

Applications of Network Analysis and Machine Learning Methods In Bioinformatics

Summary of PhD Theses

Peter Juma Ochieng

Supervisor: Miklós Krész, PhD

Doctoral School of COMPUTER SCIENCE

Department of COMPUTER ALGORITHMS AND ARTIFICIAL
INTELLIGENCE

Faculty of Science and Informatics

University of Szeged



Szeged
2025

Introduction

In the era of big data, computational methods in bioinformatics play a pivotal role in modern biological research, leveraging advanced data-driven methodologies to explore complex biological systems[3]. As biological data grows increasingly vast, innovative computational approaches are essential to uncover critical insights into gene interactions, protein structures, and disease mechanisms[6]. This thesis introduces new computational methods in network analysis and machine learning to improve bioinformatics applications. Key contributions include advanced network analysis techniques for cancer gene ranking, protein complex detection, and drug-target prediction, along with machine learning algorithms, such as Expectation-Maximization, Savitzky-Golay filtering, and Convolutional Neural Networks, designed to analyze high-dimensional biological data like gene expression and whole-exome sequencing (WES). However, there are several computational challenges inherent to the analysis of biological data:

- **Challenge 1:** Ranking cancer genes in large-scale biological networks presents significant computational challenges due to the vast number of genes and complex interactions[5]. Key problems include calculating individual gene mutation burden scores, assessing gene spreading strength and neighbor influence in the gene consensus networks, and ranking genes in the network using existing algorithms, which often exhibit instability.
- **Challenge 2:** Detecting protein complexes in dense protein-protein interaction (PPI) networks is challenging due to structural complexity, noise, and dynamic interactions[1]. The key challenge lies in identifying protein complexes based on module structure, cluster density, and network topology. Most existing computational methods overlook these structural factors, leading to inefficiencies in accurately detecting and distinguishing relevant protein complexes.
- **Challenge 3:** Predicting drug-target interactions (DTIs) in complex biological networks is challenging due to multi-target interactions, vast chemical space, and intricate biological systems[8]. A key challenge is developing an efficient network-weighted centrality measure to identify primary drug targets, as traditional drug discovery methods are resource-intensive, time-consuming, and hindered by data sparsity and high dimensionality.
- **Challenge 4:** Analyzing gene expression patterns and copy number variation (CNV) peaks from whole exome sequencing (WES) data is a highly intricate task [4]. One of the major challenges in this process is developing a clustering model that can effectively identify groups of genes exhibiting similar expression patterns. This is particularly difficult due to the continuous

nature of gene expression, which makes it challenging to define clear boundaries between different groups. Furthermore, creating an accurate filtering algorithm to analyze CNV peaks within coverage profiles is another substantial computational hurdle. Both tasks require advanced methods to ensure that the data is interpreted correctly and reliably.

- **Challenge 5:** Predicting copy number variation (CNV) bait positions in Whole Exome Sequencing (WES) data is a complex task[7]. A major challenge lies in developing an effective normalization process that enhance prediction of bait coordinates this is due to hidden systemic biases, such as GC bias, which can significantly impact CNV detection sensitivity and specificity. Also, designing a 1D convolutional neural network (CNN) model to accurately predict bait coordinates in WES kits is challenging when performing specially when optimizing GC bias normalization across target regions.

Summary of Theses Results

Thesis 1: A graph-based approach for prioritizing cancer genes

In this work, I proposed a graph-based approach for prioritizing cancer genes, starting with an overview of existing graph-based ranking algorithms to establish their mechanics and applications. Building on this foundation, First, I introduced a novel forward-looking stability measure matrix to evaluate the ranking stability of these algorithms. This matrix provides a robust framework for assessing ranking consistency, which I applied to the proposed graph-based method and related approaches. To evaluate the stability of various ranking methods, I computed the Euclidean distance between consecutive rating vectors using two approaches: rolling window and expanding window(see Figure 1).

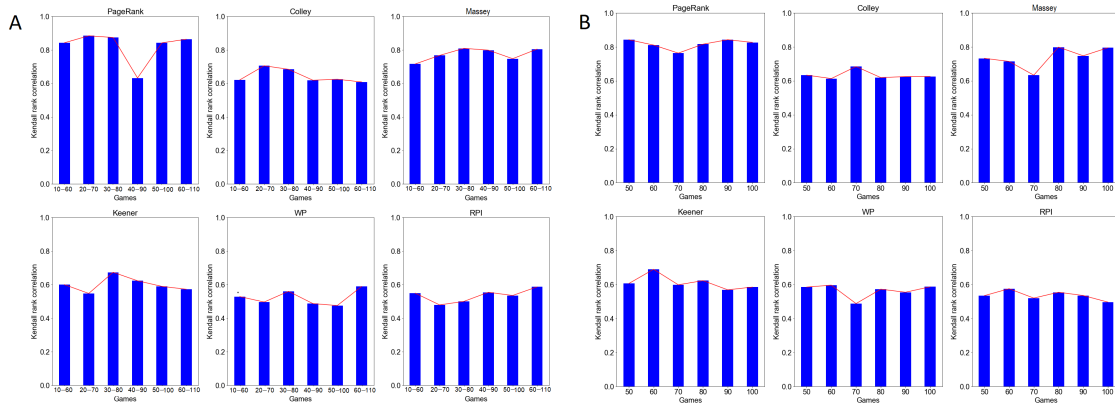


Figure 1: Rank stability. **A:** Rolling window (FLA-RW); **B:** Expanding window (FLA-EW) for different ranking methods.

For the rolling window approach, the distance values $d_{RW}^2(t)$ varied over time, with PageRank and Massey demonstrating higher stability, reflected in low average distances ($d_{RW}^2 = 0.025$ and $d_{RW}^2 = 0.029$, respectively). In contrast, methods like Colley, Keener, WP, and RPI exhibited less stability, with average distances exceeding $d_{RW}^2 \geq 0.035$. Similarly, for the expanding window approach, distance values $d_{EW}^2(t)$ increased with window size, indicating evolving rating stability over time. Again, PageRank and Massey showed greater stability, with average distances between $d_{EW}^2 = 0.025$ and $d_{EW}^2 = 0.030$, while Colley, Keener, WP, and RPI had slightly higher distances ($d_{EW}^2 = 0.035$ to $d_{EW}^2 = 0.040$), reflecting greater deviations and lower stability. Secondly, I developed a graph-based gene prioritization approach, Graph-Based Prioritization with PageRank (GBPPR), to prioritize cancer-related genes. This method integrates an adjusted normalized weighted mutation scoring technique to calculate gene-level mutation burden. It employs a new graph-based prioritization framework that leverages consensus gene interaction networks, gene spreading strength, and the influence of neighboring mutations. A modified PageRank algorithm with dynamic, gene-specific damping factors—calculated based on gene mutation enrichment scores—refines the ranking process. To evaluate GBPPR, I compared it against graph-based methods such as Colley (GBPC), Massey (GBPM), and Keener (GBPK) [5] as well as statistical methods like MutSigCV [9], and MUFFINN [2] on six cancer datasets. I evaluated metrics including rank stability, Discounted Cumulative Gain (DCG), precision, and sensitivity. Results showed that GBPPR achieved superior rank stability, particularly in identifying top-ranked genes, except for the PRAD dataset. It consistently outperformed other methods in terms of DCG indicating its ability to accurately prioritize biologically significant genes (see Figure 2).

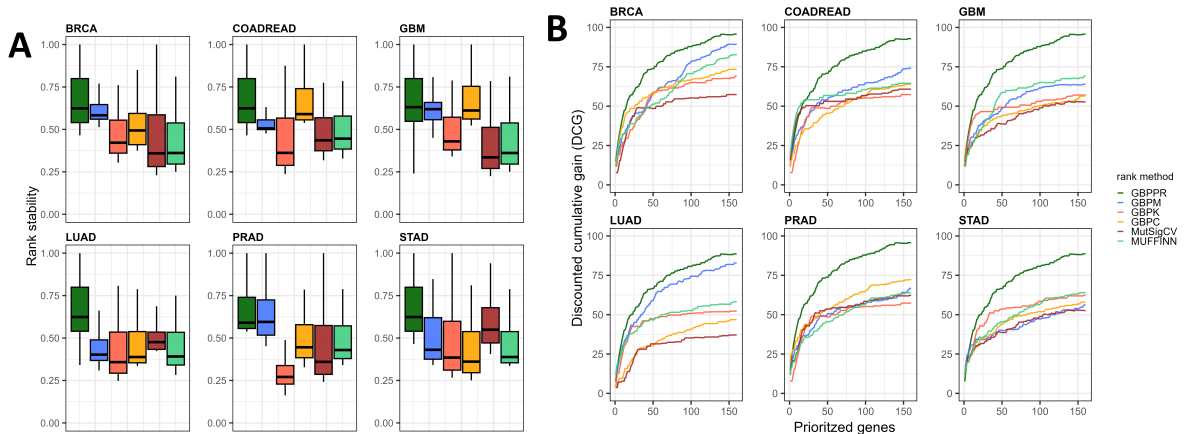


Figure 2: The performance of GBPPR and related methods. **A:** Rank stability and **B:** Discounted cumulative gain (DCG).

