiple data sets. *Scientific reports*, *6*, 20533.

Özen, A., Gönen, M., Alpaydın, E., & Haliloğlu, T. (2009). Machine learning integration for predicting the effect of single amino acid substitutions on protein stability. *BMC Structural Biology*, *9*(1), 66.

Patel, R. K., & Jain, M. (2012). Ngs qc toolkit: a toolkit for quality control of next generation sequencing data. *PloS one*, *7*(2), e30619.

Pavlidis, P., Weston, J., Cai, J., & Grundy, W. N. (2001). Gene functional classification from heterogeneous data. In *Proceedings of the fifth annual international conference on computational biology* (pp. 249–255).

Qin, J., Hu, Y., Xu, F., Yalamanchili, H. K., & Wang, J. (2014). Inferring gene regulatory networks by integrating chip-seq/chip and transcriptome data via lasso-type regularization methods. *Methods*, *67*(3), 294–303.

Ravikumar, P., Wainwright, M. J., Lafferty, J. D., et al. (2010). High-dimensional ising model selection using 1-regularized logistic regression. *The Annals of Statistics*, *38*(3), 1287–1319.

Raza, K., & Alam, M. (2016). Recurrent neural network based hybrid model for reconstructing gene regulatory network. *Computational biology and chemistry*, *64*, 322–334.

Ritchie, M. D., Holzinger, E. R., Li, R., Pendergrass, S. A., & Kim, D. (2015). Methods of integrating data to uncover genotype-phenotype interactions. *Nature Reviews Genetics*, *16*(2), 85–97.

Rogers, S., & Girolami, M. (2005). A bayesian regression approach to the inference of regulatory networks from gene expression data. *Bioinformatics*, *21*(14), 3131–3137.

Salakhutdinov, R., & Larochelle, H. (2010). Efficient learning of deep boltzmann machines. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics* (pp. 693–700).

Santra, T. (2014). A bayesian framework that integrates heterogeneous data for inferring gene regulatory networks. *Frontiers in bioengineering and biotechnology*, *2*.

Santra, T., Kolch, W., & Kholodenko, B. N. (2013). Integrating bayesian variable selection with modular response analysis to infer biochemical network topology. *BMC systems biology*, *7*(1), 57.

Schäfer, J., & Strimmer, K. (2004). An empirical bayes approach to inferring large-scale gene association networks. *Bioinformatics*, *21*(6), 754–764.

Schölkopf, B., & Smola, A. J. (2002). *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press.

Schölkopf, B., Tsuda, K., & Vert, J.-P. (2004). *Kernel methods in computational biology*. MIT press.

Schwaller, L. (2015). An introduction to graphical models.

Shawe-Taylor, J., & Cristianini, N. (2004). *Kernel methods for pattern analysis*. Cambridge university press.

Siva, N. (2008). *1000 genomes project.* Nature Publishing Group.

Sonnenburg, S., Rätsch, G., Schäfer, C., & Schölkopf, B. (2006). Large scale multiple kernel learning. *Journal of Machine Learning Research*, *7*(Jul), 1531–1565.

Srivastava, N., & Salakhutdinov, R. R. (2012). Multimodal learning with deep boltzmann machines. In *Advances in neural information processing systems* (pp. 2222–2230).

Tian, T., & Burrage, K. (2003). Stochastic neural network models for gene regulatory networks. In *Evolutionary computation, 2003. cec'03. the 2003 congress on* (Vol. 1, pp. 162–169).

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., & Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(1), 91–108.

Tong, H., Faloutsos, C., & Pan, J.-Y. (2006). Fast random walk with restart and its applications.

Troyanskaya, O. G., Dolinski, K., Owen, A. B., Altman, R. B., & Botstein, D. (2003). A bayesian framework for combining heterogeneous data sources for gene function prediction (in saccharomyces cerevisiae). *Proceedings of the National Academy of Sciences*, *100*(14), 8348–8353.

Van Vliet, M. H., Horlings, H. M., Van De Vijver, M. J., Reinders, M. J., & Wessels, L. F. (2012). Integration of clinical and gene expression data has a synergetic effect on predicting breast cancer outcome. *PloS one*, *7*(7), e40358.

