State-of-the-art in Candidate Disease Genes Prioritization and Prediction Approaches and Techniques

Fahd N. Al-Wesabi¹, Ghada M. Alqubati² and Abdulmajed Alkhuliadi²

¹College of Science and Arts in Muhayel Asir, King Khalid University, KSA. Faculty of Computer and IT, Sana’a University, Sana’a, Yemen; fwesabi@gmail.com
²Faculty of Computer and IT, Sana’a University, Sana’a, Yemen; ghada.alqubati101@gmail.com

Abstract

Objective: Candidate disease genes identification has grasped the attention of many researchers for its significant role in bioinformatics. In this review, we demonstrate several classifications of some recent identification approaches and their datasets. Methods/Findings: The approaches are classified into five categories and the datasets into two categories. Some categories are also classified into several types. In each category, we explain every approach based on its objectives, mechanism, datasets and results. Different algorithms have been used such as random walk algorithm, machine-learning algorithms or genetic algorithm. Furthermore, the common approach followed to test the performance is cross-validation approach using precision, recall and F1-metrics. During our research, we found a novelty of many methods and a noticeable improvement in some networks and algorithms. We also noticed that the major emphasis was to enhance genome datasets using different mechanisms such as integrating them or adding new features. We noticed that most researchers focus more on this aspect as they believe that the best way to improve genes prioritization and identification and get more accurate results is to have a reliable dataset including all required information. Application: This survey can be a valuable source of information. It explains and summaries every item in the classification in a simple and understandable way. Therefore, it can be used by researchers concerning with disease genes identification as it can enlighten and guide them to different techniques and dataset used in this subject.

Keywords: Biological and Topological Properties, Disease Gene Identification, Gene Expression, Gene Ontology, Phenotype, Protein Interaction Networks

1. Introduction

Nowadays, disease candidate genes prediction is a critical part of biomedical research. This is because of the many breakthroughs it could lead in medical diagnostic and therapeutic. In spite of the many computational approaches that have been developed for ranking, prioritizing, and detecting disease candidate genes, they remain a big challenge. In complex diseases, many genes work together to provoke a disease phenotype and understanding the link between these genes and disease offers a better opportunity to understand disease etiology, which allows for a development in treatment strategies.

However, even if there was a good understanding of the genetic basis of a disease, interpreting this link at the...
molecular level to understand the mechanisms leading to the disease remains a challenge in biomedical science. Besides, the limited number of known disease genes compared to the huge set of unlabeled or unknown genes makes it difficult for many methods such as machine learning techniques to learn from this limited set and to produce unbiased set of candidate genes. Therefore, in order to enhance this limited knowledge, several hybrid techniques and networks have been presented. In addition, many methodologies have tried to understand the mechanism of these known disease genes and the complex interplay between them and their proteins. They integrated variety molecular networks such as gene expression, genetic linkage, protein-protein interactions and gene-phenotype associations. In this survey, we talk about the classification of CDGPP approaches and their datasets. In the first section, we explain the approach we followed for classification. In the second section, we illustrate a number of different approaches for complex, non-complex, specific and general disease associated genes identification. We divide these approaches into five categories based on the techniques used to enhance and exploit datasets and develop more efficient ranking algorithms. For each approach, we explain its objectives, methodology and results. In the third section, we divide genomic datasets into two categories. The following section is a summary of all approaches and their datasets. We have noticed that the main interest of most of those approaches was to enhance existing methods and datasets. Generally, we can say that all of them are mutually correlated, and we believe that more methods in future would be developed taking these approaches as their ground base.

2. Literature Survey

In spite of the many computational approaches that have been developed for ranking, prioritizing and detecting disease candidate genes, they remain a big challenge. The limited number of known disease genes compared to the huge set of unlabeled or unknown genes makes it difficult for many methods such as machine learning techniques to learn from this limited set and to produce unbiased set of candidate genes. Therefore, in order to enhance this limited knowledge, several techniques and algorithms have been presented and some of them proposed approaches to improve the positive or the negative dataset used by semi-supervised methods and others offered different algorithms to enrich various data sources. For instance, a novel algorithm applied for feature selection in gene expression dataset and other methods proposed to measure semantic similarity in gene-phenotype networks and gene ontology dataset. In addition, many methodologies have tried to understand the mechanism of these known disease genes and the complex interplay between them and their proteins. They integrated a variety of sources such as gene expression, genetic linkage, protein-protein interactions and gene-phenotype associations and developed hybrid approaches with better results. The major focus of these approaches is mostly based on the enriching of current genome datasets and all related sources and algorithms. Therefore, the utilization of computational approaches complemented by these approaches can greatly develop the study of disease genes and. The main challenge in candidate disease genes identification is to find an effective model able to incorporate different molecular networks and elicit the exact knowledge required to produce a more reliable set of candidate disease genes.