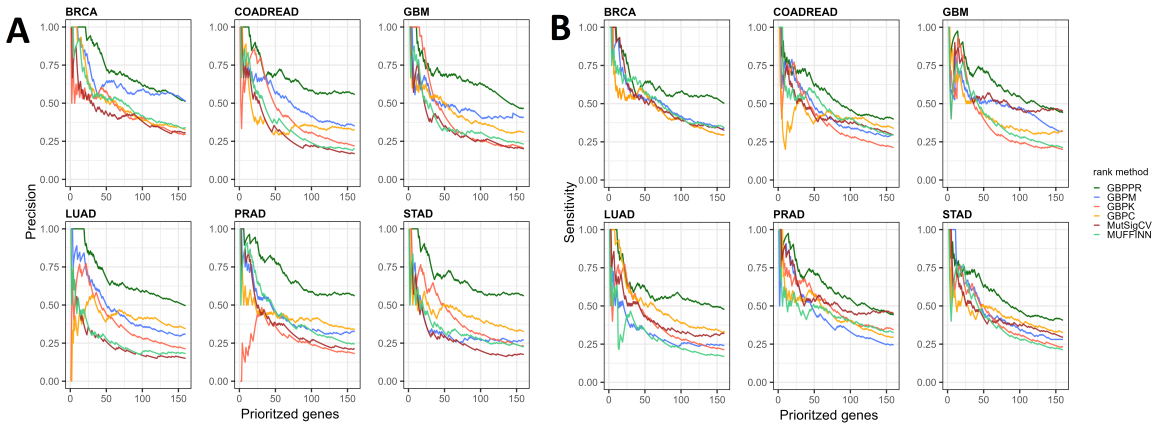


Figure 3: The performance of GBPPR and related methods. **(A):** Precision i.e. is the fraction of prioritized genes (y-axis) contained in the known driver benchmarks for a specific number of genes (x-axis) and **(B):** Sensitivity i.e. is a specific number of genes (y-axis) taking into account the fraction of prioritized genes contained in the known driver benchmarks for a particular number of genes (x-axis).

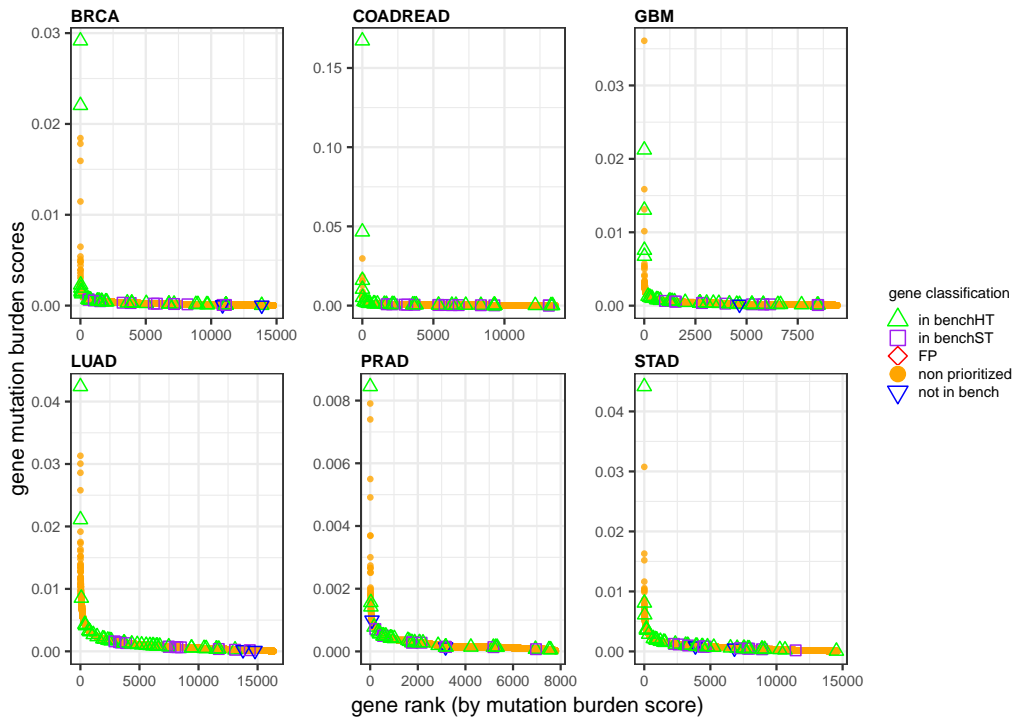


Figure 4: Long tail of top 50 prioritized gene by GBPPR with normalised mean method. The green triangle represents in benchHT (predicted hard truth benchmark genes), the purple square represents in benchST (predicted soft truth benchmark genes), the yellow node is the nonprioritized genes, the blue triangle represents prioritized genes not in both benchmarks and the red diamond is FP (predicted false positive genes).

In ranking precision, GBPPR excelled across most datasets, except for LUAD, where GBPM performed slightly better. Importantly, GBPPR exhibited higher sensitivity, effectively identifying a wide range of relevant genes with significant mutations (see Figure 3). Further, I validated the prioritized genes against both based on hard truth (e.g., NCG, CGC, IntOGen, Bailey, CGI) and soft truth (e.g., CancerMine) benchmark datasets using long tail analysis. In all cancer types, the majority of the top 50 prioritized genes overlapped with benchmark datasets, with no false positives detected. Unlike frequency-based methods, GBPPR effectively prioritized both high- and low-frequency mutated genes present in the benchmarks, highlighting its capability to identify potential driver genes beyond mutation frequency alone (see Figure 4).

Thesis 2: Network-based methods for detecting protein complexes in Protein-Protein Interaction (PPI) networks

In this work, I introduce two methods for detecting protein complexes in protein-protein interaction (PPI) networks: Markov Clustering and a novel approach called Weighted Edge Core-Attachment and Local Modularity structures (WECALM). First, I proposed the Markov Clustering technique, which partitions PPI networks into cohesive clusters by simulating flow within the network. This method identifies densely connected regions corresponding to potential protein complexes, effectively capturing intricate patterns in the PPI network. Using flow simulations, I determined the optimal inflation parameter to maximize cluster granularity. Inflation values between 1.4 and 2.5 were tested, as this parameter influences the strengthening and weakening of currents, impacting cluster formation. Results showed that setting the inflation value to $R = 1.4$ produced 13 protein complexes, while increasing it to $R = 2.5$ generated 49 clusters (see Figure 5). The excessive number of clusters at higher inflation values likely reflects over-representation, indicating that overly strengthened clusters do not accurately represent true protein complexes. In addition, results showed that 78% of 2054 interactions (involving 482 proteins) occurred between protein pairs with similar functions, supporting the hypothesis that high-quality protein complexes primarily consist of proteins with shared functions. To assess the quality of the detected complexes, I calculated the relative number of interactions between functionally similar proteins (RA) and observed a significant relationship between RA and density (d_{in}). Complexes with $d_{in} = 0.101$ or less demonstrated high statistical significance, while increasing d_{in} to 0.268 included larger, more functionally diverse complexes (see Figure 6).

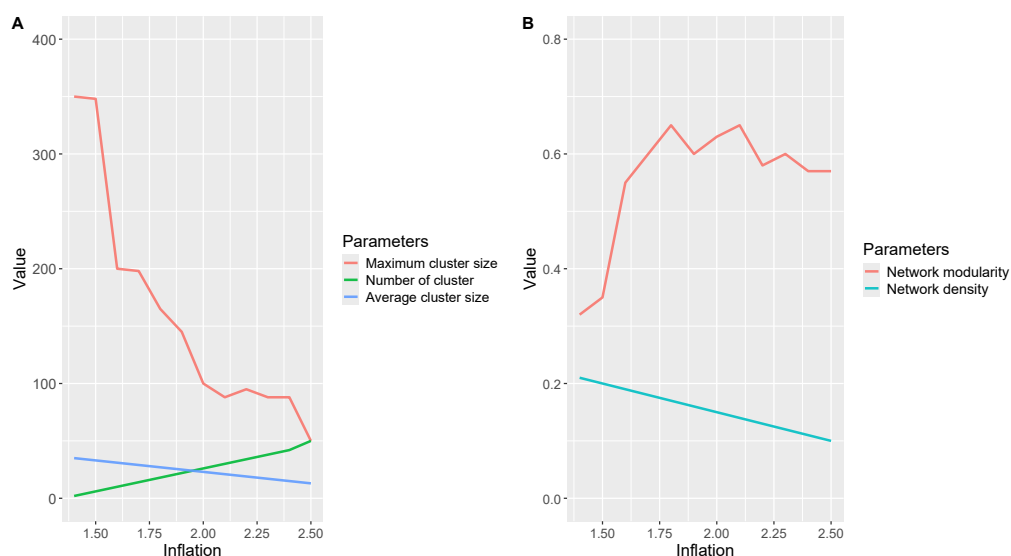


Figure 5: Effect of inflation adjustment on network parameters. **A:** Impact of inflation changes on maximum cluster size, number of cluster, and average cluster size. **B:** Impact of inflation changes on network modularity and density.

