Journal of Neonatal Surgery
ISSN(Online): 2226-0439
Vol. 14, Issue 16s (2025)
https://www.jneonatalsurg.com

OPEN ACCESS

Advances in Computational Approaches for Disease Prediction: A Comprehensive Survey of Candidate Gene Identification Using Protein-Protein Interaction (PPI) Networks and Multi-Omics Integration

V.Shanu[1], Dr.K. Chitra[2]

[1]P. hD Scholar, Department of Computer Science, Sri Krishna Adithya College of Arts &Science, Coimbatore

[2]Associate professor &Head, Department of Computer Science, Sri Krishna Adithya College of Arts & Science,Coimbatore

Cite this paper as: V. Shanu, Dr.K. Chitra, (2025) Advances in Computational Approaches for Disease Prediction: A Comprehensive Survey of Candidate Gene Identification Using Protein-Protein Interaction (PPI) Networks and Multi-Omics Integration. *Journal of Neonatal Surgery*, 14 (16s), 615-620.

ABSTRACT

Disease prediction plays a crucial role in understanding the underlying genetic factors responsible for various conditions and improving early diagnosis and treatment. Computational methods have become essential tools in identifying genes linked to diseases, aiding in the discovery of new therapeutic targets. One of the key areas of research in this field is the use of protein-protein interaction (PPI) networks and gene analysis to predict candidate genes involved in disease processes. This research focuses on the prediction of candidate genes using PPI networks, examining various computational approaches to identify genes potentially involved in specific biological processes, diseases, or traits. By leveraging the structural and functional insights provided by PPI networks, these methods infer gene involvement based on their interactions with known disease-associated proteins. Techniques range from network-based algorithms, which analyze the topological properties of PPI networks, to machine learning models that predict gene-disease associations using complex data patterns. Additionally, integrative approaches combine PPI networks with multi-omics data, such as genomics, transcriptomics, and epigenomics, to enhance the biological relevance of predictions. This survey highlights the strengths and limitations of each method, offering a comprehensive overview of the evolving landscape of candidate gene prediction.

Keywords: *Candidate gene prediction, protein-protein interaction (PPI) networks, disease prediction, network-based algorithms, machine learning, multi-omics integration, gene-disease association, computational biology, gene analysis, bioinformatics.*

1. INTRODUCTION

In the field of bioinformatics, uncovering and understanding genes linked to specific diseases is a key area of research. This work enhances our comprehension of disease processes, aids in diagnosis, and contributes to the discovery of new therapeutic approaches. A **candidate gene** refers to a gene thought to be involved in a particular disease, often selected based on its biological role, location in the genome, or prior associations with related conditions. Over the last ten years, the advancement of high-throughput genomic technologies like **next-generation sequencing (NGS)** and **microarrays** has allowed researchers to produce substantial amounts of genomic data. This extensive data has facilitated more advanced methods for determining gene similarities and improving the identification of genes that may be linked to diseases.

The exploration of candidate genes has been fundamental in genetics, traditionally guided by hypothesis-driven strategies. Researchers initially focused on genes presumed to influence diseases, chosen based on their biological functions, locations on chromosomes, or links to similar conditions. For instance, genes involved in metabolic pathways have been targets in studies of metabolic diseases. Despite yielding crucial insights, this method had notable shortcomings, particularly for complex diseases like cancer, diabetes, or neurodegenerative disorders, which involve interactions among numerous genes and environmental factors.

The introduction of high-throughput technologies such as next-generation sequencing (NGS), whole-exome sequencing (WES), and genome-wide association studies (GWAS) has dramatically expanded researchers' capabilities. Now, rather than limiting themselves to a few potential candidates, scientists are able to sequence whole genomes or exomes, producing extensive datasets that not only include genes directly associated with a disease but also map their interactive networks. This abundance of data has spurred the creation of advanced computational methods aimed at sorting through and prioritizing candidate genes from these extensive pools of genetic information.