3. Methodology

There are many approaches proposed for candidate disease genes identification. They differ according to many factors, such as target diseases, algorithms, datasets, etc. We have collected a number of recent approaches and analyzed every approach thoroughly using different criteria.

We focused in our analysis on four aspects, approach’s objectives, its implementation process, its dataset/s and its findings. Based on those items we classified those approaches and their datasets into a number of categories. In the approaches classification section, we focus on the technique followed on each category. For example,
some techniques aim to optimize the given datasets by refining them from redundancy and impurity and other techniques optimize datasets by enriching them with a number of genetic features. Based on those techniques, we define five categories. In each category, we explain its types if any and its approaches. For every approach, we give a brief explanation about its objectives, methodology, dataset and experimental results. In the datasets classification section, we focus on the types used in those approaches and also the techniques used in some of them to build a heterogeneous dataset. We define two datasets categories, homogenous and heterogeneous. In each category, we mention a number of classes in which each class includes datasets types, their sources and their components. Finally, we sum up all those approaches in a review table comparing all approaches based on their purpose, method, dataset and results.

4. Classification of CDGPP Approaches

Approaches in this section focus on achieving more effective candidate disease genes identification through refining or enriching particular datasets, integrating different datasets and developing new methods, techniques or algorithms.

4.1 Dataset Refinement Approach

Some disease gene identification approaches focus on cleansing a given dataset to get rid of all the redundant data affecting the accuracy of the results. A highly data redundancy of gene expression could impact the result of gene expression process such as clustering, which is one of the pivotal ways for cancer genes identification or classification. Therefore, proposed a hybrid feature selection algorithm called MIMAGA-selection algorithm. This algorithm aims to reduce gene expression's redundant data and dimension in order to improve the process of gene expression clustering and classification. MIMAGA-selection algorithm integrates two feature selection algorithms. The first algorithm is called Mutual Information Maximization algorithm (MIM) which is used for genetic filtration by measuring the information dependency of a random sample over other samples. The second algorithm is called Adaptive Genetic Algorithm (AGA) which has two operators, crossover and mutation. Crossover operation has a global search capability to produce new global solutions, whereas, mutation operation has a local search capability. The mutation and crossover probability variables determine if GA algorithm is able to find the optimal solution. In the adaptive Genetic Algorithm, those variables are predefined to get more accurate values. In MIMAGA-Selection, chose Extreme Leaning Machine (ELM) as the classifier and the measure for its accuracy was the fitness of AGA Fitness becomes the ELM's classification accuracy. The compared results of the proposed algorithm with other classifiers show its outperformance and efficiency in dimension and redundancies reducing of gene expression data.

4.2 Dataset Enrichment Approach

One of the obstacles in disease genes classification and predication is the lack of a reliable dataset containing all necessary information needed for more precise results. Hence, some approaches seek to fill the gaps in a given dataset and enrich it with more information that can help extract the similarity and difference aspects between different genes, which can pave the way for more successful disease identification. Every gene has many characteristics that can be used for enrichment. We mention two methods based on two features as follows:

4.2.1 Topological and Biological Similarity Enrichment

In suggested an approach to increase the efficiency of ranking candidate disease genes. Their approach depends on one dataset rather than heterogeneous one, which is Protein-protein Interaction (PPI) network. Their goal is to enhance the dataset in order to gain a better understanding of proteins functionality and their contribution to diseases. They believe that PPI networks are an efficient source for disease genes identification, but they need for more biological knowledge of individual proteins to become better. Therefore, they first collected the
keywords associated with each protein from the Universal Protein Resource (UniProt) database. These keywords describe the various biological mechanisms of the proteins. After that, they constructed a PPI-Keyword (PPIK) network by using these keywords. The protein and its keyword were represented in different nodes. PP IK network was modeled by an undirected graph consist of a set of nodes and a set of edges and the label function of the nodes that determine whether the node is protein or a keyword. Furthermore, suggested including the topological similarities to their method because they believe that proteins with similar roles could share similar topological features. However, the traditional structure of module-based methods that model topological similarities do not differentiate between different labels of nodes. For that reason, they use metagraphs because allow for common structures representation on a heterogeneous graph, where nodes with different labels are connected with each other. Based on these metagraph representations, they built classifiers by means of supervised learning. To evaluate the performance of the proposed representation and the classifiers chose three different human PPI databases, namely IntAct, NCBI and STRING. Disease labels for proteins were obtained from the UniProt and OMIM databases. They use ROC Curve (AUC) measure, which is a robust measure of the classifiers' predictive power, to evaluate the effectiveness of the proposed method, and the results were positive.