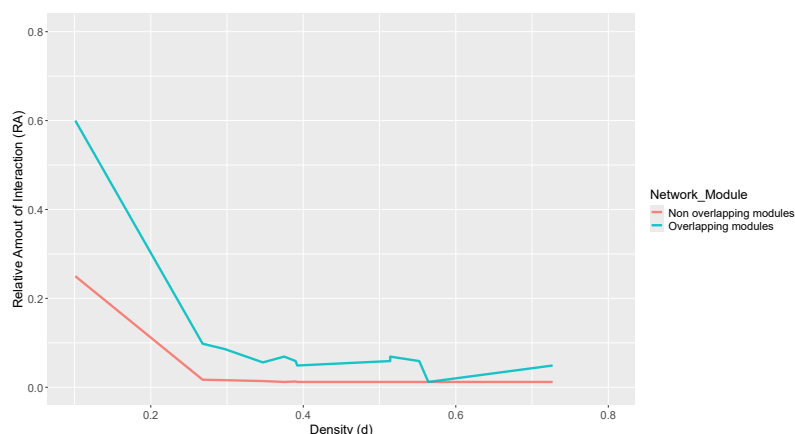


Figure 6: Relative amount of interactions involving non-overlapping and overlapping modules in MCL 34 clusters (protein complexes) against corresponding density values.

Secondly, I developed a new network-based technique for detecting protein complexes called the Weighted Edge Core-Attachment and Local Modularity (WECALM) approach. This method uses structural, weighted network analysis to identify core proteins and their attachments, enabling the detection of complex protein formations within protein-protein interaction (PPI) networks. By leveraging network topology and structural properties, WECALM efficiently captures core structures and local modularity for precise protein complex detection. To evaluate

WECALM's performance, I analyzed the effects of adjusting key parameters π , η , and ω on overlapping score, local modularity score, and core structural similarity score. Adjusting π between 0.1 and 1.0 revealed that recall, MMR, and CR scores decreased as π increased, while precision and F-measure scores were optimal at $\pi = 0.8$ and $\pi = 0.4$, respectively. Similar trends were observed when analyzing BioGRID, DIP, and CYC2008 datasets (see Figure 7).

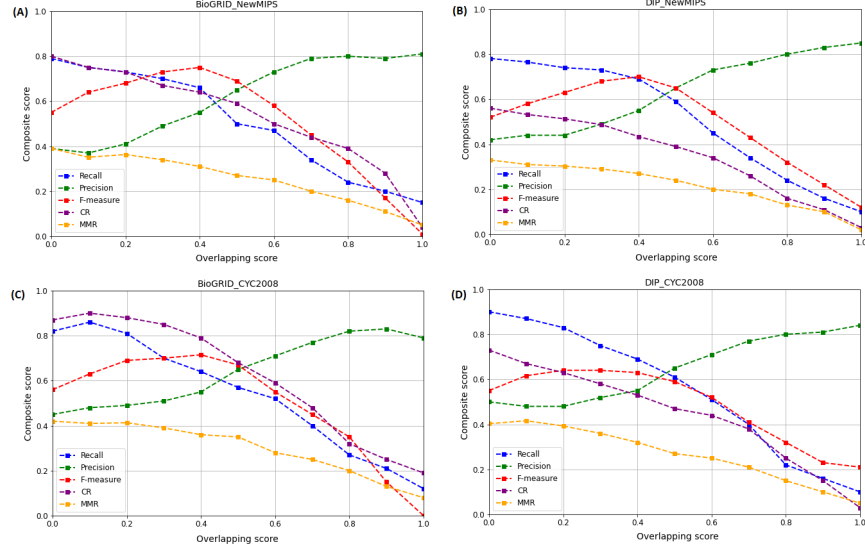


Figure 7: The effect of π on the performance of WECALM based on Recall, Precision, F-measure, CR, and MMR matrices. π is the predefined overlapping threshold.

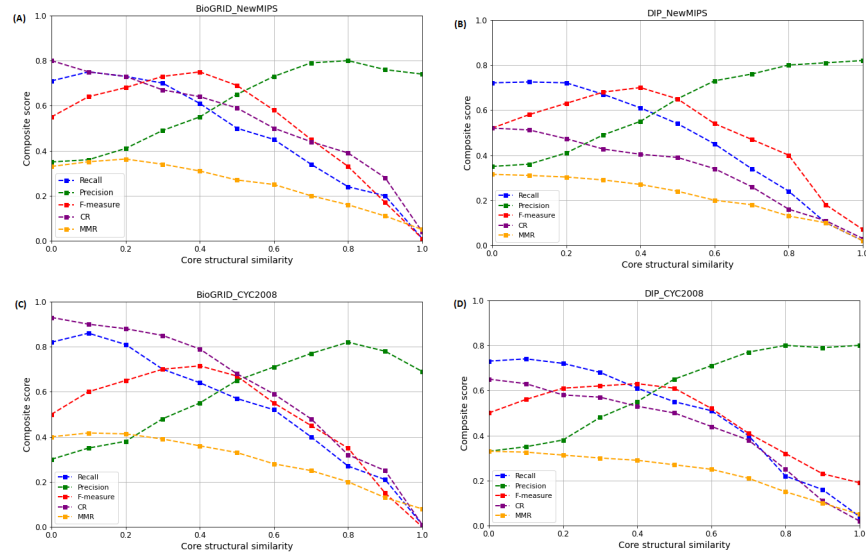


Figure 8: The effect of ω on the performance of WECALM based on Recall, Precision, F-measure, CR, and MMR matrices. ω is a predefined core structural similarity threshold.

The effect of ω on WECALM's performance was also tested, with thresholds ranging from $[0.1, 1.0)$. As ω increased, recall, MMR, and CR scores decreased, while precision and F-measure scores were highest at $\omega = 0.8$ and $\omega = 0.4$. These findings consistently held across BioGRID, DIP, and CYC2008 datasets (see Figure 8). Finally, I compared WECALM's computational complexity and performance against eight other algorithms (see Table 1). Using default settings and benchmark parameters, WECALM demonstrated low computational complexity, comparable to EWCA, while achieving the highest MMR, Sep, and ACC scores. This indicates the accuracy and efficiency WECALM approach in detecting protein complexes in PPI networks.

Table 1: Computational complexity and accuracy of WECALM and other algorithms

Dataset	Algorithm	$C_P(v)$	F-Measure	CR	MMR	Sep	ACC	CPU Run Time (s)
Human	MCL	315	0.1001	0.1759	0.0105	0.1753	0.2167	5906.34
	COACH	4484	0.2455	0.5408	0.0677	0.5216	0.2777	2851.05
	EWCA	1979	0.4048	0.5221	0.0964	0.6081	0.5221	29.37
	CFinder	449	0.1256	0.2834	0.0116	0.3912	0.2511	3896.35
	GMFTP	773	0.2651	0.4193	0.0419	0.4917	0.3852	254.67
	Core	576	0.1621	0.3267	0.1267	0.3573	0.2778	2853.14
	CALM	1108	0.5127	0.5182	0.1394	0.6894	0.5289	198.39
	ClusterONE	375	0.1026	0.3071	0.0207	0.3773	0.2975	4895.78
	CMC	672	0.1251	0.2503	0.0183	0.2975	0.3313	3904.83
	ProRank+	838	0.3651	0.2856	0.0687	0.5526	0.5613	282.66
	WECALM	2367	0.4255	0.5155	0.0981	0.6155	0.6219	28.45
Yeast	MCL	298	0.1104	0.2761	0.0117	0.1625	0.1395	4967.47
	COACH	1551	0.2083	0.5521	0.0466	0.3583	0.3117	3603.31
	EWCA	936	0.4199	0.6182	0.0982	0.5904	0.5879	18.54
	CFinder	351	0.1429	0.2749	0.0281	0.3453	0.4163	3432.07
	GMFTP	675	0.2763	0.3129	0.0309	0.5145	0.4092	229.89
	Core	402	0.2124	0.2968	0.3285	0.1517	0.3218	2543.34
	CALM	732	0.4015	0.6787	0.1433	0.6261	0.6532	154.89
	ClusterONE	317	0.2012	0.2767	0.0285	0.3371	0.3255	3989.92
	CMC	589	0.2115	0.1975	0.0198	0.2934	0.3553	2987.63
	ProRank+	516	0.2712	0.2816	0.0487	0.5471	0.5602	251.54
	WECALM	1891	0.4216	0.6394	0.0487	0.64131	0.6534	17.65

$C_P(v)$: Detected Protein Complex; **CR**: Coverage Rate ; **MMR**: Maximum Match Ratio ; **Sep**: Separation ; **ACC**: Geometrical Mean Accuracy.

Thesis 3: A weighted centrality approach for drug target prediction in complex biological networks

In this work, I developed a computational framework for predicting drug targets in complex biological networks. The framework employs a novel approach based on weighted centrality measures and advanced network construction techniques, allowing it to integrate diverse omics data and provide a more comprehensive analysis of potential drug targets. The goal is to enhance the accuracy and reliability of drug target predictions in intricate biological systems.

