BioChemMed

Technium BioChemMed Vol. 2, Issue 4 pp.38-52 (2021)
ISSN: 2734-7990
www.techniumscience.com

Approaches on Modelling Genes Interactions: A Review

Dhafar Sami Hammadi[1], Basim Mahmood[2], and Marwah M. A. Dabdawb[3]

[1]*Computer Science Dept./ University of Mosul/ IRAQ*

[2]*Computer Science Dept./ University of Mosul/ IRAQ*
BioComplex Lab., Exeter, UK

[3]*Software Dept./ University of Mosul/ IRAQ*

[1]dhafar_un@uomosul.edu.iq, [2] bmahmood@uomosul.edu.iq/ biocomplexlab.org, and [3]marwa_marwan21@uomosul.edu.iq

Abstract. The human genome is a set of humans' nucleic acid sequences that are encoded as DNA in the human chromosome pairs. These are usually treated separately as the nuclear genome and the mitochondrial genome. Studying and understanding genetic systems has a great impact on health. Therefore, in the last few decades, the world has witnessed a great revolution in genetic engineering that aims to identify new phenomena and support mitigating the effect of diseases or even find ultimate solutions. The first step in studying genetics data is modelling this data and use some analysis approaches. The main problem of researchers is finding appropriate approaches for modelling their genetics data. This work comes to review the literature and present a variety of approaches used by researchers in modelling genetics data and genes interactions. This review tries to make it easy for researchers when adopting particular modelling approaches by presenting the state-of-the-art in terms of the dataset used, strengths, and limitations.

Keywords. Biological Networks, Modelling Genetics, Genes, Protein, Interactions, Complex Networks

1. Introduction

Genetics is the study of organisms' genes and is considered one of the most growing fields in biology [1]. Recently, the genetics field has been significantly developed and caused a paradigm shift in our life. Studies in genetics proved that organisms' (e.g., human, animals, or plants) genes have a direct impact on health and diseases [2][3]. Therefore, it is crucial to understand genes disorders and genes interactions aiming to mitigate diseases and maintain health. To this end, many approaches are available and can computerize complex genetics data into more understandable forms. These forms are called models that can be simulated and analysed using computer systems. Moreover, with the advent of new computer systems and applications, the analysis of such models has become more accurate and provide more reliable results [4].

Technium BioChemMed Vol. 2, Issue 4 pp.38-52 (2021)
ISSN: 2734-7990
www.techniumscience.com

The classical approaches of the genetic systems are not easy to be implemented due to the complexity of genetics data as well as the cost and time consumed [5]. For this reason, modelling genetics systems using computer systems and computational theories is important to be performed if we look for accurate knowledge.

Other approaches can also be used in modelling genetics data and genes interactions such as Mathematical and Statistical approaches [6][7][8]. Although these approaches provide accurate and reliable analysis, they still have a lack in providing sophisticated visualization. Furthermore, graphical modelling is currently considered one of the most attractive approaches for modelling genetics data [9]. These methods have the ability to provide advanced visualizations that enable researchers to look at their data from different aspects. As one of the newest methods that can model genetics data is Complex Networks [10]. This kind of data representation is based on the Graph Theory [11] that represent data objects as nodes that are connected by edges [12]. This kind of representation is powerful since genetic data objects have complex relations and interactions with each other. Therefore, complex networks approaches are widely used to model genetics data and especially genes interactions and protein interactions.

Furthermore, strong modelling of genetic interactions in organisms could provide crucial insight into complex diseases because genetic interactions provide insight into how genotype connects to phenotype in an organism and is considered to be useful in understanding multidrug resistance in some diseases and other biological abnormalities [13].

According to the literature, there is a lack in providing comprehensive reviews that consider the issue of modelling genetics data and genes interactions. Hence, the contribution of this work, that is, it provides a comprehensive review of the state-of-the-art of modelling genetics data in a way that enables researchers to have a wider view of the current approaches and makes it easier to select the most appropriate one for their data.

The rest of this paper is as follows: Section 2 illustrate the different approaches of modelling genes interactions. In section 3 discussions were made, and finally, some conclusions were explained in section 4.

2. Modelling Genes Interactions

2.1 Traditional Approaches

In the early stages of the biology field, the gene-related analysis and modelling approaches were based on traditional and classical ways. For instance, researchers used clinical and experimental results for analysing most of the gene-related issues [14]. Nowadays, some researchers still use this kind of approach. The study of Marok et al. [15] in 2021 used clinical data to model and analyse drug-gene interactions. Their model was build using PK-Sim simulation software and feed with 67 cases. However, their approach had limitations related to the evaluation of the output of the software, which needs highly skilled and experienced workers. Another study performed by Schafer et al. [16] in 2019 used a time-series-based approach for analysing genetic information of 8 autism patients. The main disadvantage of this approach was the difficulties in tracking the genetic temporal information of patients since it cannot be obtained regularly. However, the genetic data can be useful for future investigations. Regardless of the cost and the other limitations of the classical methods, they are considered reliable and provide credited and more trusted results. Furthermore, the traditional analysis of genetics data may become a step after applying pre-analysis to the data. In other words, an analysis may be applied before performing clinical experiments, which reduce the time and the cost consumed.

2.2 Mathematical and Statistical Approaches

Mathematical and statistical approaches for modelling genetics data was used by many researchers in the literature. This kind of method is considered to be more theoretical than its practical aspects. The distinguished work of Hansen et al. [17] in 2001 is a good example of this kind of approach. They

Technium BioChemMed Vol. 2, Issue 4 pp.38-52 (2021)
ISSN: 2734-7990
www.techniumscience.com

modelled genetics data in the form of a multilinear mathematical structure. Their proposed model was dynamic and able to analyse quantitative genetics data and understand experiments in a more comprehensive way. However, this approach may not be always suitable for biological data due to the nonlinear nature of this field's data. On the other hand, statistical methods have also been involved in modelling genetics data. Two years later, a work performed by Wu et al. [18] used the interactions/multi-way interactions and expressions among genes in the form of a Graphical Gaussian Model. The approach involved some clustering algorithms to extract model information aiming to rapidly explore genes relationships. The data used in their work was in the form of microarrays that needs special treatment by developers. However, this kind of approach needs further analysis by experts. Many years later, in 2014, the study of Patton et al. [19] used posterior probabilities to describe the relations among genome data. The authors believed that this kind of approach is considered an adequate tool for identifying uncategorized relations among genomes. However, these approaches are not always suitable for some kind of genetics data due to the variety of correlations among data objects.