4.2.2 Semantic Similarity Enrichment

In have developed an approach for deciphering disease gene association. They used a novel measure called HeteSim. In their study, they argued that despite the success of network-based algorithmic approaches, most of them did not consider the semantic differences behind the paths between network nodes such as gene-phenotype heterogeneous networks, which are multisource networks representing the relationship between a person genetic characteristics and physical characteristics. They believe that disregarding these differences affects the credibility of the resulted candidate genes. Therefore, they suggested a path-based measure to calculate these differences between objects in heterogeneous network. The measure is efficient at recognizing the semantics between different paths in order to accurately identify the affiliation between genes and human phenotypes. They also proposed two multipath methods using HeteSim measure in two different methods as follows:

4.2.2.1 HeteSim MultiPath (HSMP) method

There are different paths connecting two objects in heterogeneous networks. Every path has different semantic meanings. For example, gene-human phenotype-human phenotype path differs from gene-gene-human phenotype. The first path means if there are two similar human phenotypes and one of them is associated with a gene, then the other one will have high possibility to be also associated with that gene. The latter path means if a gene is associated with human phenotype, then another similar gene will be potentially associated with that human phenotype. Hence, these semantic differences are significant in determining the similarity between human phenotype and different genes, which is a measure used to identify candidate disease genes with known phenotype properties. HSMP uses HeteSim to measure similarity between objects in heterogeneous networks. A constant $\beta$ is added to the HeteSim path scores to diminish the contributions from longer paths as authors found that short path with length less than five could contribute more than a long path and after comprehensive searching, they found that the constant $\beta$ could be used for further association prediction in their experiments.

4.2.2.2 HeteSim SVM (HSSVM) Method

Similarly, HSSVM method uses HeteSim measure. However rather than using a constant, it uses a machine learning method. The reason of using machine learning in this method is that different paths make different contribution to the relevance score, thus, the weight of this contribution is measured by a machine learning method. In this method, the associations between genes and human phenotypes are used as the positive set and gene-phenotype pairs for which no associations existed are used as the unlabeled set. Then the HeteSim scores are
used for each feature based on 66 constrained paths that were used to construct 66 features for each gene-phenotype pair.

After the comparison of the two methods performance with four other methods, Katz, CATAPULT, PRINCE and ProDiGe, found that the two methods had outperformed all of the four methods. They also found that the overlap ratio between the associated phenotypes and phenotype of the top 10 predicted genes of the disease phenotypes, which were reported to contain a very high degree of overlap, was lower comparing to other methods ratio.

Another approach is proposed in based on Gene Ontology (GO), which contains structured vocabularies representing several features and properties about genes. Gene ontology can be used to enrich genomic data with through GO annotation. However, the process of genes annotation by using those vocabularies is not an easy task. Hence, have proposed a semantic method called Gene Ontology Hierarchy Preserving Hashing (HPHash). The goal of this method is to accurately predict the semantics similarity between a huge number of GO terms and given genes and this is by preserving the GO terms hierarchy captured by HPHASH. After annotating genes and creating a matrix of compressed gene terms, by using a number of hash functions, the method uses these terms to define the semantic similarity between genes to and predict functions of a gene based on annotations of its semantic neighbors. These predictions are then mapped back to the original GO terms space, thus the associations between genes and massive GO terms are eventually made. To evaluate the ability of HPHash to replenish GO annotations of partially annotated genes, chose some semantic similarity measures and gene function prediction methods. They compared HPHash performance with Baseline (Ham), Baseline (Cos), InterGFP (BMA), clusDCA (G), HashGO and HPHash (G), clusDCA and HashGO. They applied eight evaluation metrics including MicroAvgF1, MacroAvgF1, RankingLoss, AvgPrecision, AvgAUC, MCC (Matthew correlation coefficient), Fmax and Smin. They used two GO annotations files of three species in different dates.

They used the old file in 2016 to train gene function prediction models and the updated file in 2017 to test HPHash prediction performance. The experimental results demonstrated that HPHash excels all other methods in all evaluation metrics for all the three spices. Moreover, it surpassed other methods in the ability to predict completely un-annotated genes, whereas other methods cannot predict functions for genes whose annotations are completely unknown, since the similarity between a completely un-annotated gene and other genes is zero.

4.3 Reliable Heterogeneous Dataset Construction Approach

Some approaches start with combining several datasets to construct one reliable heterogeneous dataset and then they develop a ranking method and apply it on the new dataset to classify disease genes and achieve better result.