I applied the proposed framework in the field of Network Pharmacology to explore the mechanisms of action of *Reuelia* herbal medicine in treating rheumatoid arthritis. To evaluate its performance, I tested the framework using four real network datasets: Freeman’s Electronic Information Exchange System (EIES) network, USAir97 network, Groad network, and Newman’s scientific collaboration network. The spreading influence of top-ranked nodes was assessed using the Susceptible-Infected (SI) model, comparing our method, EWNC, to other centrality measures like degree centrality (DC), betweenness centrality (BC), and closeness centrality (CC).

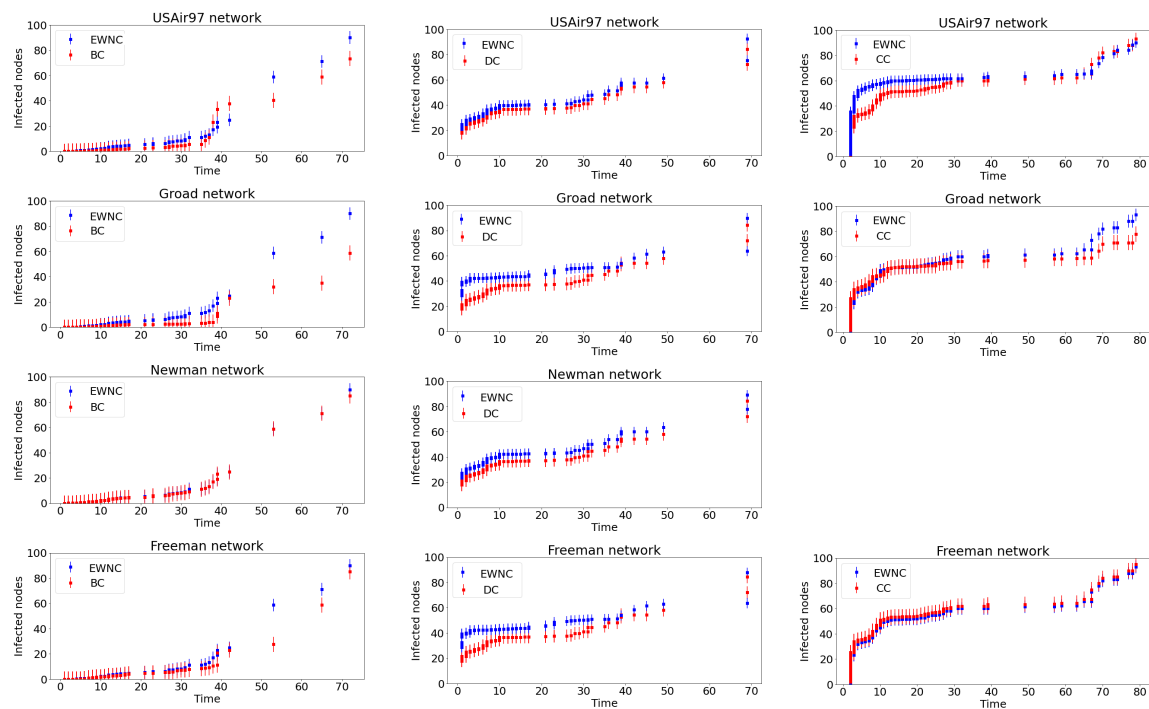


Figure 9: The comparison of EWNC with other classical centrality measures in real networks. The figures show a comparison of the cumulative number of infected nodes as a function of time-step.

In the USAir97 network, EWNC consistently outperformed the other methods, showing a higher number of infected nodes. It also demonstrated faster propagation in the early stages. In the Groad network, EWNC showed superior performance in DC and BC, with fewer errors compared to other measures. For Newman’s network, EWNC performed similarly to DC but outperformed BC in terms of spreading ability and accuracy. In Freeman’s network, EWNC surpassed DC and BC and showed a minimal error compared to CC, suggesting that it is more accurate at identifying influential nodes(see Figure 9). Next, I conducted a Network Pharmacology analysis for *Reuelia* herbal medicine, identifying 1,519 nodes based on weighted centrality scores. These included herbs, composite compounds, putative targets,

known therapeutic targets for rheumatoid arthritis (RA), and other human proteins interacting with these targets. The network revealed 57 important nodes and 417 edges with high centrality scores, indicating the interactions among *Reuelia* herbs, compositive compounds, known RA therapeutic targets, and putative targets involved in RA treatment (see 2).

Table 2: Major identified targets acting on rheumatoid arthritis

Compositive Compounds	ϕ^{DC}	ϕ^{BC}	ϕ^{CC}
Glycidyl oleate	0.5342	0.0485	0.5581
Glyceryl-2 Palmitate	0.5109	0.0517	0.5892
Oleoyl chloride	0.4589	0.0528	0.4706
9-Octadecenoic acid (Z)-, methyl ester	0.4316	0.0525	0.4752
Methyl stearate	0.4446	0.0489	0.4468
Hexadecanoic acid, methyl ester	0.4472	0.0532	0.4656
Hexadecanoic acid, 2,3-bis(acetyloxy)propyl ester	0.4912	0.1413	0.4842
Glyceryl diacetate-2-Oleate	0.4509	0.0517	0.4892
Known RA Targets			
Nitric oxide synthase	0.4982	0.0565	0.5681
Interleukin-1 β	0.4106	0.0493	0.4855
Delta-type opioid receptor	0.5417	0.1199	0.4944
Tumor necrosis factor	0.4626	0.1696	0.4896
Prostaglandin G/H synthase 2	0.4933	0.3967	0.5411
Prostaglandin G/H synthase 1	0.4551	0.5040	0.4794
Growth-regulated alpha protein	0.4126	0.0888	0.5249
C-C chemokine receptor type 5	0.4026	0.0589	0.4609
Interleukin-6	0.4244	0.0671	0.5158
Cannabinoid receptor 2	0.4700	0.0401	0.4352
Putative Targets			
TNF	0.2705	0.0921	0.0588
CXCL10	0.1728	0.0116	0.5640
IL10	0.1728	0.0125	0.5433
IL6	0.3004	0.0930	0.5402
CXCL1	0.1044	0.0151	0.5432
CXCL8	0.1273	0.0160	0.5439
FOXP3	0.4106	0.0149	0.3748
CCR2	0.3004	0.0161	0.4762
CCR5	0.3004	0.0279	0.4922
IL10	0.1273	0.0145	0.5302
CCL2	0.2495	0.0190	0.4548
IL1B	0.1405	0.0253	0.5838
IL17	0.2495	0.0195	0.5374
IL13	0.2212	0.0252	0.5347
CSF2	0.2905	0.0996	0.5101

Finally, I validated 39 major RA-related targets from the imbalanced multi-level network. Gene Ontology and pathway enrichment analyses revealed significant biological processes and molecular functions, including cytokine-mediated signaling, inflammatory responses, and receptor bindings. Additionally, the top ten targeted genes by drugs validated by KEGG pathways analysis show that the majority of targeted genes were involved in IL-17 signaling, TNF signaling, and NF-kappa B signaling, all of which are critical in rheumatoid arthritis.

Thesis 4: Computational frameworks for analyzing time-course gene expression patterns and Copy Number Variations (CNV) coverage profiles

In this work, I developed a new computational framework to analyze time-course gene expression patterns and copy number variations (CNV) coverage profiles. First, I propose a new Rejection Control Expectation Maximization (RCEM) model designed for clustering time-course gene expression data. The RCEM algorithm improves the traditional Expectation-Maximization (EM) approach by integrating a rejection-control mechanism that enhances robustness and convergence, particularly in the presence of noise or outliers. By filtering outliers, this method ensures that rare or noisy data points do not skew parameter estimates, thereby ensuring more stable and accurate clustering. When applied to mRNA levels of 808 *Drosophila melanogaster* genes simulated over 100 time points, the algorithm identified three distinct clusters. Cluster 1, comprising 173 genes, peaked during embryogenesis and adulthood; Cluster 2, with 130 genes, peaked during the larval to pupal transition; and Cluster 3, also with 130 genes, peaked during the egg to larva transition (see Figure 10).

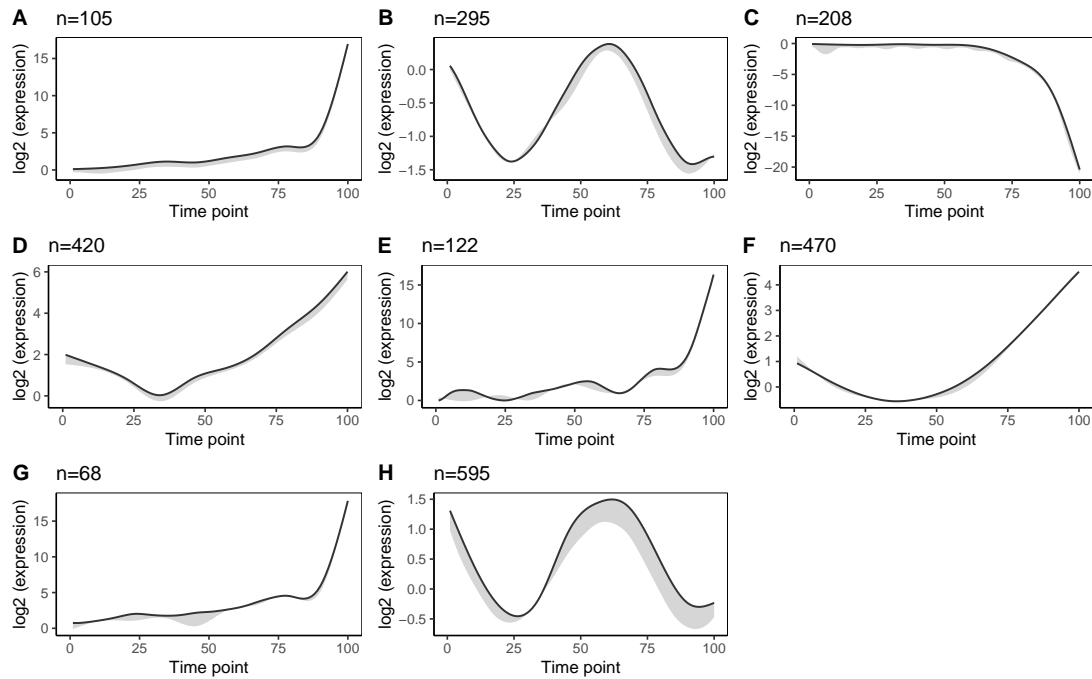


Figure 10: Estimated mean expression curves (solid lines) and 95% confidence bands (grey bands) for eight of 17 clusters discovered by RCEM in the *Drosophila melanogaster* time course microarray data.