2. METHODS FOR PREDICTING CANDIDATE GENES VIA PPI NETWORKS

A. Network-Based Approaches

1. **Random Walk with Restart (RWR):** This algorithm performs a simulated random walk on the PPI network to prioritize genes or proteins that are close in the network topology to known disease-associated genes.
2. **Network Propagation:** Network propagation techniques distribute information, such as disease relevance scores, from established disease genes throughout the PPI network to rank other genes based on their potential role in the disease.
3. **Network Clustering:** Groups of proteins that are densely connected in PPI networks are likely to share similar functions or be part of the same pathways. Identifying these clusters can help discover potential candidate genes.
4. **Shortest Path Algorithms:** Identifying the shortest paths between disease-related proteins and other proteins in the PPI network can highlight genes that may be involved in related biological processes.

B. Machine Learning Approaches

1. **Supervised Learning:** Models that are trained on known gene-disease associations can predict new candidate genes by analyzing their position and relationships within PPI networks.
2. **Graph Neural Networks (GNNs):** GNNs are machine learning models specifically designed to process graph-structured data. They have been applied to PPI networks to predict novel interactions and potential candidate genes.
3. **Feature-Based Approaches:** These approaches extract features such as network centrality, clustering coefficients, and other graph metrics from PPI networks to classify or rank genes as candidates for disease involvement.

C. Integrative Approaches

1. **Gene Expression and PPI Networks:** By combining gene expression data with PPI networks, candidate gene predictions can be refined by focusing on genes and interactions that are active in particular tissues or conditions.
2. **Multi-Omics Data Integration:** Integrating PPI data with other omics datasets—such as genomics, transcriptomics, proteomics, metabolomics, and epigenomics. This holistic view enhances candidate gene prediction by considering various layers of biological information, capturing both genetic and environmental factors contributing to disease.
3. **Disease-Specific Networks:** By constructing PPI networks that are specific to certain diseases or conditions, researchers can more effectively identify candidate genes that are relevant to those specific diseases.

3. TOP OF FORM

Review on Network Based Approaches

Since 2020, advancements in network-based approaches have greatly improved the accuracy and scalability of predicting candidate genes using Protein-Protein Interaction (PPI) networks. These approaches leverage graph-based algorithms and machine learning techniques to analyze PPI data, helping identify genes associated with diseases or biological traits.

(i)Random Walk with Restart (RWR)

Zhao et al. (2020) introduced a context-specific RWR model that adapts the restart probability based on tissue-specific gene expression data. Their study improved the identification of tissue-relevant candidate genes for complex diseases, such as cancer and metabolic disorders.

Morshed et al. (2021) enhanced the traditional RWR by combining it with genomic and transcriptomic data, showing increased accuracy in predicting disease genes in neurodegenerative disorders such as Alzheimer's.

(ii)Network Propagation

Köhler et al. (2021) developed a novel network propagation method that combines PPI data with functional annotations, such as gene ontology terms and pathway involvement, to prioritize candidate genes for rare diseases. Their method significantly improved predictions for understudied diseases.

Zhu et al. (2020) applied a network propagation framework that integrates epigenetic information with PPI data to identify candidate genes for complex traits. This hybrid approach improved predictions for developmental disorders.

(iii). Network Clustering

Dong et al. (2022) used a dynamic network clustering method that considers temporal changes in protein interactions across

different developmental stages. Their work demonstrated improved prediction of stage-specific candidate genes in cancer and neurological disorders.

Jiang et al. (2020) proposed an unsupervised deep learning-based clustering algorithm to detect functional gene modules in PPI networks. By using autoencoders, they achieved better performance in identifying candidate genes for autoimmune diseases.

(iv) Shortest Path Approaches

Liu et al. (2021) introduced a weighted shortest path algorithm that incorporates edge weights based on functional similarity between proteins, improving the prediction of candidate genes for cardiovascular diseases.

Sun et al. (2020) applied a shortest path-based approach to study Parkinson's disease. They introduced a pathway-specific analysis, allowing for the identification of genes involved in specific biological processes related to neurodegeneration.

Liu et al. (2021) introduced a weighted shortest path algorithm that incorporates edge weights based on functional similarity between proteins, improving the prediction of candidate genes for cardiovascular diseases.