In suggested using a genomic data extracted from Human brain tissue called expression quantitative trait loci (eQTLs). Besides, they used the human brain data including 25,866 significant SNP-gene association pairs of 3709 genes from see QTL database to construct GGCRN. For every gene, they extracted SNPs that regulate it and called them significant related SNPs. If one SNP regulates two genes, they called it a common SNP. They also considered two genes to be co-regulated if a specific proportion of SNPs regulates both genes. They identified 181,906 co-regulated gene pairs of 2830 genes with nonzero co-regulation and used them to build GGCRN. After that, they combined GGCRN with Protein–protein interaction network, known as Human Protein Reference Database (HPRD). Then they applied Random Walk with Restart algorithm (RWR), which is random walk that allows for the restart of the walk at every time it steps at source node s with probability r, on the three networks, the HPRD PPI network, which describes interaction networks in the human proteome, the GGCRN and the Union network and compare the results. They found that applying the RWR algorithm to identify associated genes
with Alzheimer disease on the integrated network had achieved better results than applying it on (HPRD) PPI network only. After performing a numerical experiment, they added a stop condition of walking no more than 100,000 times. They considered candidate genes that have a steady-state probability greater than the initial known disease genes. After performing RWR, they found that the results of RWR were affected by the restart probability, r. The results showed that when r was zero, they got too many candidate genes and when r was more than 0.1, they obtained too few candidate genes. Then, they set r to 0.01, 0.015, 0.02, 0.025, 0.03, 0.035, 0.04, 0.045 and 0.05, but still did not know exactly how many disease genes existed. Therefore, they defined a standard for mining risk genes of mining less than 2 times of the known initial genes. They examined 14 initial genes in the HPRD PPI and Union networks and 4 initial genes in the co-regulation network. The results showed that with the co-regulation network only 41 candidate genes were extracted with r set at 0.015 and 2 candidate genes with r set at 0.02. Hence, because GGCRN is small, 4 initial genes were used, using it only is inapt for mining process. Additionally, the compared results of the HPRD PPI and Union network demonstrated that with r set at 0.01, the numbers of candidate genes extracted were still insufficient. Whereas, with an r value larger than 0.015, the number resulted by using HPRD PPI network is higher than the number resulted by using Union network. This draws a conclusion that HPRD PPI needs for a higher restart probability, r, to get same results as the union one. Therefore, using Union network gets better results and the GGCRN constructed by eQTL data is a useful resource for mining disease genes.

The proposed method presented in\textsuperscript{13} was developed to Prioritize Disease Risk Genes (PDRG) in a disease related metabolic network. In\textsuperscript{13} suggested enriching the network by incorporating three types of data, gene expression, protein interactions and functional annotations. In\textsuperscript{13} first constructed disease-related metabolic network. The network consists of a number of known disease genes and their neighbors, considered as candidate genes. Furthermore, the network contained a total of 5776 proteins and 589,199 interaction relationships. Then they developed prioritization algorithm named PDRG (Prioritize Disease Risk Genes) to compute the disease risk score of genes in the disease-related metabolic network by utilizing three resources: Interaction, expression and function data, calculated the score of similarity between every gene and its neighbor node in the constructed network. After that, they combined the similarity scores of the three resources taking the relative importance of each resource into consideration. Finally, they calculated the disease risk score of every gene based on the combined similarity and the expression difference of each gene in normal and disease samples. The PDRG results were compared to the results of two other methods, ToppGene and ToppNet and the comparison showed a better performance of PDRG than the two methods. Guo et al. believe that their method could be applied successfully to other metabolism-related diseases, which in turn will help to achieve accurate diagnosis and effective therapies.

In\textsuperscript{11} have proposed an improvement for their previous two disease genes prediction algorithms as they had a limitation in their results. The first algorithm was called RWRDP. It is a random walk with restart method used to classify the candidate genes in a given PPI network through detecting the similarity in their diffusion profiles with disease genes by means of one of two indicators, Cosine Angle (COS) or Linear Correlation Coefficient (LCC). The algorithm has an iterative mechanism. It transfers from a node to a neighbor node with same probability. At each step, the random walk has a probability to return to the start node and restart. The nodes that have high COS or LCC values will have higher probability to be classified with the disease. The second algorithm was called RWRHN. This algorithm is based on a random walk approach. However, it aims to enrich the given dataset with more information to build a reliable heterogeneous network in order to achieve better results. It applies random walk algorithm on the original PPI dataset and it uses two resistances variables, α and β and a correlation coefficient to build a matrix including the topological similarities between genes. To construct the
reliable heterogeneous dataset, the PPI network at first is reconstructed by using that matrix. Then the resulted network is combined with a phenotype similarity network and the identified associations between diseases and genes. In [11] at first tried to improve the first algorithm by adding a function called Least Square Linear Regression (LLR), to enhance disease genes prediction. They called the new algorithm by RWRDPPLL. The proposed function is applied as a prioritizing function to compute LLR of the diffusion profiles of disease genes and other set of genes in PPI network. However, according to observations, the linear regression function showed some limitation in ranking candidate genes. In [11] proposed then suggested an improvement for the second algorithm. In their new method called RWRHN-PLR-SS they applied piecewise linear regression to RWRHN to build four matrices of protein sequence, topological, phenotype and disease-genes association similarities. After that, they applied RWRHN on these matrices for ranking. During the experimental stage, [11] split the dataset into six splits and used accuracy, precision, recall, F-Measure, prediction error, AUC and ROC values for the comparison. For all dataset splits, the results showed that the last proposed algorithm had achieved higher performance compared to RWRDP, RWRHN and RWRDPPLL algorithms. For example, for the 200 dataset split, the proposed RWRHN-PLR-SS algorithm performed 50% better than RWRDP, 14.58% better than RWRHN and 39.58% better than RWRDPPLL. Besides, according[11], a further improve can be done to the proposed RWRHN-PLR-SS algorithm by integrating it with more information such as, biological and gene expression information.