I also compared the clustering performance of the proposed model with related methods such as Fuzzy c-Means and K-Means methods (see Figure 11). Using

metrics like the Rand Index, Adjusted Rand Index (ARI), misclassification error rates, and overall success rates, the proposed model demonstrated superior accuracy. It achieved an ARI of 0.9667, significantly outperforming Fuzzy c-Means (ARI of 0.8897) and K-Means. Additionally, the model exhibited a lower interquartile range (IQR) of 0.005649, compared to 0.22629 for Fuzzy c-Means. In terms of misclassification rates, the RCEM model outperformed both Fuzzy c-Means and K-Means with a low misclassification rate of 0.1289, while Fuzzy c-Means and K-Means had higher rates. Furthermore, the proposed method achieved a high success rate of 98.71%, accurately recovering the correct number of clusters and estimating the mean expression curves closely aligned with their true functional forms. The computational complexity of the model is approximately $O(n^2 + t^2)$, indicating both high clustering accuracy and efficiency. The results also highlight the error rates for different clustering methods across various centroid distance calculations, demonstrating the effectiveness of the proposed model.

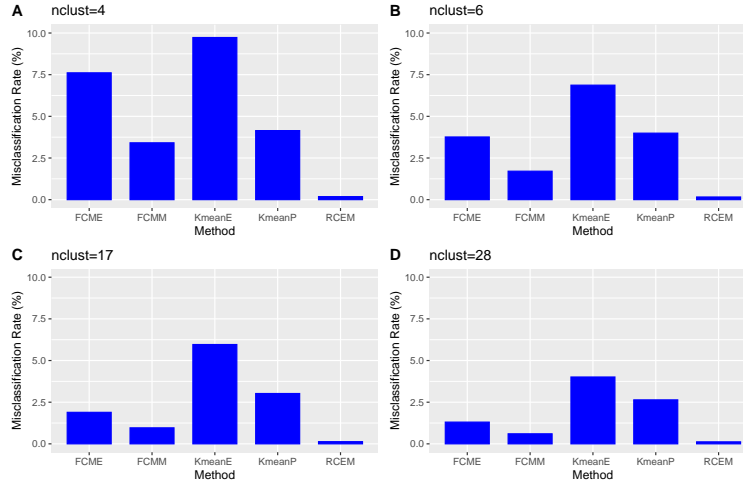


Figure 11: Error rate for 808 simulated *Drosophila melanogaster* mRNA gene datasets at different number of clusters.

Secondly, I proposed an adaptive Savitzky-Golay filter designed to predict and analyze copy number variations (CNVs) in whole-exome sequencing (WES) data. The filter is specifically tailored to address the dynamic complexity of CNVs, enhancing the accuracy of their detection by smoothing the data and reducing noise. To evaluate the effect of filter order on smoothing performance, I calculated the Minimum Mean Squared Error (MMSE) defined by equation: $MMSE = r_k \sigma^{2k} f(S_i)^{\frac{1}{2k}}$, where r_i is noise coefficient, σ is the noise power, $f(S_i)$ is the peak distribution function and k is the filter order. In simulations, the original peaks were corrupted by Gaussian noise at different noise power levels. The results showed that a higher filter order led to a lower MMSE, but also increased computational burden due to the least square (LS) fitting (see Table 3). The adaptive Savitzky-Golay filter, by selecting the polynomial

order automatically, effectively reduced computational complexity while smoothing peaks with high variation (see Table 3).

Table 3: Effect of filter order on the estimation error

Peak	σ	$k_1 = 2$		$k_2 = 3$		$k_3 = 4$		$k_4 = 5$	
		m_1	MMSE	m_2	MMSE	m_3	MMSE	m_4	MMSE
$f(S_1)$	0.050	11	0.00010	21	0.00006	31	0.00005	41	0.00003
	1.000	33	0.00551	63	0.00326	93	0.00216	123	0.00116
$f(S_2)$	0.050	11	0.00008	21	0.00007	31	0.00006	41	0.00004
	1.000	33	0.00672	63	0.00421	93	0.00321	123	0.00213
$f(S_3)$	0.050	11	0.00009	21	0.00008	31	0.00007	41	0.00001
	1.000	33	0.00841	63	0.00554	93	0.00414	123	0.00394

σ , noise power, k filter order, m window length, MMSE Minimum Mean Square Error

In addition, I conducted a comprehensive comparison of the adaptive Savitzky-Golay filter with several other filtering methods, such as Fourier transform, Gaussian Kernel, Epanechnikov Kernel, and Lowess smoothing. The analysis focused on noise suppression and peak height fidelity. The adaptive SG filter outperformed the others in noise suppression, demonstrating superior performance in estimating the optimal window length (see Figure 12). In terms of peak height fidelity, the adaptive filter, along with Lowess and Fourier filters, maintained around 90% of the original peak height, while the Gaussian kernel filter showed poor performance, especially with high-frequency noise. Additionally, the adaptive Savitzky-Golay filter exhibited the lowest root mean squared error (RMSE), indicating minimal estimation bias.

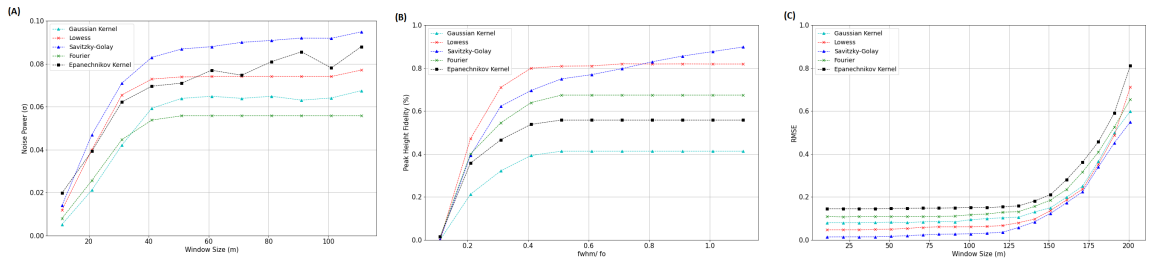


Figure 12: Comparison of performance of Adaptive Savitzky-Golay filtering with peer filtering methods. (A): Comparison based on noise power; (B): Comparison based on peak height fidelity; (C): Comparison based on smoothing bias.

For CNV peak analysis, I applied the adaptive Savitzky-Golay filter to WES data, aiming to detect CNV peaks in coverage profile data and distinguish between mutated and normal genomic regions. I used two experimental datasets: one based on the coverage depth profile (LONG dataset) and another with short reads (SHORT dataset). I observed that the adaptive filter was highly effective in extracting

significant CNV features from both long and short read datasets, even when peak distributions were asymmetrical due to GC bias which indicates substantial feature extraction across all genomic segments indicating the potential of Adaptive SG for detection and analysis of CNV peak coverage profiles (see Figure 13).