2.3 Network-Based Approaches

Recent years have witnessed a great revolution in the use of network science approaches for modelling a different kinds of problems including genetics data [20][21]. For instance, the highly distinguished work of Barabasi et al. [22] in 2011 (Network Medicine) stated the main concepts for modelling biological networks and presented their main features. They described the basics that should be taken into considerations when modelling biological networks. The work also described the main models in network medicine including disease networks, disease-gene networks, gene-gene interactions networks, protein-protein interactions networks, and network pharmacology. Hence, this section presents a variety of approaches that model genetics problems into network models. For example, modelling gene-gene interactions can be formed as follows:

Assume the following set of gene-gene interactions:

#1: "Gene_1(Chromosome_X)" **interacts with** "Gene_2(Chromosome_X)"
#2: "Gene_1(Chromosome_X)" **interacts with** "Gene_3(Chromosome_X)"
#3: "Gene_1(Chromosome_X)" **interacts with** "Gene_1(Chromosome_Y)"
#4: "Gene_2(Chromosome_X)" **interacts with** "Gene_1(Chromosome_20)"
#5: "Gene_3(Chromosome_X)" **interacts with** "Gene_1(Chromosome_5)"
#6: "Gene_1(Chromosome_X)" **interacts with** "Gene_1(Chromosome_20)"

Then, the first step is to implement each gene as a node. After that, the edges are created based on the above-listed interactions (see Figure 1).

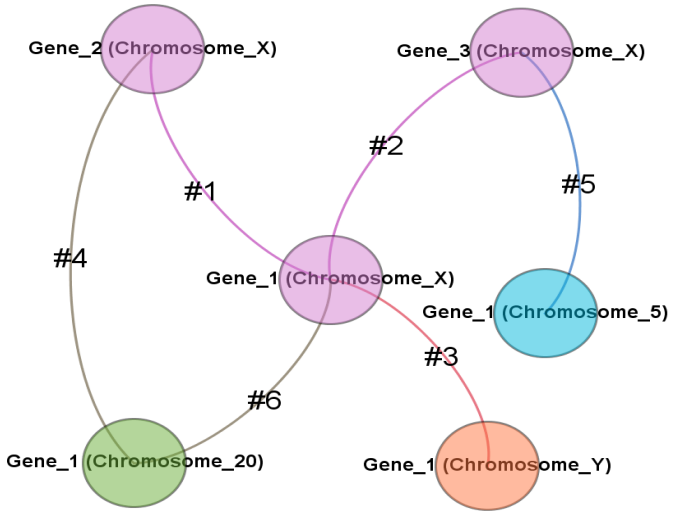

Fig. 1. An example illustrating how the gene-gene interactions network is generated.

Technium BioChemMed Vol. 2, Issue 4 pp.38-52 (2021)
ISSN: 2734-7990
www.techniumscience.com

2.3.1 Expression/Co-expression Networks

Kanakoglo et al. [23] in 2020 modelled gene expressions in the form of a network with nodes and edges and called "Differential Gene Expression Analysis (DGAs)". Their model was a protein-protein interactions network model, which means each protein (or the coded gene) is considered as a node, if there is an interaction between two proteins, an edge is created between them. The model was able to identify irradiated and non-irradiated cells. The same modelling approach was also used by Liu et al. [24] in 2020. They modelled gene expressions/co-expressions in the form of a network model. Their model was able to identify the hub genes, which enable researchers to understand and verifying genes. The limitations of their model were the small size of data, which leads to having 16 hub genes only. Practically, this kind of modelling needs to have huge datasets aiming to have more comprehensive network models that reflect reliable results. Care et al. [25] in 2019 used gene expressions and gene correlations in modelling genetics data. The authors aimed to investigate specific diseases using the modelled data. Their model was able to provide information that can be used in clinical experiments. The main issue in their approach was the ability to deal with multiple datasets in the same network model. Therefore, their model was only able to deal with small-scaled datasets.

2.3.2 Correlations Networks

Correlations in gene expressions/co-expression were also utilized in modelling genetics data in the literature. The work of Nayak et al. [26] in 2009 involved the correlations among gene co-expressions. The network was built using the correlation matrix among gene co-expressions. The goal of their model was to provide useful information about the relations between human genes and diseases. The model considered genes as nodes and the edges were created based on the correlations of the co-expression among genes. The main limitation of this kind of modelling is the accuracy of the correlations, which needs special attention by researchers. Similarly, Michalopoulos et al. [27] in 2012 used the same modelling approach in [26] but with a large scale of data. They applied some clustering algorithms to extract the genes with similar co-expression that perform similar processes in the human body. Their developed model was efficient in distinguishing the genes of similar functions. Moreover, their model was also able to predict information about the human genes alongside their relations and functions.

2.3.3 Gene-Gene (Protein-Protein) Interactions-Based Networks

The recent trend in modelling genetics data is using network science approaches. A network model consists of nodes and edges, where nodes are genes and the edges are created if there is an interaction between two genes. The strategy of building network models may vary from one research to another. The work of Santibanez et al. [28] in 2020 used a specific method in modelling genetics data. They considered 3 types of interactions among genes, namely, protein-protein, protein-DNA (GRN), and protein-metabolite. This strategy is considered to be complex, but it is efficient in modelling genes since it includes a lot of information encoded in the network model. This information can be retrieved using network measurements (e.g., community detection, centrality metrics, clustering coefficient, etc.). In the same context, Doncheva et al. [29] in 2018 proposed a network model that was based on protein-protein interactions but with some differences in creating network edges. The authors followed a particular strategy when creating edges that were based on giving a confidence score to each pair of nodes and then determine the weight of the edges within the network. The work was performed using the Cytoscape network visualization tool and STRING dataset. The aim of the work was to ease the process of querying and retrieving information about diseases and their related genes using a specific GUI for users. Another distinguished work proposed by Goh et al. [30] in 2007 suggested an approach that was based on the concepts of the bipartite graph to generate a network model. The model consisted of two disjoint sets of nodes; the first included genetic disorders, while the second one included the diseases genes in the human genome as shown in Figure 2 [30]. Their approach is considered an important tool for visually distinguishing some facts about genes and disorders. However, the limitation of this tool was the incompleteness of the data, which limits the extraction of information from the network model.