Sun et al. (2020) applied a shortest path-based approach to study Parkinson's disease. They introduced a pathway-specific analysis, allowing for the identification of genes involved in specific biological processes related to neurodegeneration.

Network-Based Approaches		
Year and Author	**Methods**	**Contribution**
Zhao et al. (2020)	context-specific Random Walk with Restart (RWR)	Integration of tissue-specific data.
Morshed et al. (2021)	RWR by combining it with genomic and transcriptomic data,	Use of genomic and transcriptomic datasets in conjunction with PPI networks.
Köhler et al. (2021)	novel network propagation Method	Incorporation of epigenetic and functional annotation data.
Zhu et al. (2020)	network propagation with epigenetic information	predictions for developmental disorders.
Dong et al. (2022)	dynamic network clustering	neurological disorders.
Jiang et al. (2020)	unsupervised deep learning-based clustering algorithm	identifying candidate genes for autoimmune diseases.
Liu et al. (2021)	weighted shortest paths	Incorporation of weighted edges for functional similarity.
Sun et al. (2020)	Pathway-centric shortest path approach	specific biological processes related to neurodegeneration.

4. REVIEW ON MACHINE LEARNING BASED APPROACHES

Machine learning (ML) approaches have become increasingly significant in the prediction of candidate genes using protein-protein interaction (PPI) networks, owing to their ability to handle high-dimensional data and uncover complex patterns. Since 2020, there have been substantial advancements in applying supervised learning, deep learning, and other machine learning techniques to PPI networks for predicting disease-related genes.

(i)supervised learning

Chen et al. (2020) developed a supervised learning framework using PPI network features alongside gene expression data. Their method successfully identified novel candidate genes for several complex diseases by applying logistic regression and support vector machines (SVMs).

Li et al. (2021) proposed a Random Forest-based model that integrates protein interaction networks and gene function annotations. The model was applied to cardiovascular disease datasets, improving the prediction of candidate genes related

to disease pathways.

(ii) Graph Neural Networks (GNNs)

Zitnik et al. (2021) introduced a GNN-based method, **GraphSAGE**, which was applied to multi-layer PPI networks to predict novel gene-disease associations. Their method showed superior performance compared to traditional ML techniques, especially in large, complex datasets.

Peng et al. (2022) applied a GNN model, **DeepGraphGO**, which integrates PPI data with gene ontology (GO) annotations to predict candidate genes for neurodegenerative diseases. Their approach significantly improved predictions in Alzheimer's and Parkinson's disease datasets.

(iii) Feature-Based Machine Learning Approaches

Cao et al. (2021) developed a feature-based Random Forest model, leveraging features like node centrality, closeness, and clustering coefficients. The model achieved high precision in predicting candidate genes for cancer and autoimmune diseases.

Sun et al. (2020) created a hybrid machine learning model combining feature extraction with an SVM classifier. They extracted graph metrics, such as node connectivity and degree distribution, from PPI networks and used them to predict genes related to cardiovascular diseases.

Machine Learning Based Approaches		
Year and Author	**Methods**	**Contribution**
Chen et al. (2020)	supervised learning framework	Use of diverse network features like node centrality and clustering.
Li, L., Gao, Z., & Zhang, X. (2021).	Random Forest-based model	applied to cardiovascular disease
Zitnik, M., Wang, C., & Leskovec, J. (2021).	*graph neural networks and multi-layer PPI networks*	Use of deep learning techniques to directly model interactions in PPI networks.
Peng, J., Tang, Y., & Liu, Q. (2022).	*DeepGraphGO*	Integration of gene ontology annotations.
Cao, Y., Zhu, Z., & Li, J. (2021).	*A feature-based Random Forest approach*	predicting candidate genes for cancer and autoimmune diseases.
Sun, Y., Zhang, Y., & Ma, Z. (2020).	*hybrid machine learning model*	*predicting cardiovascular disease genes*

5. REVIEW ON INTEGRATIVE APPROACHES

Integrative approaches combine multiple data types—such as PPI networks, gene expression data, genomics, transcriptomics, and metabolomics—to enhance the accuracy of candidate gene prediction. By leveraging diverse biological information, these methods provide a more holistic view of gene-disease associations, improving the identification of disease-relevant genes.