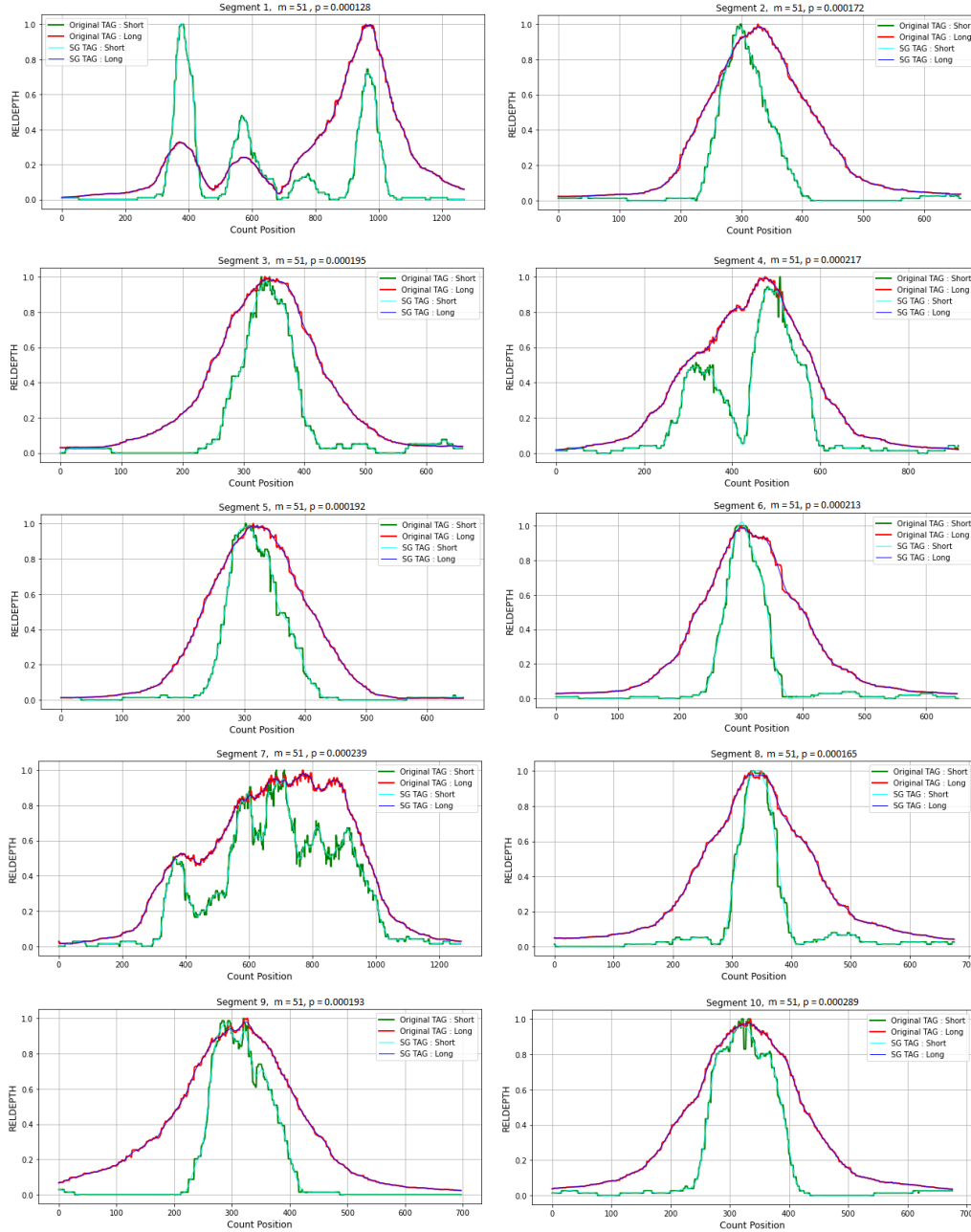


Figure 13: Show CNV peaks from different genomic segments filtered using Adaptive Savitzky–Golay filters. Green and red solid lines are the original short and long tag CNV peaks respectively; cyan and blue solid lines are the smoothed the short and long tags CNV peaks respectively.

Thesis 5: Deep learning frameworks for predicting Copy Number Variations (CNV) bait positions in Whole Exome Sequencing (WES) kits

In this work I introduce a deep learning framework for predicting CNV bait positions in whole exome sequencing (WES) kits by analyzing sequence data, coverage profiles, and on target region information. In many cases, the genomic positions of bait (capture sequences) used in WES kits are not disclosed to users as they are considered trade secrets. However, knowing these bait positions is crucial for the accurate prediction of copy number variants, as it enables better quantitative normalization based on the capture efficiency of baits and target sequences with different GC content. Additionally, it allows for the more precise correction of systematic biases present in coverage profile data. To predict bait coordinates, I tested 10 different NN models. In this work, I trained eight 1D CNN models and two NN models based on dense layers, using various hyperparameters. First, I tested the performance of the proposed CNN model and other different dense layer models by evaluating training and validation accuracy, along with the effects of hyperparameter optimization. Here, I considered testing the models with different windows (data segments) of 500 and 1000 base pairs and trained model with 100 epochs with 1 million training and validation examples. I noted that CNN models with batch normalization significantly outperformed dense models, both in terms of accuracy and training time (see Table 4).

Table 4: Summary of the tested 10 models, including the neural network architecture, input data, and the applied optimizations

Model	Model Architecture	Hyper-parameter optimizations	Input data
1	Conv1D	batch normalization, causal padding, reverse weight	COV, SEQ, ONTARGET
2	Dense	batch normalization, reverse weight	COV, SEQ, ONTARGET
3	Conv1D	causal padding, reverse weight	COV, SEQ, ONTARGET
4	Dense	reverse weight	COV, SEQ, ONTARGET
5	Conv1D	batch normalization, causal padding	COV, SEQ, ONTARGET
6	Conv1D	batch normalization, reverse weight, valid padding	COV, SEQ, ONTARGET
7	Conv1D	batch normalization, reverse weight, same padding	COV, SEQ, ONTARGET
8	Conv1D	batch normalization, causal padding, reverse weight	COV, SEQ
9	Conv1D	batch normalization, causal padding, reverse weight	COV, ONTARGET
10	Conv1D	batch normalization, causal padding, reverse weight	SEQ, ONTARGET

Also, during neural network training, I computed the mean loss for both the training and validation datasets at the end of each epoch, saving models with the lowest loss. Using these models, I predicted 1 million random examples from the evaluation subset and compared them to the ground truth (bait positions provided by the kit manufacturer) to derive key performance metrics. As outlined in the thesis, I trained the network on a randomly selected, uneven number of examples

with varying complexity, based on the number of baits in each data segment (see Figure 14: **A**). In addition, I evaluated loss, accuracy, precision, sensitivity, specificity, F1 score, and MCC to assess models performance. I found that all CNN models outperformed the baseline (dense layer NN models), especially in cases with sparse baits. For models without batch normalization, performance was unstable, further the influence of batch normalization during the model performance (see Figure 14: **B**). Furthermore, I evaluated the effect of different padding options, input data sources, and spatial context on model performance. I found that combining experimental coverage, on-target information, and sequence data produced the best results. Interestingly, excluding sequence data slightly reduced performance, particularly in situations with multiple near-equal solutions. Spatial context played a key role in improving prediction accuracy.

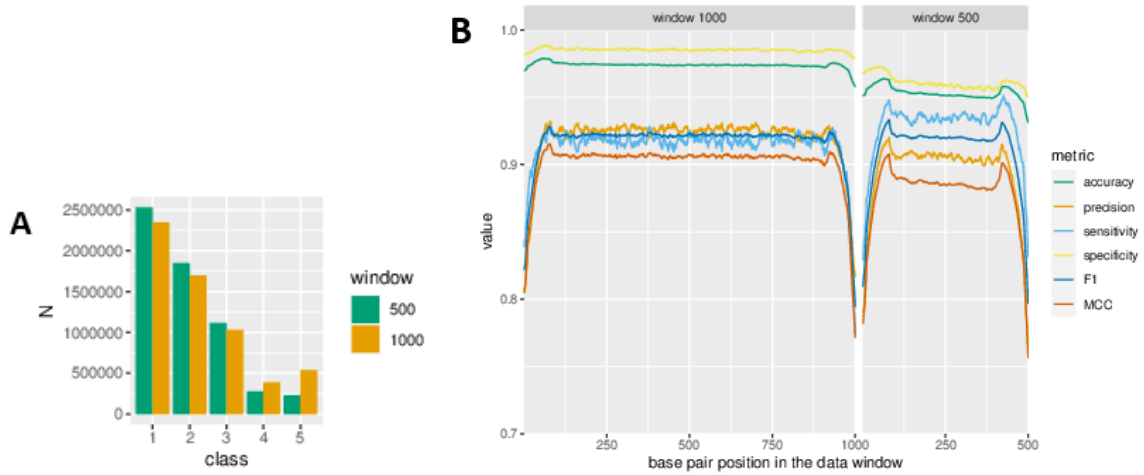


Figure 14: Shows **A:** The distribution of different classes of 1 M train examples of the 500 and 1000 base pairs data sets. The classes (1-5) represent the bait count in the data segment (exactly 1, 1-2, 2-3, 3-4, 4 or more. **B:** The mean of the evaluation metrics based on the position in the spatial data segment.

Finally, I analyzed 1 million predictions and observed that 1D CNN model with batch normalization, causal padding and reverse weight achieved the best predictions ($F1 = 1.0$) for a significant portion of the data (approximately 240,000 for 500-base pair windows and 225,000 for 1000-base pair windows). Even in worst-case scenarios, predictions were mostly accurate, with only slight overlaps or flanking of true bait positions. In contrast, the dense models performed much worse, with far fewer exact predictions. However still in the worst-case scenarios, the predicted bait positions are mainly overlapping and/or flanking the true bait positions. In the case of the Dense models only 421 (window 500) and 592 (window 1000) were predicted exactly ($F1 = 1.0$) score and overall they also show much worse predictions (see Figure 15:A and A, bottom panes).