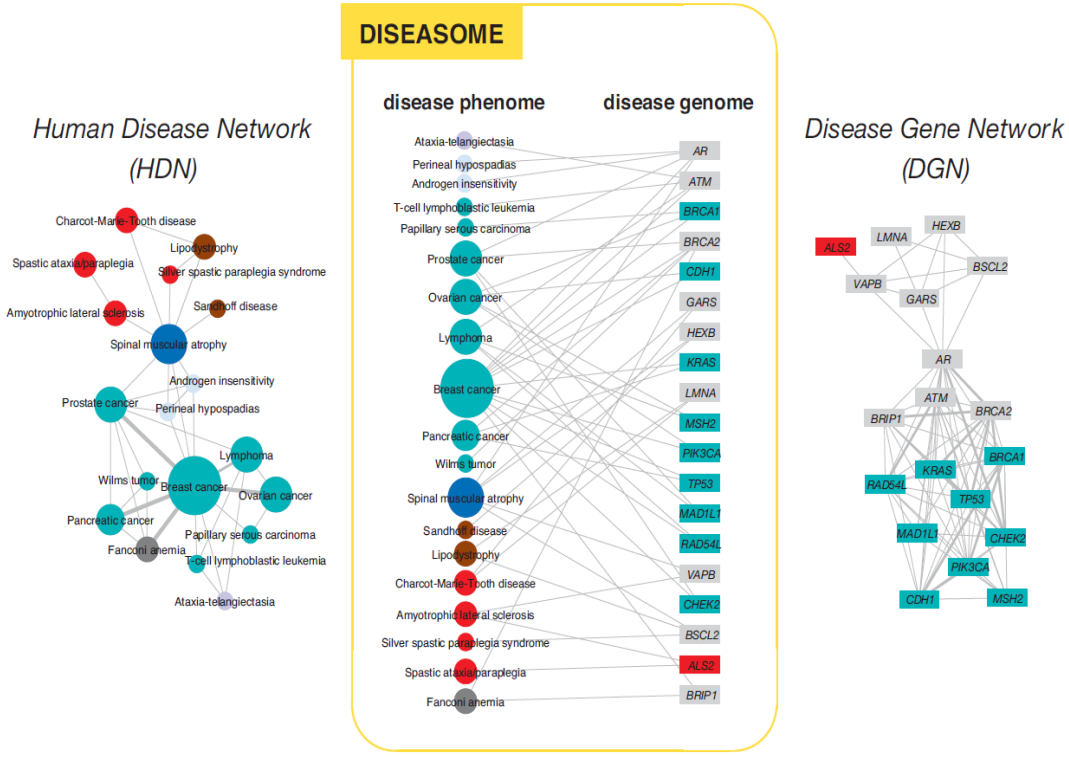

Fig. 2. Generating a network model for genes disorders and diseases genes [30].

The bipartite graph approach is considered to be effective when having a complete set of data [31]. Other types of network modelling can be used such as Boolean Networks, which is a graph that represents the genes as nodes and the edges among them is driven by their regulatory interactions [32][33]. Here, the expressions of genes are represented in the form of 0 and 1. The expression of a gene is set to 1 if it is above the threshold, otherwise, it is 0. Moreover, Boolean Network models are considered simple to implement and have the ability to simulate dynamic models, which is desired in biology. Schwab et al. [34] in 2020 developed a Boolean Network model aiming to study the dynamic complex behaviour of genetics systems and reveal facts on the interactions among proteins and the overexpression to diseases such as cancers. Their model is demonstrated in Figure 3, which shows how a biological phenomenon can be modelled in the form of a network. However, the main limitation of this kind of modelling is the inability to imitate real biological systems due to the difficulties in matching the simulation timing parameters, which leads to having unreliable results.

Another kind of approach called Gene Regularity Networks (GRNs) has been given a lot of attention in recent years for modelling genes data [35]. A GRN network consists of molecular regulators with interactions among them. GRNs have the ability to determine the functions of the cell through gene expression levels in the mRNA [36]. However, most of the GRNs approaches in the literature use limited labelled genes interactions in building genetics models, which leads to having unreliable results with no exploitation of the GRNs information. Hence, the study of Mignone et al. [37] in 2020 came to overcome the aforementioned limitations. They suggested a novel idea that involved unsupervised machine learning techniques on unlabeled interactions among genes. The approach was able to work in an unlabeled setting and can identify unknown functional relationships among genes.

The use of mathematical tools under the control theory in analyzing protein-protein interactions was extensively studied by Vinayagam et al. [38] in 2016. The study aimed to identify disease-causing mutations, prioritizing cancer genes, and identifying disease genes and their drug targets. In the study, the authors represented proteins as nodes and a direct edge is created between two nodes if they interact

Technium BioChemMed Vol. 2, Issue 4 pp.38-52 (2021)
ISSN: 2734-7990
www.techniumscience.com

with each other and the direction is driven by the signal flow of the interacted proteins. The weight of the edges is controlled by the level of confidence in the predicted direction. The proposed model was able to deal with large-scale datasets. The main limitation of such models is that the true functional form of complex biological systems is not known.

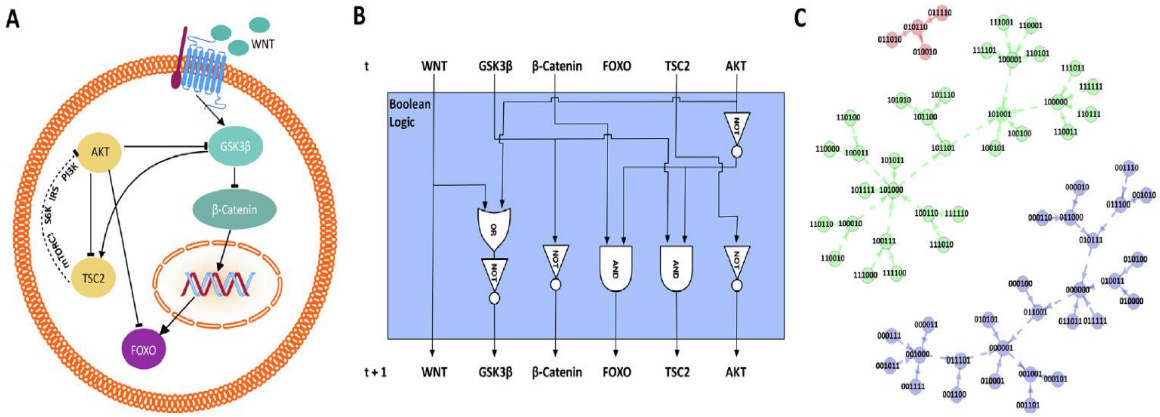

Fig. 3. (A) a genetic phenomenon, (B) Boolean network model, (C) a complete dynamic network model [34]