4.4 Semi-supervised Learning Approach

In [6] have proposed a new method based on PU-learning. They argued in their paper that disease genes that share the same phenotype features tend to have similarity in their biological features.

They also claimed that the associated genes are more likely to have similar phenotype or functional properties as their diseases genes to which they are connected. Hence, new diseases can be derived from the previous one by using machine-learning methods. The proposed method uses gene expression profiles and applies HMM, which is a probabilistic model for probabilistic distribution representation. The structure of the model was comprised of two stages: The learning stage and the predicting stage. The learning stage consisted of four steps. First, a combination of K-means++ and K-means algorithms was used to build the clustering model for each disease by using a matrix built from measuring semantic similarity between disease types based on a gene ontology. Second, discretion genes expression levels was applied for each disease by using K-mean algorithm for a sophisticated vector quantization. The first and the second step were executed simultaneously. Third, mapping the quantized expression profiles was achieved for each disease to their corresponding disease type clusters. Gene expression profiles in the last step of the learning stage are interpreted as an observation sequence in HMM models in which a model per cluster of each disease is learned. Furthermore, the corresponding threshold of each learned HMM is calculated from the training set of each cluster to boost the prediction process. In the prediction stage, the output of the second step of the learning stage was used to map unlabeled genes with the closest value. Besides, both learned HMM models and their threshold predict unlabeled genes. The proposed method achieved better results, ranged from 2% to 18% more accuracy, compared to other methods, such as PEGPUL and EPU methods.

Another diseases genes identification approach using PU-learning technique was proposed by Vasighizar and Jalili in [3]. They developed a method called PUGP to improve the performance of machine learning technique by building a reliable set of non-disease genes called the negative set. In their proposal, they followed a three-step approach. The first phase is for the extraction of the reliable negative data. In [2] argued that using a more reliable set is better than using randomly unknown genes as negative data. In this step, C-PUGP uses the clustering, with a predefined number of clusters and One Class classification approach to extract the reliable negative genes set called \{RN\} from unlabeled set called \{U\}. It calcu-
lates the similarities between the unlabeled genes in \{U\} and the positive genes in the set called \{P\} based on the idea that those genes in \{U\}, which are very dissimilar to the genes in \{P\}, are likely to be reliably negative. In the second phase, \(^2\) considered the set \{P\} as the positive or disease genes training set and the set \{RN\} as the negative or non-disease genes training set and used them to learn some binary classifiers by using several techniques such as decision tree and SVM. In the final phase, called the prioritization phase, the best-trained classifier was used to rank or label the unlabeled set and define the positive set including the candidate disease genes. The scoring function used to measure the similarity degree between the unknown genes and known disease genes was learned by SVM method. In\(^2\) used two evaluation strategies, which are Leave-one-out Cross-validation (LOOCV) and 10-fold Cross-validation and the results revealed that the proposed method was better in all of the evaluation metrics.

4.5 Hybrid Algorithms and Multi-dataset Approach

In\(^1\)\(^2\) that incorporating different dataset resources and methods leads to better results when ranking candidate genes. Based on this assumption, they explored several datasets, which they called multi-genomic datasets, such as gene ontology, protein sequences, protein-protein interaction networks, biological pathways and they chose seven best classifiers for disease genes classification. They developed software for ranking disease genes called GPS. The mechanism of GPS consists of two steps. In the first step, the seven rankers are applied on different reliable datasets to classify disease genes based on different features such as functional or biological similarity, pathway similarity or semantics similarity.

In the second step, GPS gathers all results of these local rankers and applies a global ranker to get an overall outcome. Besides, in the assessment step, GPS achieved better results compared to some other methods such as DIR and GPEC.