Vert, J.-P. (2002). A tree kernel to analyse phylogenetic profiles. *Bioinformatics*, *18*(suppl_1), S276–S284.

VOHRADSKÝ, J. (2001). Neural network model of gene expression. *the FASEB journal*, *15*(3), 846–854.

Wang, Mezlini, A. M., Demir, F., Fiume, M., Tu, Z., Brudno, M., ... Goldenberg, A. (2014). Similarity network fusion for aggregating data types on a genomic scale. *Nature methods*, *11*(3), 333–337.

Wang, C., Komodakis, N., & Paragios, N. (2013). Markov random field modeling, inference & learning in computer vision & image understanding: A survey. *Computer Vision and Image Understanding*, *117*(11), 1610–1627.

Wani, N., & Raza, K. (2017). Raw sequence to target gene prediction: An integrated inference pipeline for chip-seq and rna-seq datasets. *bioRxiv*, 220152.

Weaver, D. C., Workman, C. T., & Stormo, G. D. (1999). Modeling regulatory networks with weight matrices. In *Biocomputing'99* (pp. 112–123). World Scientific.

Wei, Z., & Li, H. (2007). A markov random field model for network-based analysis of genomic data. *Bioinformatics*, *23*(12), 1537–1544.

Xu, R., Hu, X., & Wunsch, D. C. (2004). Inference of genetic regulatory networks with recurrent neural network models. In *Engineering in medicine and biology society, 2004. iembs'04. 26th annual international conference of the ieee* (Vol. 2, pp. 2905–2908).

Xu, R., Venayagamoorthy, G. K., & Wunsch II, D. C. (2007a). Modeling of gene regulatory networks with hybrid differential evolution and particle swarm optimization. *Neural Networks*, *20*(8), 917–927.

Xu, R., Wunsch II, D., & Frank, R. (2007b). Inference of genetic regulatory networks with recurrent neural network models using particle swarm optimization. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *4*(4), 681–692.

Xu, Z., Chang, X., Xu, F., & Zhang, H. (2012). $l\_\{1/2\}$ regularization: A thresholding representation theory and a fast solver. *IEEE Transactions on neural networks and learning systems*, *23*(7), 1013–1027.

Yamanishi, Y., Vert, J.-P., & Kanehisa, M. (2004). Protein network inference from multiple genomic data: a supervised approach. *Bioinformatics*, *20*(suppl_1), i363–i370.

Yang, A. Y., Sastry, S. S., Ganesh, A., & Ma, Y. (2010). Fast 1-minimization algorithms and an application in robust face recognition: A review. In *Image processing (icip), 2010 17th ieee international conference on* (pp. 1849–1852).

Zaykin, D. V., Zhivotovsky, L. A., Westfall, P. H., & Weir, B. S. (2002). Truncated product method for combining p-values. *Genetic epidemiology*, *22*(2), 170–185.

Zhang, B., Gaiteri, C., Bodea, L.-G., Wang, Z., McElwee, J., Podtelezhnikov, A. A., ... others (2013). Integrated systems approach identifies genetic nodes and networks in late-onset alzheimers disease. *Cell*, *153*(3), 707–720.

Zhou, X., Wang, X., Pal, R., Ivanov, I., Bittner, M., & Dougherty, E. R. (2004). A bayesian connectivity-based approach to constructing probabilistic gene regulatory networks. *Bioinformatics*, *20*(17), 2918–2927.

Zhu, J., Zhang, B., Smith, E. N., Drees, B., Brem, R. B., Kruglyak, L., ... Schadt, E. E. (2008). Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nature genetics*, *40*(7), 854–861.

Žitnik, M., & Zupan, B. (2015a). Data fusion by matrix factorization. *IEEE transactions on pattern analysis and machine intelligence*, *37*(1), 41–53.

 doi:10.20944/preprints201804.0352.v1

Žitnik, M., & Zupan, B. (2015b). Gene network inference by fusing data from diverse distributions. *Bioinformatics*, *31*(12), i230–i239.

Zoppoli, P., Morganella, S., & Ceccarelli, M. (2010). Timedelay-aracne: Reverse engineering of gene networks from time-course data by an information theoretic approach. *BMC bioinformatics*, *11*(1), 154.