Furthermore, modelling genes expressions may lead to generate networks with a lot of false-positive relations (edges) [39]. This specific issue was investigated in the work of Pio et al. [40] in 2020. They proposed an approach based on the genes expressions that was able to utilize indirect relations among genes and overcome the aforementioned issue by removing these edges. The approach is also used to predict community structure accurately. The authors proved that their model was reliable and can work accurately even with the existence of noisy data. Figure 4 illustrates how the authors modelled their network [40].

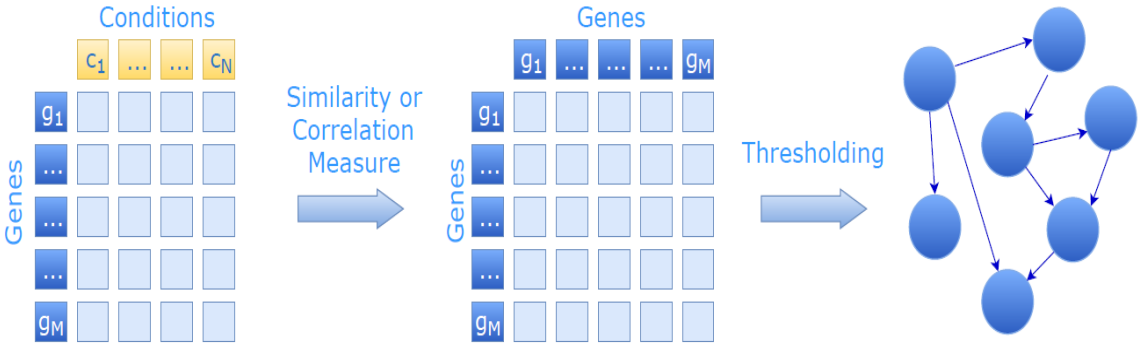

Fig. 4. Network reconstruction based on gene expression data [40].

In the same of the above context, a study performed by Lim et al. [41] investigated the issue of suspected edges in protein-protein interactions networks. They used similar the previous strategies in modelling genetics data but they incorporated literature-curated and evolutionarily conserved interactions. Their modelling approach was successful in identifying novel protein-protein interactions. This means the traditional modelling of protein-protein interactions can be developed by embedding more features and information within the network model. For instance, the pathway of genes can be

Technium BioChemMed Vol. 2, Issue 4 pp.38-52 (2021)
ISSN: 2734-7990
www.techniumscience.com

involved in the analysis of genetics data. A biological pathway can be defined as a series of molecular actions that eventually lead to a change in cells [42]. Reyna et al. [43] in 2020 involved coding/non-coding mutations of cancer genomes in developing a genetic model. They aimed to prioritize the rare (less frequently) mutations in protein-coding genes. Their model was able to identify new component that is frequently affected by non-coding and coding mutations. Figure 5 demonstrates the approach in details. In the same context, a study performed by Liu et al. [44] in 2010 investigated candidate genes that may cause rare diseases such as rare cancers (e.g., glioma). They involved genes interactions of a rare cancer and pathways in modelling a biological network. The study enabled researchers to understand the biological properties of genes and provide pathway maps for future investigations of rare cancers. The main limitation of this kind of modelling is the accuracy of the relationships among network genes since gene-gene interactions may be resulted from various cellular conditions.

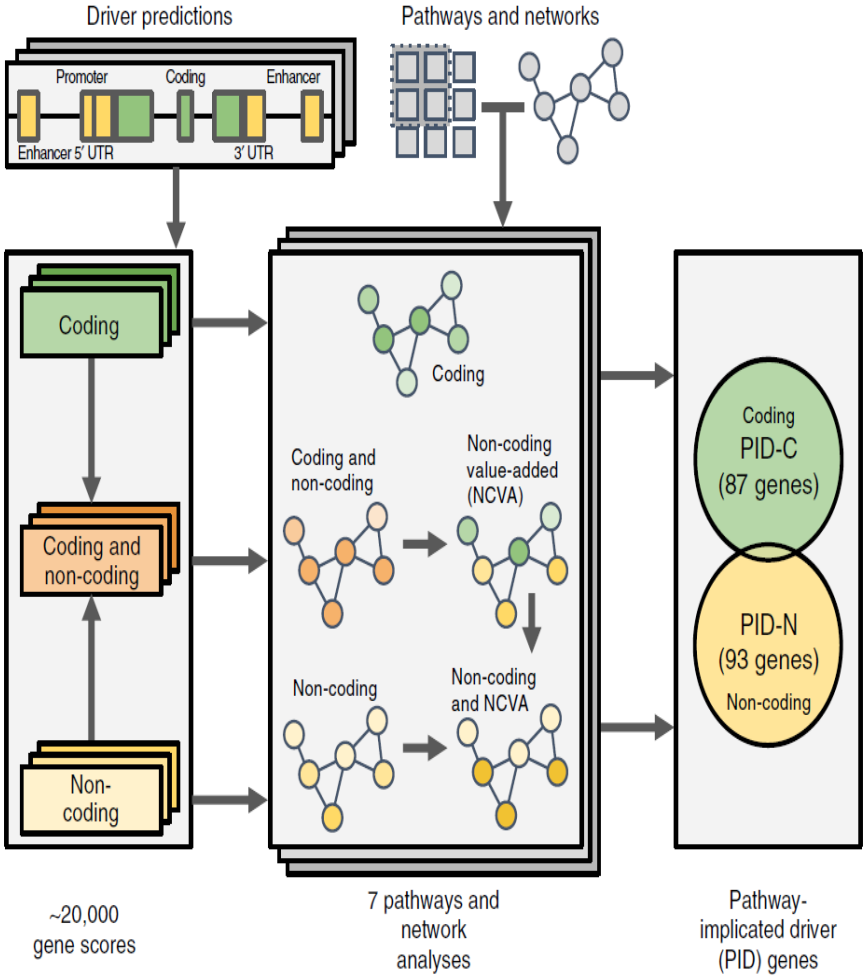

Fig. 5. Pathways and network analysis modelling approach [43].