5. Classification of Genomic Datasets

There is a wide range of genomic datasets. We can divide them into two categories as follows:

5.1 Homogenous Dataset

Homogenous dataset is a database including a collection of genome data of the same type. We can mention four types as follows:

5.1.1 SNP Database

One promising method for studying the genetic basis of complex disorders is a genome-wide approach to gene mapping known as genotyping. A person’s genotype refers to his or her own arrangement of the DNA letters, A, T, C or G, in a particular region of their genome. The arrangement may be different from one person to the next. If one person is sick with an inherited disorder and another is not, the cause may lie in differences in a particular spot in their DNA. This can be referred as genetic variations, which are commonly known as Single Nucleotide Polymorphisms (SNPs) and are considered as valuable resources for disease genetic research\(^23\). Number of SNPs has become available. For example, dbSNP is a free database for genetic variations and it is available in three formats: ASN.1, Fasta and XML. Moreover, SNP data can be obtained from Trust Case Control Consortium (WTCCC)\(^22,23\).

5.1.2 Protein-protein Interaction Network

One of the pivotal resources of genetic information is protein-protein interaction networks\(^24,25\). The information derived from the interaction between proteins is essential to decipher the essence of diseases in molecular level\(^26\). The fact that proteins connected with a specific disease are expected to have a deep interaction between them makes this type of information greatly valuable\(^27,28\) and tracing these proteins (gene products) could lead to the genes associated to disease emergence\(^3\). The interac-
tion between proteins is represented mathematically in the network. Many sources offer PPI datasets such as STRING database, which is an online web server that integrates a variety of resources from prediction and known interactions, and the Human Protein Reference Database (HPRD).

5.1.3 Protein Families Dataset
As most existing PPI networks are noisy and incomplete\textsuperscript{29,30} and proteins with the same protein domain as a disease protein could be associated with the disease as well\textsuperscript{29 and 30}. Pfam database is comprised of protein families of same domains from Pfam website.

5.1.4 Gene Expression Dataset
Gene expression data is the process by which the instructions in DNA are converted into a functional product, such as a protein. The proteins are gene products. Single strand of DNA contains thousands of genes. The size of each protein is defined from the length of their genes, while their function is determined from their shape. In other words, different protein shapes means different functionality.

5.2 Heterogeneous Dataset
Another category of genomic dataset comprises of different types of datasets collected together to construct one reliable dataset. We can mention some of them as follows:

5.2.1 Phenotype-protein-protein Interaction Dataset
The heterogeneous network\textsuperscript{5} used in their experiment was constructed from three different networks: 1. Protein-protein interaction PPI, 2. Gene-phenotype associations and 3. Phenotype-phenotype. These three networks were first collected from different sources as follows:

Gene-phenotype association network: The gene-phenotype associations were collected from nine different species: Human, Plant, Worm, Fruit fly, Mice, Yeast, Escherichia, Zebra fish and Chicken. The dataset was collected from different literature and public databases.

Phenotype-phenotype network: The dataset was extracted from MimMiner, which is a text-mining approach to evaluate the similarities between human phenotypes from the OMIM database.

After collecting networks\textsuperscript{5}, constructed the union network by connecting the gene interaction network and phenotype similarity network using the bipartite graph of the gene-phenotype association.

5.2.2 PPI-genomic Similarity Dataset
In\textsuperscript{11} have proposed an improved ranking method on a heterogeneous dataset. The dataset was built by fusing PPI network downloaded from HPRD with topological protein similarity, phenotype similarity network extracted from MimMiner, identified diseases-genes associations collected from OMIM database and protein sequence similarity extracted from PROSITE database.

5.2.3 Disease-oriented Dataset
Some datasets are constructed to identify a particular class or category of disease genes or proteins. For example, In\textsuperscript{13} has proposed a method to prioritize metabolism-related diseases such as obese genes in a disease related metabolic network. To construct the network, included different types of genomic information: Gene expression, protein interactions and functional annotations. The obesity gene expression profile was downloaded from the Gene Expression Omnibus (GEO), which contains 12 obese samples and 11 controls. One hundred and ninety-five obesity disease genes were identified in the Online Mendelian Inheritance in Man (OMIM) and (DO) databases. The information of protein interaction was downloaded from (STRING), an online web server that integrates a variety of resources from prediction and known interactions.
Metabolism-related basic metabolites, proteins and reactions were extracted and organized from several large data platforms (HMDB, HumanCyc, EHMN, Reactome, BioGRID and KEGG).

6. Comparative Study

In this section, we present a comparison among various disease genes identification approaches based on their goal, dataset, method and performance outcome as it is shown in Table 1.

