Received: 19 June 2024 | Revised: 21 August 2024 | Accepted: 3 September 2024 | Published online: 10 September 2024

RESEARCH ARTICLE

Medinformatics
2024, Vol. 1(4) 211–221
DOI: 10.47852/bonviewMEDIN42023658

BON VIEW PUBLISHING

ACO-Guided Genetic Analysis: Exploring the Role of Genes in Alzheimer's Disease

Chengjiang Zhu[1], Lin Yang[1], Xudong Huang[2,3], Kunpei Jin[1], Xinping Pang[4], Xianghu Song[1], Yue Sun[1], Chonghao Gao[1], Yanyu Wei[5,]* and Chaoyang Pang[1,]*

[1]College of Computer Science, Sichuan Normal University, China

[2]Department of Psychiatry, Massachusetts General Hospital, USA

[3]Harvard Medical School, USA

[4]School of Science, Xi'an Jiaotong-Liverpool University, China

[5]School of Electronic Science and Engineering, University of Electronic Science and Technology of China, China

Abstract: Alzheimer's disease (AD), a neurodegenerative disorder significantly impairing cognitive function and quality of life, necessitates an in-depth study of its pathogenic mechanisms. Despite extensive research, the disease's pathological mechanisms harbor numerous unknown elements. Analyzing DNA microarray data is crucial for elucidating gene co-expression relationships during AD's pathological processes. This study aimed at exploring the applicability of the ant colony optimization (ACO) algorithm in identifying linearly co-expressed genes in AD. This study utilized gene expression profile data of AD patients collected from the Gene Expression Omnibus (GEO) database, employing the ACO algorithm to explore the co-expression relationships among AD pathogenic genes in different stages. Through the analysis of co-expression relationships, we identified four related genes (SDHA, NDUFA10, SDHC, and GPI). By using the co-expressed genes as features, we applied the support vector machine (SVM) algorithm for AD classification. The results of these experiments revealed that each model achieved average AUC values of 0.90, 0.91, 0.86, and 0.92, respectively. Our research findings reveal that the ACO algorithm provides a new perspective for in-depth exploration of the pathogenic mechanisms of AD.

Keywords: ant colony optimization, Alzheimer's disease, gene order, co-expression relationships, support vector machine

1. Introduction

Alzheimer's disease (AD) is one of the most common pathological types among the elderly [1]. The disease is characterized by an early onset of significant memory decline, gradually leading to a loss of daily life capabilities, accompanied by mental symptoms and behavioral disorders [2–4]. The progression of AD can be categorized into four stages: control stage, incipient stage, moderate stage, and severe stage [5]. In the later stages, patients may face severe conditions such as difficulty swallowing and bedridden status, with complications like infections increasing the risk of mortality [5, 6]. According to statistics from Western countries, the incidence of AD in individuals aged 65 and older is approximately 5%, while in those aged 85 and older the incidence can exceed 30% [7]. Globally, an estimated 50 million people suffer from AD, with approximately 10 million patients in China [8]. Projections indicate that by 2050, there will be approximately 130 million AD patients worldwide [8, 9].

The pathological mechanism of AD is extraordinarily complex and remains a focal point of extensive research [10]. Nevertheless, researchers have gained crucial insights into this disease. Its primary hallmark involves the abnormal accumulation of Aβ protein and Tau protein in the brain, leading to neuronal damage and cell death [11]. Aβ protein aggregates in the form of amyloid plaques, while Tau protein forms tangles that disrupt normal neuronal function [11]. The pathogenesis of AD encompasses multiple intricate factors, including protein abnormalities, cellular damage, inflammation, oxidative stress, neurotransmitter alterations, and vascular issues [12, 13]. These factors interact with each other, ultimately resulting in neuronal damage and cell death [12, 14, 15].

To explore the complex factors involved in the pathogenesis of AD, the observation of gene expression levels in different stages of AD patients is made possible through a new technique known as DNA microarray technology [16]. This technology allows gene information to be expressed as digital data, referred to as expression levels [17, 18]. For instance, if two genes serve similar functions, their expression levels are likely to be similar as well [19]. Therefore, by analyzing DNA microarray data, gene characteristics can be identified. DNA microarray technology is high throughput, enabling the simultaneous generation of expression levels for thousands of genes, whereas in conventional

***Corresponding authors:** Yanyu Wei, School of Electronic Science and Engineering, University of Electronic Science and Technology of China, China. Email: yywei@uestc.edu.cn and Chaoyang Pang, College of Computer Science, Sichuan Normal University, China. Email: cypang@sicnu.edu.cn

biological experiments, only one gene can be studied at a time. This advantage accelerates gene research and has popularized DNA microarray technology in gene analysis, particularly in the study of AD-related genes over the past decade [20–22].

2. Materials and Methods

2.1. Data source and data organization

All gene expression datasets utilized in this study were obtained from the Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/browse). The GSE5281 dataset was selected for analysis. All data are freely available online. In fact, these databases originate from different platforms, brain regions, age groups and genders, which facilitates the identification of genes exhibiting common significant differences.