In biological networks, large-scale data is preferred since it can provide deep knowledge about the problem of interest [45]. Therefore, large-scale network models produce more accurate results. This is because network-based models depend on both; objects information and the relations among network objects. Large-scale network models can be built using large scale data or by combining more than one network model together in one giant component [46]. Besides, several network models can be analysed

Technium BioChemMed Vol. 2, Issue 4 pp.38-52 (2021)
ISSN: 2734-7990
www.techniumscience.com

together and extract facts based on the collective analysis of these models. This specific situation was performed by Huang et al. [47] in 2018. They analysed 21 gene-gene interactions network models together. The results showed many novel interactions in a more efficient performance compared to other network models. The study also proved that giant network models always provide better performance in identifying phenomena in genes or protein interactions.

Network measurements can be powerful in analysing genetics data. Each network measurement can reflect a particular fact about a particular node of interest in the network. These measurements can be used in both network and node levels as follows:

Network Level [48][49]:

- Network Distribution: it shows the probability distribution of nodes degrees in the network. It can be considered as an indicator of the connectivity of network nodes (e.g., genes, protein, drug, disease, etc.).
- Average Degree: an indicator of the average connections of network nodes.
- Average Path Length: it shows the average number of shortest steps among network pairs.
- Diameter: the distance between the two farthest nodes within the network.
- Density: reflects the number of potential edges within the network to the actual number of edges within the same network.

Node Level [48][49]:

- Clustering Coefficient: reflects the tendency of a node to cluster with other network nodes in a network.
- Closeness Centrality: reflects how close a node is to other network nodes.
- Betweenness Centrality: shows how well-positioned a node is within the shortest paths of network nodes. This measurement tells us how important a node's position is in the flow of information in a network.
- Eigen Centrality: reflects how well-connected a node is to the highly connected nodes within a network.

The aforementioned measurements can play a significant role in analyzing genetics data since they deeply consider the relations among network nodes, which is of interest to researchers. In this regard, Miryala et al. [50] in 2021 utilized network measurements when analyzing their network of protein-protein interactions. They involved average path length, clustering coefficient, closeness centrality, betweenness centrality, and degree centrality of nodes in investigating and analyzing the topological features of genes. Their study was able to provide new strategies against some types of infectious diseases.

3. Discussions

As can be seen in the previous sections, genetics data can be modelled using a variety of approaches. The traditional approaches are based on classical clinical experiments that can be performed in laboratories and their results are considered more reliable but with high cost and time consumed. Usually, an experiment may be performed more than one time due to a lack of parameters or the use of the "trial and error" concept, which costs a lot. Therefore, researchers try to perform some analysis as a proactive step before performing experiments. These pre-analysis procedures are currently a trend in the literature since they provide researchers with most of the required knowledge before starting their experiments. The pre-analysis procedures can be performed by converting biological problems to computerized models. These models can be mathematical-based, statistical-based, network-based, or even a combination of them. It is always up to researchers when trying to model their problems. However, the literature presents a large number of approaches that support researchers and guide them to the best path. Hence, this work makes it easier for researchers when they decide to adopt particular approaches for their data. This work discussed most of the available approaches that are used in the literature for modelling genetics data. In addition to the description provided in the previous section, we summarize all the presented approaches in Table 1. The table provides information for readers about the

approaches presented in terms of the dataset used, the main contribution of each approach, the limitations of the approach, and other information.

Table 1. Summarizing the approaches presented in this work

References	Dataset Used	Method	Strengths	Limitations
Traditional Approaches				
Marok et al. (2021) [15]	67 clinical studies data. https://github.com/Open-Systems-Pharmacology and used GetData Graph Digitizer 2.26.0.20.	The model involved Pharmacokinetic (PBPK) model of bupropion including its DDI-relevant metabolites, and also involved clinical drug-gene interaction (DGI) and DDI data.	The model is flexible in simulating various DDGI scenarios.	Metabolizers receiving the CYP2B6 inducer rifampicin, should be carefully evaluated in clinical studies before considering it in the model.
Schafer et al. (2019) [16]	A total of 8 patients with idiopathic autism spectrum disorder (ASD) and five unaffected individuals.	A time-series approach was developed to track the cortico-neuronal development.	The approach was useful for future mechanistic investigations that try to find the convergence of genetic variants that contribute to ASD risk.	-
Mathematical and Statistical Modelling				
Hansen et al. (2001) [17]	A variety of quantitative genetic data	The model captures directly epistatic effects measurable in a reference genotype.	Has many useful mathematical properties and can represent epistatic effects of any order and on multiple phenotypic traits.	The multilinear model may be rejected by Biologically significant non-linearity.
Wu et al. (2003) [18]	Sets of genes in gene expression data collected by microarrays.	Graphical Gaussian model and log-linear model.	The approach complements the typical clustering approaches that are used to analyze microarray data.	The model requires further research, such as how to better deal with sparse data when either structural zero cells are present or if it contains many small cell values.
Patton,et al. (2014) [19]	A set of Arabidopsis thaliana gene expression data and seven sets of simulated data.	They used posterior probabilities for network features that are based on multiple hierarchical replications.	Provides a very useful tool to the biological community to help identify potential unrecognized relationships in genome-wide transcript abundance datasets.	Higher signal partial correlation may impact the model.
Expression-Co-Expression Network Modelling				
Kanakoglou et al. (2020) [23]	ENA datasets	The model is based on a Differential Gene Expression Analysis (DGEA).	The model is able to investigate the effects of high dose ionizing radiation on healthy human tissue using quantitative analysis of gene expression.	An incorrect base call may affect the results of the model.