7. Results and Discussion

Many researchers have conducted different identification approaches and all of them share one goal, which is achieving more reliable and accurate results helping to predict and identify many complex diseases. Most of these approaches seek to enhance genome datasets as many researchers believe that the lack of a reliable dataset including all genetic information required for a correctly identification of associated disease genes is the main challenge in this area of bioinformatics. A wide range of approaches have been proposed to optimize datasets through integrating different data sources to construct a reliable heterogeneous dataset or through enriching a given dataset with more features or refining it from unnecessary data affecting the results. From the approaches we have analyzed, we found that researchers usually start with constructing the genomic dataset and then suggest a technique or algorithm using the constructed dataset in

<table>
<thead>
<tr>
<th>No</th>
<th>Approach</th>
<th>Goal</th>
<th>Dataset</th>
<th>Methods</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>The HeteSim measure for semantic differences of paths between different objects.</td>
<td>Improve candidate genes identification.</td>
<td>Heterogeneous network made of protein-protein interaction, gene-phenotype associations, and phenotype-phenotype.</td>
<td>HeteSimMultiPath method with β constant and HeteSim SVM (HSSVM) method with machine learning.</td>
<td>The two methods outperformed other four methods PRINCE, ProDiGe, Katz, CATAPULT and PRINCE in the cross-validation test and overlap ratio test.</td>
</tr>
<tr>
<td>2</td>
<td>Hidden Markov Model (HMM) based PU-learning using gene expression profiles.</td>
<td>Develop a novel method for semi-supervised learning to improve the prediction process of candidate disease genes.</td>
<td>DNA microarray gene expression profiles (GEP), Gene Ontology (GO) and GENECARD and OMIM disease genes data.</td>
<td>K-means++ and K-means algorithms to build the clustering model, K-mean algorithm for gene expression profile quantization, and learned HMM models and their threshold to predict unlabeled genes.</td>
<td>10-fold cross validation technique to compare the precision of the proposed method with other 7 methods (Xu’s, Smalter’s, ProDiGe, PUDI, EPU, SFM and PEGPUL) shows the method’s outperformance.</td>
</tr>
</tbody>
</table>
### C-PUGP: A Novel Positive Unlabeled Learning Technique based upon Clustering and One-Class Classification Algorithm

Overcome the lack of negative set in the machine learning methods by constructing a reliable negative set and applying a semi-supervised method to predict and rank candidate disease genes more efficiently. Disease gene datasets, positive set, and unlabeled set. In addition, the dataset consists of some significant features extracted from proteins sequences.

A three steps approach to rank candidate disease genes by using clustering technique and one-class classification method. The first step is for constructing the reliable negative set of non-disease genes, the second step is for training the classifiers based on first step outcome, and the final step is to predict the correct labels of unknown genes and extract the candidate disease genes set.

10-fold cross-validation and leave-one-out cross-validation (LOOCV) evaluation tests are implemented, and the experimental outcomes indicate that the proposed method produces better results. Compared to the existing methods, the increase of performance of the method is 11.7% better than the best method in terms of F-measure. Additionally, results show about 6% increase in the prioritization results.

---

### A Hybrid Feature Selection Algorithm for Gene Expression Data Classification

Reduce gene expression dimension and redundancy to improve the accuracy of gene expression classification applied for tumor genes prediction. Six gene expression datasets.

MIMAGA-Selection algorithm integrate both mutual information maximization (MIM) used for genetic filtration, and adaptive genetic algorithm (AGA) used for feature selection with extreme leaning machine (ELM) classification accuracy.

With the number of genes up to 20,000, MIMAGA-Selection algorithm is always able to reduce the gene number to below 300 with reasonably high classification accuracies, and compared to three algorithms, ReliefF, sequential forward selection (SFS) and MIM on same datasets with same target genes number, it demonstrates higher accuracy rates.
### Table 1 Continued