(i)Gene Expression and PPI Networks

Chen et al. (2020) integrated PPI networks with gene expression data from disease-specific tissues to predict candidate genes for Alzheimer's disease. By focusing on differentially expressed genes in brain tissues, they identified novel genes with potential roles in disease progression.

Wang et al. (2021) used a similar approach to identify genes involved in breast cancer by combining PPI networks with RNA-seq expression profiles. This method improved the prediction of genes driving tumor growth.

(ii)Multi-Omics Data Integration

Wu et al. (2021) combined genomic, transcriptomic, and proteomic data with PPI networks to predict candidate genes for cardiovascular diseases. Their model integrated data from genome-wide association studies (GWAS), gene expression profiles, and PPI networks, leading to the discovery of novel risk genes for heart disease.

Liu et al. (2020) applied a multi-omics integration approach, incorporating epigenomic and metabolomic data with PPI networks, to predict genes involved in metabolic disorders. The study demonstrated that combining these datasets provides

deeper insights into gene-disease associations.

(iii) Epigenomics and PPI Networks

Xu et al. (2021) developed an integrative approach that combines PPI networks with chromatin accessibility data (ATAC-seq) to predict genes involved in cancer. By incorporating both interaction and regulatory data, they identified genes that are likely to be functionally significant in tumor progression.

Zhang et al. (2020) combined DNA methylation data with PPI networks to predict candidate genes in breast cancer. Their method effectively highlighted epigenetically regulated genes that interact with known oncogenes.

Integrative Approaches		
Year and Author	**Methods**	**Contribution**
Chen et al. (2020)	*Integration of PPI networks and gene expression data*	Focus on tissue- and disease-specific gene expression data to increase prediction relevance
Wang et al. (2021)	Integration of PPI networks with RNA-seq expression	Improved ability to prioritize genes that are differentially expressed in specific disease contexts.
Wu et al.(2021)	Multi-omics data integration with PPI networks	Integration of GWAS, epigenomics, and metabolomics with PPI networks.
Liu et al. (2020)	epigenomic and metabolomic data with PPI networks,	Successful application in complex, multifactorial diseases such as cardiovascular and metabolic disorders.
Xu et al. (2021)	*Integrating protein-protein interaction networks and chromatin accessibility*	Use of chromatin accessibility data (e.g., ATAC-seq) to enhance gene prediction.
Zhang et al. (2020)	Integration of DNA methylation and protein interaction networks	*identifies candidate genes in breast cancer*

6. CONCLUSION

In recent years, the prediction of candidate genes via protein-protein interaction (PPI) networks has advanced through network-based, machine learning, and integrative approaches, each offering distinct advantages. **Network-based methods** leverage the topological features of PPI networks, using techniques like Random Walk with Restart, network propagation, clustering, and shortest path algorithms to identify genes closely connected to known disease genes. While effective, these approaches are sometimes limited by incomplete or noisy PPI data. **Machine learning approaches** enhance prediction accuracy by training models, including Graph Neural Networks (GNNs), on known gene-disease associations, allowing them to capture complex patterns within PPI networks. These methods, though powerful, depend on the availability of sufficient labeled data. Finally, **integrative approaches** combine PPI networks with multi-omics data (e.g., genomics, transcriptomics, epigenomics), offering a more comprehensive understanding of gene function in specific disease contexts. These approaches excel in providing biologically relevant insights but can be computationally demanding. The future of candidate gene prediction lies in integrating these methods, combining the predictive power of machine learning with the multi-layered biological context provided by integrative approaches for more accurate and meaningful predictions.

REFERENCES

[1] Zhao, X., Li, Z., Yang, M., & Fang, J. (2020). Predicting candidate disease genes by integrating tissue-specific gene expression data into a random walk with restart model on protein–protein interaction networks. BMC Bioinformatics, 21(1), 283.