Liu et al. (2020) [24]	6 unprocessed datasets for gene expression profiles from the Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo/) including: GSE6004, GSE58545, GSE27155, GSE53157, GSE60542, and GSE33630.	The use of gene co-expression network.	The first model that studies PTC using a combination of RRA and WGCNA. The model also provides new insights into current diagnosis and pathogenesis for PTC.	More studies are needed to validate these research results that may be biased due to the small sample sizes.
Care et al. (2019) [25]	The Cancer Genome Atlas (TCGA), cancer samples for (BRCA, CRC).	4. Gene Correlation Network Analysis (PGCNA).	The model provides an approach to extract useful networks that can be effectively applied to diverse clinical and experimental datasets.	The model may have an issue when having large-scale data.
Correlations/Co-Expression Network Modelling				
Nayak et al. (2009) [26]	Catalogue of Published Genome-Wide Association Studies (http://www.genome.gov/26525384) (http://www.ncbi.nlm.nih.gov/pubmed/ Entrez Gene information bases (http:// www.ncbi.nlm.nih.gov/locales/entrez	The model used correlations in expression levels of more than 8.5 million human gene pairs in immortalized B cells to infer gene co-expression networks.	The co-expression networks were offered information on the role of human genes in disease and normal processes.	Some genes were paired randomly as opposed to being paired based on correlation patterns.
Michalopoulos et al. (2012) [27]	Affymetrix's Human Genome U133 Plus 2.0 Array Chip, Simple Omnibus Format in Text (SOFT).	Human Gene Correlation Analysis (HGCA) (co-expression-based).	The model was a powerful tool for discovering genes associated with similar functions based on their co-expression patterns.	-
Gene-Gene/Protein-Protein Interactions Networks Modelling				
Santibáñe et al. (2020) [28]	Gene Expression Omnibus, https://www.ncbi.nlm.nih.gov/bioproject/, and EcoCyc, https://ecocyc.org/	Developing software (Atlas) that converts genome graphs and gene regulatory, interaction and metabolic networks into dynamic models.	Their model can evaluate silico modifications, such as gene knockouts. The model also could be applied to the dynamic modelling of natural and synthetic networks of any bacteria.	Gene regulatory networks are static models, and dynamic models are difficult to obtain due to their size, complexity, stochastic dynamics and interactions with other cell processes.

Doncheva et al. (2018) [29]	STRING database.	The model is based on a protein-protein interactions network and developed in the form of software.	The app in this research supported several types of queries to retrieve information using the Cytoscape tool.	Issues in visualizing the MS-based proteomics data. Another issue is related to data on post-translational modifications.
Goh et al. (2007) [30]	The list of disease genes, disorders and associations between them, from Online Mendelian Inheritance in Man (OMIM).	The model used the concepts of a bipartite graph.	The model offers a rapid visual reference of the genetic links between disorders and disease genes.	The incompleteness of the OMIM and some noise in the data.
Schwab et al(2020) [34]	This article used application examples and guidelines to work with Boolean network models.	Boolean Networks modelling.	The network model was used to uncover regulatory interactions leading to protein overexpression in cancers.	There is an issue related to matching the timing of the model to the real biological system, which may lead to unrealistic results.
Mignone et al. (2020) [37]	Gene Expression Omnibus (GEO) (www.ncbi.nlm.nih.gov/geo/), this adopted the dataset includes 6 different organs (liver, brain, lung, heart, marrow, skin, and bone).	The model is based on a transfer learning approach.	The model was able to exploit the knowledge about a source GRN to improve the reconstruction of a target regulatory network, it was also able to exploit a large number of unlabeled examples.	Limitations when big data is involved.
Vinayagam et al. (2016) [38]	Dataset S2 Dataset S1 Dataset S4 http://www.pnas.org" Atlas Program (TCGA) https://www.cancer.gov	The model characterized the structural controllability of a large directed human PPI network.	This network model allowed the classification of proteins as "indispensable", "neutral", or "dispensable" and the ability to identify drug targets and new disease genes.	The lack of information on the true functional form of the underlying dynamics of a complex biological network makes it more difficult for this kind of model to obtain accurate results.
Pio et al. (2020) [40]	SynTReN and DREAM5 datasets.	Causal and predictive Network model with community structure.	The model can predict the existence of unseen edges in the network. It also allows limiting further investigations on few promising gene interactions.	-
Lim et al. (2006) [41]	CGI-DCI http://www.cell.com/cgi/content/full/125/4/801/DC1/.	Protein-protein interactions network based on inherited cerebellar ataxias.	The model was able to uncover many previously unsuspected connections between the different ataxia-causing proteins.	It may have a high rate of false-positive interactions.

Reyna et al. (2020) [43]	ICGC/TCGA Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium.	A network model that is based on pathways, coding and non-coding cancer driver mutations.	A new component of known cancer pathways that are recurrently affected by both coding and non-coding mutations is identified by this model.	A large percentage of non-coding mutations in the cohort cannot be detected due to a lack of power, especially in cancer types with a limited number of patients. Another limitation was the limited availability of transcriptomic data.
Liu et al. (2010) [44]	The Genetic Association Database (allergies and infections) (http://geneticassociationdb.nih.gov/).	Gene-gene interaction network and pathways.	Investigated the candidate genes that have been linked to many complex genetic disorders. Also, investigated gliomagenesis' biological properties and the pathway maps for the understanding of glioma disease.	This study network may not represent the actual causal relationship between genes because many different gene-gene interactions are resulting from various cellular/experimental conditions.
Vázquez et al. (2003) [51]	Yeast PIN	Protein-protein interaction network.	The model was a first step in identifying the evolutionary dynamics leading to the development of protein functions and interactions.	It takes a lot of runs to reveal the average behaviour behind statistical fluctuations because the model follows a random-growth process.
Huang et al. (2018) [47]	21 human genome-wide interaction networks	Sparse composite network model with high efficiency and performance.	An important benchmarking process of 21 molecular networks.	-
Miryala et al. (2021) [50]	29P. mirabilis strains from the National Center for Biotechnology Information Genome.	Network model with network science measurements.	Provided a better understanding of the molecular basis of multidrug resistance mechanisms in P.mirabilis.	-

4. Conclusions

This work reviewed the literature in terms of the approaches used for modelling genetics data. Many methods and approaches have been adopted for modelling this kind of data. The paper started with the traditional approaches, then moved to mathematical and statistical modelling. Thereafter, the paper described the most popular approaches for modelling genetics data, which are network science approaches. This kind of approach was the focus of this work since it is currently considered an efficient way for genetic analytics. Network-based models have the ability to investigate the relations among

network objects and extract useful information that can be of benefit for genetics researches. Network modelling is not an easy task to perform since it needs a lot of attention when building a model (e.g., the accuracy of the data used and the strategy of creating nodes and edges). However, these models may have limitations due to a variety of reasons as discussed earlier. Therefore, when adopting a particular network model (e.g., Boolean model), the limitations of this kind should be considered as a step before adopting the approach. Also, the size of data should be considered because simulating large-scale data needs high-performance hardware that may not be available in some academic settings. Moreover, adopting a particular approach should take into consideration the availability and the ability of the software. Finally, we strongly believe that network science approaches are the most powerful tool for analysing genetics data especially when machine learning is involved (e.g., unsupervised learning). The use of such techniques will add a lot to the approach.