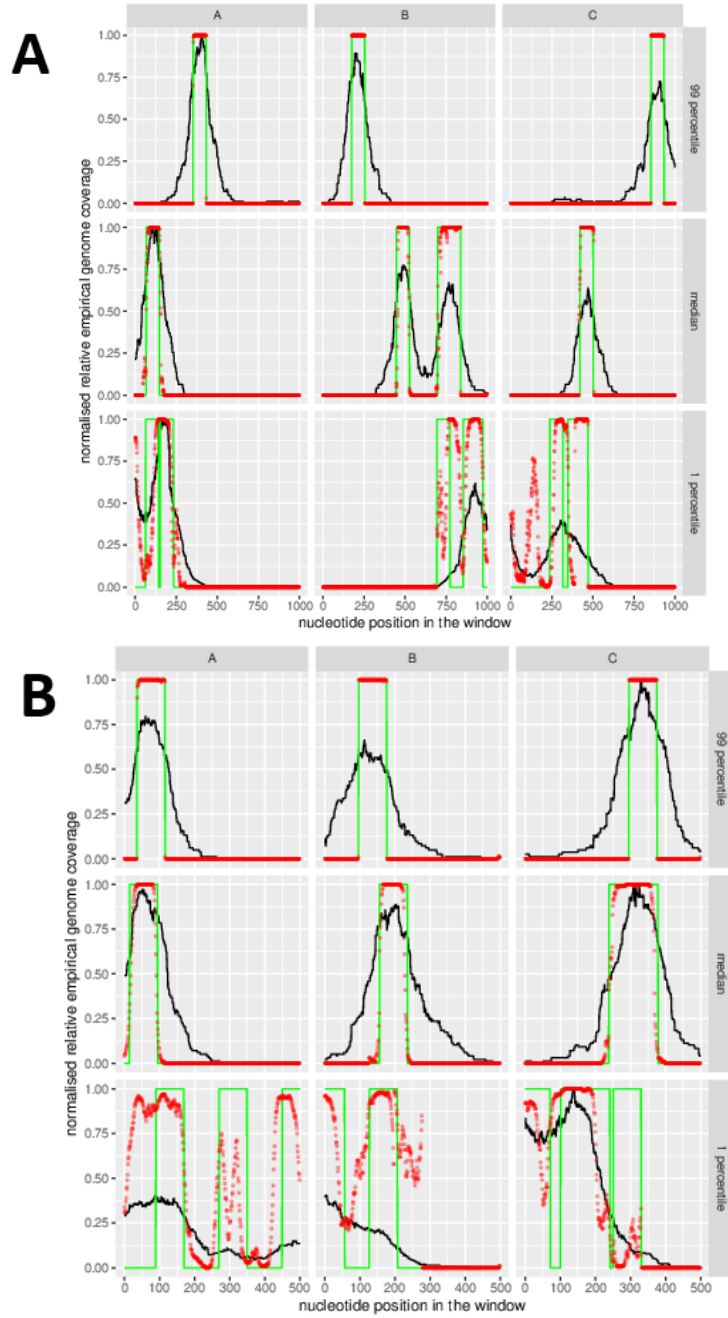


Figure 15: Visualization of predicted examples of model 1 based on the ranking of the F1 scores of 1 M individual predictions using the 1000 (A) and 500 (B) base pair data segment respectively. Subpanels 99th, median, and 1st percentile are 3 different (A, B, C) prediction examples with their appropriate F1 score rank. The black lines show the relative coverage in the data segment, the red dots are the probability of bait positions, and the green line represents the bait positions (value 1 for bait position, value 0 for no bait position). The lines are used to better visualize the start and end of bait regions in the data segment).

Thesis Publications

Table 5 summarizes the publications associated with this thesis and their relationship to the corresponding research topics.

Table 5: *Relation between theses and the corresponding publications.*

No	Publication	Year	Thesis				
			1	2	3	4	5
1	[J1] Information	2022	*				
2	[J2] Applied Network Science	2024	*				
3	[C1] SOR	2023	*				
4	[J3] JOP	2017		*			
5	[J4] Applied Sciences	2023		*			
6	[J5] Applied Network Science	2023			*		
7	[C2] IEEE Access	2017				*	
8	[J6] Information	2023				*	
9	[J7] BMC Bioinformatics	2024					*

Journal publications

- [J1] **Peter Juma Ochieng**, András London, and Miklós Krész. A Forward-Looking Approach to Compare Ranking Methods for Sports. *Information*, 13(5), 232, 2022. doi:10.3390/info13050232
- [J2] **Peter Juma Ochieng**, József Dombi, Miklós Krész, Tibor Kalmar, and Zoltán Maróti. A Graph-Based Approach for Prioritization of Related Cancer Genes. *Applied Network Science*, 2024. **In Review.**
- [J3] **Peter Juma Ochieng**, Wisnu Ananta Kusuma, and Toto Haryanto. Detection of Protein Complexes from Protein-Protein Interaction Networks Using Markov Clustering. *Journal of Physics*, 835(1):1-13, 2017. doi:10.1088/1742-6596/835/1/012001
- [J4] **Peter Juma Ochieng**, József Dombi, Miklós Krész, and Tibor Kalmar. A Special Structural-Based Weighted Network Approach for the Analysis of Protein Complexes. *Applied Sciences*, 13(11), 6388, 2023. doi:10.3390/app13116388
- [J5] **Peter Juma Ochieng**, Abrar Hussain, József Dombi, and Miklós Krész. An efficient weighted network centrality approach for exploring mechanisms of action of the Ruellia herbal formula for treating rheumatoid arthritis. *Applied Network Science*, 8, 7 (2023). doi:10.1007/s41109-022-00527-2

- [J6] **Peter Juma Ochieng**, Zoltán Maróti, József Dombi, Miklós Krész, József Békési, and Tibor Kalmar. Adaptive Savitzky–Golay Filters for Analysis of Copy Number Variation Peaks from Whole-Exome Sequencing Data. *Information*, 14(2), 128, 2023. doi:10.3390/info14020128
- [J7] Zoltán Maróti, **Peter Juma Ochieng**, Miklós Krész, József Dombi, and Tibor Kalmar. Optimizing Sequence Data Analysis Using Convolution Neural Networks for the Prediction of CNV Bait Positions. *BMC Bioinformatics*, 25, 389 (2024). doi:10.1186/s12859-024-06006-y

Conference proceedings

- [C1] **Peter Juma Ochieng**, József Dombi, András London, Miklós Krész, Tibor Kalmar, and Zoltán Maróti. Graph-Based Prioritization of Related Cancer Genes. In *The 17th International Symposium on Operations Research in Slovenia (SOR)*, 2023.
- [C2] **Peter Juma Ochieng**, Sri Ita Tarigan, and Hendrik Didik. A Clustering Model for Identification of Time-Course Gene Expression Patterns. In *1st International Conference on Biomedical Engineering (IBIOMED)*, IEEE, 1-6, 2016. doi:10.1109/IBIOMED.2016.7869819

Further related publications

- [C3] **Peter Juma Ochieng**, Kani, Hastuadi Harsa. Fingerprint authentication system using back-propagation with downsampling technique. In *2016 2nd International Conference on Science and Technology-Computer (ICST)*, IEEE, 182-187, 2016. doi:10.1109/ICSTC.2016.787737
- [C4] **Peter Juma Ochieng**, Taufik Djatna, Wisnu Ananta Kusuma. Tandem repeats analysis in DNA sequences based on improved Burrows-Wheeler transform. In *International Conference on Advanced Computer Science and Information Systems (ICACISIS)*, IEEE, 117-122, 2015. doi:10.1109/ICACISIS.2015.7415159
- [J8] **Peter Juma Ochieng**, Wisnu Ananta Kusuma, Mohmad Rafi, Tony Sumaryada. Deciphering the action mechanism of Indonesia herbal decoction in the treatment of type II diabetes using a network pharmacology approach. *Int J Pharm Pharm Sci*, 9(3), 243-53, 2017. doi=59/ijpps.2017v9i3.16413
- [J9] **Peter Juma Ochieng**, Tony Sumaryada, Daniel Okun. Molecular docking and pharmacokinetic prediction of herbal derivatives as maltase-glucoamylase inhibitor. *Asian J Pharm Clin Res*, 10(9), 392-98, 2017.