[2] Morshed, M. N., Mollah, M. N. H., & Mazumder, H. (2021). An integrative approach to identifying candidate genes for Alzheimer's disease using protein–protein interaction network and machine learning. Scientific Reports, 11(1), 17822.

[3] Köhler, S., Gargano, M., Matentzoglu, N., & Robinson, P. N. (2021). Network propagation approaches for prioritizing candidate genes in rare disease research. American Journal of Human Genetics, 108(3), 439-452.

[4] Zhu, X., Liao, Z., Li, Y., & Liu, S. (2020). Integrating epigenomic data and protein-protein interaction network for candidate gene prioritization of complex diseases. Frontiers in Genetics, 11, 582749.

[5] Dong, Z., Cui, J., Liu, Q., & Shen, Y. (2022). Dynamic clustering of protein interaction networks improves candidate gene prioritization for stage-specific diseases. PLoS Computational Biology, 18(3), e1009803.

[6] Jiang, P., Zhang, Y., & Chen, J. (2020). Unsupervised deep learning for detecting gene modules in protein–protein interaction networks. Bioinformatics, 36(9), 2639-2645.

[7] Liu, Z., Wang, C., & Wang, Y. (2021). Predicting disease genes using weighted shortest paths in protein-protein interaction networks. Bioinformatics, 37(12), 1730-1738.

[8] Sun, Y., Liu, L., & Huang, T. (2020). Pathway-centric shortest path approach for prioritizing candidate genes for Parkinson's disease. BMC Bioinformatics, 21(1), 527.

[9] Chen, Y., Yang, P., & Sun, X. (2020). Prediction of candidate genes for complex diseases based on protein–protein interaction networks and gene expression data. BMC Genomics, 21, 512.

[10] Li, L., Gao, Z., & Zhang, X. (2021). Prediction of cardiovascular disease-related candidate genes based on protein interaction networks and gene function annotations. Scientific Reports, 11(1), 9736.

[11] Zitnik, M., Wang, C., & Leskovec, J. (2021). Predicting gene-disease associations with graph neural networks and multi-layer PPI networks. Nature Communications, 12(1), 1257.

[12] Peng, J., Tang, Y., & Liu, Q. (2022). DeepGraphGO: Predicting candidate genes in neurodegenerative diseases by integrating PPI networks with graph neural networks and gene ontology annotations. Bioinformatics, 38(6), 899-906.

[13] Cao, Y., Zhu, Z., & Li, J. (2021). A feature-based Random Forest approach for predicting cancer candidate genes using protein-protein interaction networks. BMC Medical Genomics, 14(1), 214.

[14] Sun, Y., Zhang, Y., & Ma, Z. (2020). A hybrid machine learning model for predicting cardiovascular disease genes based on PPI network features. Journal of Translational Medicine, 18, 300.

[15] Chen, L., Zhang, X., & Liu, M. (2020). Integration of PPI networks and gene expression data reveals candidate genes for Alzheimer's disease. Journal of Alzheimer's Disease, 75(3), 801-812.

[16] Wang, Y., Zhao, Z., & Zhang, D. (2021). Integrating gene expression data with protein-protein interaction networks for breast cancer candidate gene prediction. BMC Bioinformatics, 22(1), 312.

[17] Wu, H., Zhao, Y., & Xu, J. (2021). Multi-omics integration with protein interaction networks for prioritizing candidate genes in cardiovascular diseases. Scientific Reports, 11(1), 10988.

[18] Liu, B., Luo, X., & Lin, D. (2020). Multi-omics data integration with PPI networks for predicting candidate genes in metabolic disorders. BMC Medical Genomics, 13, 172.

[19] Xu, L., Wang, Y., & Sun, C. (2021). Integrating protein-protein interaction networks and chromatin accessibility data for predicting cancer-related genes. Bioinformatics, 37(5), 678-684.

[20] Zhang, M., Shi, W., & Wang, L. (2020). Integrative analysis of DNA methylation and protein interaction networks identifies candidate genes in breast cancer. BMC Genomics, 21(1), 492.

.