References

[1] Hedrick, P. (2011). *Genetics of populations*. Jones & Bartlett Learning.
[2] Benton, M. L., Abraham, A., LaBella, A. L., Abbot, P., Rokas, A., & Capra, J. A. (2021). The influence of evolutionary history on human health and disease. *Nature Reviews Genetics*, *22*(5), 269-283.
[3] Schmid-Hempel, P. (2021). *Evolutionary parasitology: the integrated study of infections, immunology, ecology, and genetics*. Oxford University Press.
[4] Del Vecchio, C., Verrilli, F., & Glielmo, L. (2018). Modelling and stability analysis in human population genetics with selection and mutation. *Mathematical Methods in the Applied Sciences*, *41*(4), 1492-1508.
[5] Hidirov, B. N. (2008). Mathematical and computer modelling regulatorika of hierarchical molecular genetic systems. *Scientiae Mathematicae Japonicae*, *67*(2), 229-240.
[6] Génin, E. (2020). 49th European Mathematical Genetics Meeting (EMGM) 2021. *Human Heredity*, *85*(2), 69-100.
[7] Rouzine, I. M. (2020). *Mathematical Modeling of Evolution*. De Gruyter.
[8] Basavarajaiah, D. M., & Murthy, B. N. (2020). Statistical Genetics and Its Application in Drug Trail. In *Design of Experiments and Advanced Statistical Techniques in Clinical Research* (pp. 179-211). Springer, Singapore.
[9] Scutari, M. (2015). Graphical Modelling in Genetics and Systems Biology. In *Foundations of Biomedical Knowledge Representation* (pp. 143-158). Springer, Cham.
[10] Costa, L. D. F., Rodrigues, F. A., & Cristino, A. S. (2008). Complex networks: the key to systems biology. *Genetics and Molecular Biology*, *31*(3), 591-601.
[11] Goh, Kwang-Il, Michael E. Cusick, David Valle, Barton Childs, Marc Vidal, and Albert-László Barabási. "The human disease network." *Proceedings of the National Academy of Sciences* 104, no. 21 (2007): 8685-8690.
[12] Vidal, Marc, Michael E. Cusick, and Albert-László Barabási. "Interactome networks and human disease." *Cell* 144, no. 6 (2011): 986-998.
[13] Huang, J., Niu, C., Green, C. D., Yang, L., Mei, H., & Han, J. D. J. (2013). Systematic prediction of pharmacodynamic drug-drug interactions through protein-protein-interaction network. *PLoS computational biology*, *9*(3), e1002998.
[14] Laird, N. M., & Lange, C. (2011). *The fundamentals of modern statistical genetics* (p. 167). New York: Springer. ISSN: 1431-8776
[15] Marok, F. Z., Fuhr, L. M., Hanke, N., Selzer, D., & Lehr, T. (2021). Physiologically Based Pharmacokinetic Modeling of Bupropion and Its Metabolites in a CYP2B6 Drug-Drug-Gene Interaction Network. *Pharmaceutics*, *13*(3), 331.
[16] Schafer, S. T., Paquola, A. C., Stern, S., Gosselin, D., Ku, M., Pena, M., ... & Gage, F. H. (2019). Pathological priming causes developmental gene network heterochronicity in autistic subject-derived neurons. *Nature neuroscience*, *22*(2), 243-255
[17] Hansen, T. F., & Wagner, G. P. (2001). Modeling genetic architecture: a multilinear theory of gene interaction. *Theoretical population biology*, *59*(1), 61-86.
[18] Wu, X., Ye, Y., & Zhang, L. (2003). Graphical modeling based gene interaction analysis for microarray data. *ACM SIGKDD Explorations Newsletter*, *5*(2), 91-100.

[19] Patton, K. L., John, D. J., Norris, J. L., Lewis, D. R., & Muday, G. K. (2014). Hierarchical probabilistic interaction modeling for multiple gene expression replicates. *IEEE/ACM transactions on computational biology and bioinformatics*, *11*(2), 336-346.

[20] Gstaiger, M., & Aebersold, R. (2009). Applying mass spectrometry-based proteomics to genetics, genomics and network biology. *Nature Reviews Genetics*, *10*(9), 617-627. https://doi.org/10.1038/nrg2633

[21] Thornton-Wells, T. A., Moore, J. H., & Haines, J. L. (2004). Genetics, statistics and human disease: analytical retooling for complexity. *TRENDS in Genetics*, *20*(12), 640-647. https://doi.org/10.1016/j.tig.2004.09.007

[22] Barabási, A. L., Gulbahce, N., & Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. *Nature reviews genetics*, *12*(1), 56-68, doi:10.1038/nrg2918.

[23] Kanakoglou, D. S., Michalettou, T. D., Vasileiou, C., Gioukakis, E., Maneta, D., Kyriakidis, K. V., ... & Michalopoulos, I. (2020). Effects of High-Dose Ionizing Radiation in Human Gene Expression: A Meta-Analysis. *International journal of molecular sciences*, *21*(6), 1938.

[24] Liu, Y., Chen, T. Y., Yang, Z. Y., Fang, W., Wu, Q., & Zhang, C. (2020). Identification of hub genes in papillary thyroid carcinoma: robust rank aggregation and weighted gene co-expression network analysis. *Journal of translational medicine*, *18*, 1-14

[25] Care, M. A., Westhead, D. R., & Tooze, R. M. (2019). Parsimonious Gene Correlation Network Analysis (PGCNA): a tool to define modular gene co-expression for refined molecular stratification in cancer. *NPJ systems biology and applications*, *5*(1), 1-17.

[26] Nayak, R. R., Kearns, M., Spielman, R. S., & Cheung, V. G. (2009). Coexpression network based on natural variation in human gene expression reveals gene interactions and functions. *Genome research*, *19*(11), 1953-1962.