<table>
<thead>
<tr>
<th>5</th>
<th>PPI-Keywords (PPIK) network consisting of proteins and their keywords describing their biological properties to enrich PPI network.</th>
<th>Gain better understanding of protein functionality and their association to diseases.</th>
<th>Protein-protein interaction (PPI) network.</th>
<th>Integrating biological feature to PPIN and using metagraph representations of topological similarity to build more accurate classifiers using supervised learning technique.</th>
<th>The classifiers outperform diffusion-based methods and module-based methods by 13.8–32.9% for overall disease protein prediction. They also outperform RWR, PRINCE and the module-based methods by 6.6–14.2% for predicting breast cancer genes.</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>A heterogeneous network and Piecewise linear regression-protein sequence similarity (RWRHN-PLR-SS) to improve previous algorithms.</td>
<td>Reduce the prediction error of candidate disease genes in their previous two classifiers to achieve better results.</td>
<td>Heterogeneous network constructed of protein–protein interaction (PPI), phenotype similarity network, the Gene–phenotype relationships, and the protein sequence similarity.</td>
<td>Random walk-based algorithm on a reliable heterogeneous network of sequence similarity using Piecewise linear regression.</td>
<td>The experimental results show that the performance of RWRHN-PLR-SS is better than their two previous methods and the first new method.</td>
</tr>
<tr>
<td>7</td>
<td>Integrating functional annotations, protein interactions and gene expression information to compute similarity between genes in a disease related metabolic network.</td>
<td>Prioritize disease risk genes in a disease related metabolic network and broadly apply the proposed method to other metabolism-related diseases to help achieve more accurate diagnosis and effective therapies.</td>
<td>The obesity gene expression profile, 195 obesity disease genes, Protein interaction information, and Metabolism-related basic metabolites, proteins and reactions extracted and organized from several large data platforms.</td>
<td>Prioritization algorithm named PDRG (prioritize disease risk genes) to compute the disease risk score of genes in the disease-related metabolic network by utilizing three resources: interaction, expression and function data.</td>
<td>From the top 50 known obese genes, PDRG method identifies 42 disease risk genes, while ToppNet and ToppGene identify only 8 and 13. Additionally, based on the metabolic-PPI network, the results demonstrate that PDRG method is better than ToppNet and ToppGene in terms of literature validation rate and recall rate.</td>
</tr>
</tbody>
</table>
### Table 1 Continued

<table>
<thead>
<tr>
<th></th>
<th>A two-step approach to apply the best-performing methods and reliable multi-genomic data sources by exploring a wide range of datasets and algorithms.</th>
<th>Solving gene prioritization problem by incorporating different dataset resources and methods.</th>
<th>1) Network-based data: PPI databases included in HIPPIE v2.0 and FLN datasets. 2) Functional annotation-based data: The kyoto encyclopedia of genes and genomes (KEGG), Reactome pathways and GO of three domains, CC, MF, and BP.</th>
<th>Seven local rankers are used to prioritize candidate genes on the reliable multi-genomic data sources. Then simple rank average method is applied to build global ranking of all local rankers’ results.</th>
<th>GPS provides a better performance than other tools such as DIR, GeneDistiller, GPEC, and Endeavour in all evaluation metrics like AUC, TOP5 and TOP30.</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>Gene Ontology Hierarchy Preserving Hashing (HPHash) based semantic method for gene function prediction.</td>
<td>Predict correct GO annotations for genes from a massive set of GO terms as defined by GO to correctly predict the biological functions of genes used in candidate disease genes prioritization.</td>
<td>Two GO annotations of three species, one as a training set and the other as the testing set. Also contemporary ontology files from Gene Ontology. 2 GO.</td>
<td>HPHash method measures the semantics similarity using GO terms for genes annotation. It then uses a hierarchy preserving hashing technique to keep the hierarchical order between GO terms, and to optimize a series of hashing functions to compact the gene-term association matrix into a low-dimensional one and perform semantic similarity based gene function prediction in the low-dimensional space.</td>
<td>HPHash outperforms other semantic similarity and function prediction methods such as Baseline (Ham), Baseline (Cos), InterGFP (BMA) and clusDCA (G). Moreover, it surpasses other methods in the ability to predict completely un-annotated genes.</td>
</tr>
</tbody>
</table>
order to improve the results. Moreover, different experimental approaches were used to test the performance of these approaches. One of them is called cross validation in which researchers divide dataset into a number of subset (folds) and repeat the experiment for a number of repetitions. In each iteration they use one fold for testing and the rest for training. After that, they compare the results of all experiments with other existing methods to measure their proposal performance by using some of accuracy measures. Another experimental approach split dataset into a number of splits ranging from 200 split to 1200 split. For every split, the proposed method is applied and compared with other methods. For precision measurement, researchers usually used precision, recall and F-measure to compare their methods with other methods. They also used AUC measure to evaluate the prediction ability of the proposed classifier compared with other existing classifiers. From the experimental results of those approaches, we found that their methods had succeed to outperform other methods by a percentage ranges from 6% to 17%. Some of them also exceeded previous methods by more than 50%.

8. Conclusions

Many techniques of disease-associated genes identification have been proposed. They are geared by the fundamental part of this technology in medical improvement. However, the scarcity of available information of known diseases has considered a big barrier. Hence, different approaches and algorithms have been developed to overcome this challenge.

This survey explains different identification techniques and datasets and it illustrates a comparison between techniques based on their methodology, dataset and their experimental findings compared to other methods. We believe that the major focus of these approaches is based on enriching current genome datasets and all related sources and algorithms. Most of them are reciprocally interactive, and we think that based on them, many effective approaches could be developed.

9. References