Bibliography

- [1] B. Chen, W. Fan, J. Liu, and F.-X. Wu. Identifying protein complexes and functional modules—from static PPI networks to dynamic PPI networks. *Briefings in Bioinformatics*, 15(2):177–194, 2014.
- [2] A. Cho, J. E. Shim, E. Kim, F. Supek, B. Lehner, and I. Lee. MUFFINN: Cancer gene discovery via network analysis of somatic mutation data. *Genome Biology*, 17(1):1–16, 2016.
- [3] C. S. Greene, J. Tan, M. Ung, J. H. Moore, and C. Cheng. Big data bioinformatics. *Journal of Cellular Physiology*, 229(12):1896–1900, 2014.
- [4] R. H. Jones. Bayesian information criterion for longitudinal and clustered data. *Statistics in Medicine*, 30(25):3050–3056, 2011.
- [5] A. N. Langville and C. D. Meyer. *Who’s #1?: The Science of Rating and Ranking*. Princeton University Press, 2012.
- [6] Y. Li and L. Chen. Big biological data: Challenges and opportunities. *Genomics, Proteomics Bioinformatics*, 12(5):187–189, 2014.
- [7] O. A. Montesinos López, A. Montesinos López, and J. Crossa. Random forest for genomic prediction. In *Multivariate Statistical Machine Learning Methods for Genomic Prediction*, pages 633–681. Springer, 2022.
- [8] S. K. Ng, T. Krishnan, and G. J. McLachlan. The EM algorithm. In *Handbook of Computational Statistics: Concepts and Methods*, pages 139–172, 2012.
- [9] X. Xu, Z. Qi, D. Zhang, M. Zhang, Y. Ren, and Z. Geng. DriverGenePathway: Identifying driver genes and driver pathways in cancer based on MutSigCV and statistical methods. *Computational and Structural Biotechnology Journal*, 21:3124–3135, 2023. 5

Summary in Hungarian

A disszertáció a komplex biológiai adatok elemzésre kifejlesztett új bioinformatikai módszereket mutatja be. Először egy új gráf-alapú keretrendszer kerül bemutatásra a potenciálisan rákot okozó gének priorizálására. Az innovációk között szerepel egy új hálózati rangsorolás stabilitási számítás, új normalizált, súlyozott mutáció burden számítási módszerek, valamint egy új gráf alapú priorizációs technika bevezetése, amely a konszenzus interakciós hálózatokban az egymással szomszédos gének mutáció terjedési erősségének (spreading strength) és befolyásának (influence) számításán, és egy módosított PageRank algoritmus alkalmazásán alapul. Az új módszertan javítja a gének rangsorolási stabilitását, és a prioritizált gének érzékenyebb és specifikusabb analízisét, így hozzájárul a rákban potenciálisan szerepet játszó gének pontosabb azonosításához. Másodszor, két megközelítést dolgoztak ki a fehérjekomplexek detektálására a fehérje-fehérje interakciós (PPI) hálózatokban. A Markov-klaszterezési módszer sűrűn kapcsolódó régiókat azonosít, míg a Weighted Edge Core-Attachment and Local Modularity (WECALM) megközelítés a magfehérjéket és kapcsolódásaikat detektálja, javítva a fehérje komplexek kimutatásának pontosságát és hatékonyságát. Harmadszor, egy számítógépes keretrendszer került kifejlesztésre gyógyszer-célpontok előrejelzésére összetett biológiai hálózatokban. A súlyozott centralitásmérőket és hálózatépítési technikákat alkalmazva az új keretrendszer integrálja az omikai adatokat a gyógyszer hatás mechanizmusok tanulmányozásához, amelyet jelenleg a Reuelia gyógynövény reumás ízületi gyulladásra gyakorolt farmakológiai hatásainak vizsgálatában is alkalmaznak. Negyedszer, módszereket dolgoztak ki az időbeli génexpresszió és a kópia szám eltérések (CNV) pontosabb elemzésére. A Rejection Control Expectation Maximization (RCEM) modell finomítja a zajos időbeli génexpressziós klaszterezést, míg egy adaptív Savitzky-Golay szűrő javítja a CNV detektálásának pontosságát az egész-exom szekvenálási adatokban. Végül egy 1D Konvolúciós Neurális Hálózat (CNN) keretrendszer került bemutatásra a célzott hibridizációs (targeted capture) teljes exome (WES) adatok elemzésében. Ez a mélytanulási megközelítés az experimentális lefedettségi profilok, a szekvencia adatok, és a target régiók elemzésével prediktálja a gyártók által legtöbbször nem közölt oligo capture baitek legvalószínűbb genom koordinátáit. A prediktált oligo capture koordináták lehetővé teszik a lefedettségi adatokban jelen levő szisztematikus torzítások (például a GC bias) jobb normalizációját, így érzékenyebb és specifikusabb CNV detekciót tesz lehetővé, előmozdítva a WES adatok pontosabb analízisét. Ezek az eredmények robusztus számítógépes eszközöket biztosítanak a rákkutatáshoz, a gyógyszerfejlesztéshez és a genomi elemzésekhez, új betekintést nyújtva az összetett biológiai rendszerek működésébe.

Declaration

In the PhD dissertation of Peter Juma Ochieng, titled "*Applications of Network Analysis and Machine Learning Methods in Bioinformatics*", Peter Juma Ochieng and his supervisor share the following joint and undivided contributions:

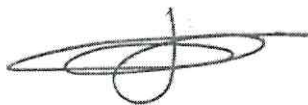
- Developed a graph-based approach for prioritizing cancer genes; designed network-based methods to detect protein complexes in protein-protein interaction (PPI) networks; proposed a new weighted centrality method for exploring drug targets and mechanisms of action; developed computational frameworks for analyzing time-course gene expression and copy number variation (CNV) peaks in whole-exome sequencing (WES) data; designed a deep learning model based on a 1D convolutional neural network (CNN) for predicting CNV bait positions [J1, J2, J3, J4, J5, J6, J7, C1, C2].

In the PhD dissertation of Peter Juma Ochieng, the following results represent his individual and decisive contributions:

- **Thesis 1:** Developed a forward-looking stability measure matrix to evaluate ranking stability and introduced the Graph-Based Prioritization with PageRank (GBPPR) method for prioritizing cancer-related genes [J1, J2, C1].
- **Thesis 2:** Developed protein complex detection methods for PPI networks, utilizing the Markov Clustering approach and the Weighted Edge Core-Attachment and Local Modularity (WECALM) structures [J3, J4].
- **Thesis 3:** Deployed a weighted network centrality approach for predicting drug targets in complex biological networks [J5].
- **Thesis 4:** Formulated computational methods for analyzing time-course gene expression patterns and CNV peaks in WES data [C2, J6].
- **Thesis 5:** Designed a deep learning model using a 1D Convolutional Neural Network (CNN) to optimize sequence data analysis for predicting CNV bait positions [J7].

This work cannot be used to obtain any academic research degree other than the submitted PhD dissertation of Peter Juma Ochieng.

Szeged, 2025.03.03



Peter Juma Ochieng
PhD candidate

MIKLÓS
FERENC KRÉSZ

Miklós Krész
Supervisor

Digitally signed by
MIKLÓS FERENC KRÉSZ
Date: 2025.03.05
16:27:37 +01'00'

The head of the Doctoral School of Computer Science declares that the statement above was sent to all co-authors, and none of them raised any objections.

Szeged, 2025.03.03



Mark Jelasity, DSc.
Head of Doctoral School

- [J1] Peter Juma Ochieng, Andr'as London, and Mikl'os Kr'esz. A Forward-Looking Approach to Compare Ranking Methods for Sports. *Information*, 13(5), 232, 2022. doi:10.3390/info13050232
- [J2] Peter Juma Ochieng, J'ozsef Dombi, Mikl'os Kr'esz, Tibor Kalmar, and Zolt'an Mar'oti. GBP-PR: A Graph-Based Computational Approach for Prioritizing Significantly Related Cancer Genes Across Multiple Cancer Types. *Applied Network Science*, 2023. In Review.
- [J3] Peter Juma Ochieng, Wisnu Ananta Kusuma, and Toto Haryanto. Detection of Protein Complexes from Protein-Protein Interaction Networks Using Markov Clustering. *Journal of Physics*, 835(1):1-13, 2017. doi:10.1088/1742-6596/835/1/012001
- [J4] Peter Juma Ochieng, J'ozsef Dombi, Mikl'os Kr'esz, and Tibor Kalmar. A Special Structural-Based Weighted Network Approach for the Analysis of Protein Complexes. *Applied Sciences*, 13(11), 6388, 2023. doi:10.3390/app13116388
- [J5] Peter Juma Ochieng, Abrar Hussain, J'ozsef Dombi, and Mikl'os Kr'esz. An efficient weighted network centrality approach for exploring mechanisms of action of the Ruellia herbal formula for treating rheumatoid arthritis. *Applied Network Science*, 8, 7 (2023). doi:10.1007/s41109-022-00527-2
- [J6] Peter Juma Ochieng, Zolt'an Mar'oti, J'ozsef Dombi, Mikl'os Kr'esz, J'ozsef B'ek'esi, and Tibor Kalmar. Adaptive Savitzky–Golay Filters for Analysis of Copy Number Variation Peaks from Whole-Exome Sequencing Data. *Information*, 14(2), 128, 2023. doi:10.3390/info14020128
- [J7] Zolt'an Mar'oti, Peter Juma Ochieng, Mikl'os Kr'esz, J'ozsef Dombi, and Tibor Kalmar. Optimizing Sequence Data Analysis Using Convolution Neural Networks for the Prediction of CNV Bait Positions. *BMC Bioinformatics*, 25, 389 (2024). doi:10.1186/s12859-024-06006-y
- [C1] Peter Juma Ochieng, J'ozsef Dombi, Andr'as London, Mikl'os Kr'esz, Tibor Kalmar, and Zolt'an Mar'oti. Graph-Based Prioritization of Related Cancer Genes. In *The 17th International Symposium on Operations Research in Slovenia (SOR)*, 2023.
- [C2] Peter Juma Ochieng, Sri Ita Tarigan, and Hendrik Didik. A Clustering Model for Identification of Time-Course Gene Expression Patterns. In *1st International Conference on Biomedical Engineering (IBIOMED)*, IEEE, 1-6, 2016. doi:10.1109/IBIOMED.2016.7869819