[27] Michalopoulos, I., Pavlopoulos, G. A., Malatras, A., Karelas, A., Kostadima, M. A., Schneider, R., & Kossida, S. (2012). Human gene correlation analysis (HGCA): a tool for the identification of transcriptionally co-expressed genes. *BMC research notes*, *5*(1), 1-11

[28] Santibáñez, R., Garrido, D., & Martin, A. J. (2020). Atlas: Automatic modeling of regulation of bacterial gene expression and metabolism using rule-based languages. *Bioinformatics*.

[29] Doncheva, N. T., Morris, J. H., Gorodkin, J., & Jensen, L. J. (2018). Cytoscape StringApp: network analysis and visualization of proteomics data. *Journal of proteome research*, *18*(2), 623-632.

[30] Goh, K. I., Cusick, M. E., Valle, D., Childs, B., Vidal, M., & Barabási, A. L. (2007). The human disease network. *Proceedings of the National Academy of Sciences*, *104*(21), 8685-8690.

[31] Zha, H., He, X., Ding, C., Simon, H., & Gu, M. (2001, October). Bipartite graph partitioning and data clustering. In *Proceedings of the tenth international conference on Information and knowledge management* (pp. 25-32). https://doi.org/10.1145/502585.502591

[32] Shmulevich, I., Dougherty, E. R., & Zhang, W. (2002). From Boolean to probabilistic Boolean networks as models of genetic regulatory networks. *Proceedings of the IEEE*, *90*(11), 1778-1792. https://doi.org/10.1109/JPROC.2002.804686

[33] Cheng, D., Qi, H., & Li, Z. (2010). *Analysis and control of Boolean networks: a semi-tensor product approach*. Springer Science & Business Media. ISSN: 0178-5354

[34] Schwab, J. D., Kühlwein, S. D., Ikonomi, N., Kühl, M., & Kestler, H. A. (2020). Concepts in Boolean network modeling: What do they all mean?. *Computational and structural biotechnology journal*, *18*, 571-582.

[35] Davidson, E., & Levin, M. (2005). Gene regulatory networks. *Proceedings of the National Academy of Sciences*, *102*(14), 4935-4935. https://doi.org/10.1073/pnas.0502024102

[36] Karlebach, G., & Shamir, R. (2008). Modelling and analysis of gene regulatory networks. *Nature reviews Molecular cell biology*, *9*(10), 770-780. https://doi.org/10.1038/nrm2503

[37] Mignone, P., Pio, G., D'Elia, D., & Ceci, M. (2020). Exploiting transfer learning for the reconstruction of the human gene regulatory network. *Bioinformatics*, *36*(5), 1553-1561.

[38] Vinayagam, A., Gibson, T. E., Lee, H. J., Yilmazel, B., Roesel, C., Hu, Y., ... & Barabási, A. L. (2016). Controllability analysis of the directed human protein interaction network identifies disease genes and drug targets. *Proceedings of the National Academy of Sciences*, *113*(18), 4976-4981.

[39] Wang, D. J., Shi, X., McFarland, D. A., & Leskovec, J. (2012). Measurement error in network data: A re-classification. *Social Networks*, *34*(4), 396-409. https://doi.org/10.1016/j.socnet.2012.01.003

[40] Pio, G., Ceci, M., Prisciandaro, F., & Malerba, D. (2020). Exploiting causality in gene network reconstruction based on graph embedding. *Machine Learning*, *109*(6), 1231-1279 .

[41] Lim, J., Hao, T., Shaw, C., Patel, A. J., Szabó, G., Rual, J. F., ... & Zoghbi, H. Y. (2006). A protein–protein interaction network for human inherited ataxias and disorders of Purkinje cell degeneration. *Cell*, *125*(4), 801-814.

[42] Bossy-Wetzel, E., Schwarzenbacher, R., & Lipton, S. A. (2004). Molecular pathways to neurodegeneration. *Nature medicine*, *10*(7), S2-S9. https://doi.org/10.1038/nm1067

[43] Reyna, M. A., Haan, D., Paczkowska, M., Verbeke, L. P., Vazquez, M., Kahraman, A., ... & Raphael, B. J. (2020). Pathway and network analysis of more than 2500 whole cancer genomes. *Nature communications*, *11*(1), 1-17.

[44] Liu, Y., Shete, S., Hosking, F., Robertson, L., Houlston, R., & Bondy, M. (2010). Genetic advances in glioma: susceptibility genes and networks. *Current opinion in genetics & development*, *20*(3), 239-244

[45] Tunali, V. (2021). Large-Scale Network Community Detection using Similarity-Guided Merge and Refinement. *IEEE Access*. https://dio.org/10.1109/ACCESS.2021.3083971

[46] Schulz, M., Klipp, E., Uhlendorf, J., & Liebermeister, W. (2006). SBMLmerge, a system for combining biochemical network models. *Genome Informatics*, *17*(1), 62-71. https://doi.org/10.11234/gi1990.17.62

[47] Huang, J. K., Carlin, D. E., Yu, M. K., Zhang, W., Kreisberg, J. F., Tamayo, P., & Ideker, T. (2018). Systematic evaluation of molecular networks for discovery of disease genes. *Cell systems*, *6*(4), 484-495

[48] Mahmood, B. M., Sultan, N. A., Thanoon, K. H., & Khadhim, D. S. (2020). Collaboration Networks: University of Mosul Case Study. *AL-Rafidain Journal of Computer Sciences and Mathematics*, *14*(1), 117-133. https://doi.org/10.33899/csmj.2020.164679

[49] Dabdawb, M., & Mahmood, B. (2021). On the Relations among Object-Oriented Software Metrics: A Network-Based Approach. *International Journal of Computing and Digital Systems*. http://dx.doi.org/10.12785/ijcds/100182

[50] Miryala, S. K., Anbarasu, A., & Ramaiah, S. (2021). Gene interaction network approach to elucidate the multidrug resistance mechanisms in the pathogenic bacterial strain Proteus mirabilis. *Journal of Cellular Physiology*, *236*(1), 468-479

[51] Vázquez, A., Flammini, A., Maritan, A., & Vespignani, A. (2003). Modeling of protein interaction networks. *Complexus*, *1*(1), 38-44.