The GSE5281 dataset was obtained from brain tissue slices using the GPL570 platform, which included brain tissues from the olfactory cortex, hippocampus, temporal gyrus, posterior cingulate gyrus, superior frontal gyrus, and primary visual cortex. Gene expression analysis was performed on laser-captured cells using the Affymetrix U133 Plus 2.0 chip (approximately 55,000 transcripts). The dataset comprises 54,675 genes and 161 samples, of which 74 are control samples and 87 are samples of AD patients. These samples were divided into two independent matrices for research and analysis: the control group and the AD group. The data from the control group are presented in a matrix denoted as M_{ctrl} in this paper, with dimensions of 74 columns by 54,675 rows. These 74 columns are generated by 74 gene chips, with each column containing 54,675 data points. Each data point represents the expression level of a specific gene (the $i-th$ row data corresponds to the $i-th$ gene). The same organizational method is employed to create the M_{ad} matrix, which contains 87 columns of data from the disease stage. Each row in the M_{ctrl} matrix represents a 74-dimensional vector, comprising 74 expression levels generated under different conditions. The *i-th* row vector corresponds to the *i-th* gene, containing hidden information about that gene. Each row vector within the same matrix can be characterized as a k-dimensional vector. The squared Euclidean distance between two vectors, $X = (x_1, x_2, \ldots, x_k)$ and $Y = (y_1, y_2, \ldots, y_k)$, is defined as follows:

$$D_E(X, Y) = \sum_{i=1}^{k} (x_i - y_i)^2 \tag{1}$$

If two genes have similar functions, their expression levels are likely to be similar as well. Therefore, analyzing gene expression levels can identify co-expressed genes, which is crucial for understanding complex biological processes. Specifically, we evaluate the differences between gene expression levels by calculating the squared Euclidean distance between them. A smaller squared Euclidean distance indicates a smaller difference in expression levels between the two genes, suggesting a higher functional similarity. This method enables more precise identification of potential co-expressed genes, thereby providing a solid foundation for subsequent biological validation and functional studies.

2.2. Gene order

Based on DNA microarray data, a concept called gene order is presented. All genes are arranged in a line (sequence or circle), and similar genes are arranged together, the resulting sequence is called gene order. The similarity of any two genes is measured by the difference in their expression levels. The smaller the difference, the more similar the two genes are. Under different experimental conditions, different expression levels are generated. For the same gene, there is a multidimensional vector corresponding to it, and each component is an expression level generated under different experimental conditions. The distance of the vectors measures the similarity of two of the corresponding genes under all experimental conditions. The smaller the distance, the more similar the two corresponding genes are.

Gene order can be characterized as the route of the traveling salesman problem (TSP) [23], where each gene is characterized as a virtual city and its corresponding vector is the position of the virtual city. where TSP route refers to the route where a salesman visits each city exactly once and returns to the starting city [23]. The optimal gene order refers to the shortest TSP route [23]. Clearly, optimal gene order is an optimal linear permutation of all genes, where similar genes are clustered together and the clustering is globally optimal [23]. Thus, optimal gene order is attractive and some experts have paid attention to its computational method in the last decade [23]. As in Dahiya and Sangwan [24], HK Tsai et al. improved Genetic Algorithm (GA) to calculate gene order.

2.3. Co-expression gene screening

If a gene is very close in distance to another gene, it may be co-expressed with that gene. The degree of co-expression is typically measured by the distance.

For matrix M_{ctrl}, the corresponding optimal gene order is calculated using the ant colony optimization (ACO) method, denoted as G_{ctrl}. Let's assume that G_{ctrl} is the following sequence, which includes all the genes:

$$G_{ctrl} = \{g_1, g_2, g_3, \ldots . g_i, \ldots, g_n\} \tag{2}$$

where g_i denotes the $i-th$ gene and the number of genes is n.

Suppose gene g_i corresponds to vector $X(g_i)$ (i.e., the $i-th$ line vector in matrix), and the vector corresponding to gene g_j is denoted by $X(g_j)$.

Calculate distance $D_E(X(g_i), X(g_j))$ between gene g_i and g_j. Let

$$f_{ctrl}(i, j) = D_E(X(g_i), Xg_j)) \tag{3}$$

where $i, j = 1, 2, \cdots, n$.

Clearly, value of $f_{ctrl}(i, j)$ measures the degree of co-expression between gene g_i and g_j. The smaller it is, the stronger the co-expression is.

All function values form a data set

$$\{f_{ctrl}(1), f_{ctrl}(2), \ldots, f_{ctrl}(i), \ldots, f_{ctrl}(n)\} \tag{4}$$

And the above data set forms a curve of function, which is denoted by $f_{ctrl}(i, j)$ Clearly, the degree of co-expression between g_i and arbitrary gene is mapped into function $f_{ctrl}(i)$.

Figure 1 shows the curve of the function $f_{ctrl}(t)$. Figure 1 shows that the degree of co-expression between any gene and METTL3 is clearly and accurately displayed. Compared to the heat map of gene order, the function curve acts like a magnifying glass, providing a clear view of the co-expression details. This characteristic compensates for the shortcomings of gene order and other clustering methods.

On the other hand, it is easy to determine the co-expressed genes of a specific gene like METTL3 using the curve: if a gene is positioned around METTL3 in the gene order (i.e., it is a neighbor of METTL3) and has a relatively small function value, then it is considered a co-expressed gene. Let's assume that all co-expressed genes form the set C_{ctrl}.

Using the method described above, for the disease data (i.e., matrices M_{ad}), we can generate corresponding functions, denoted

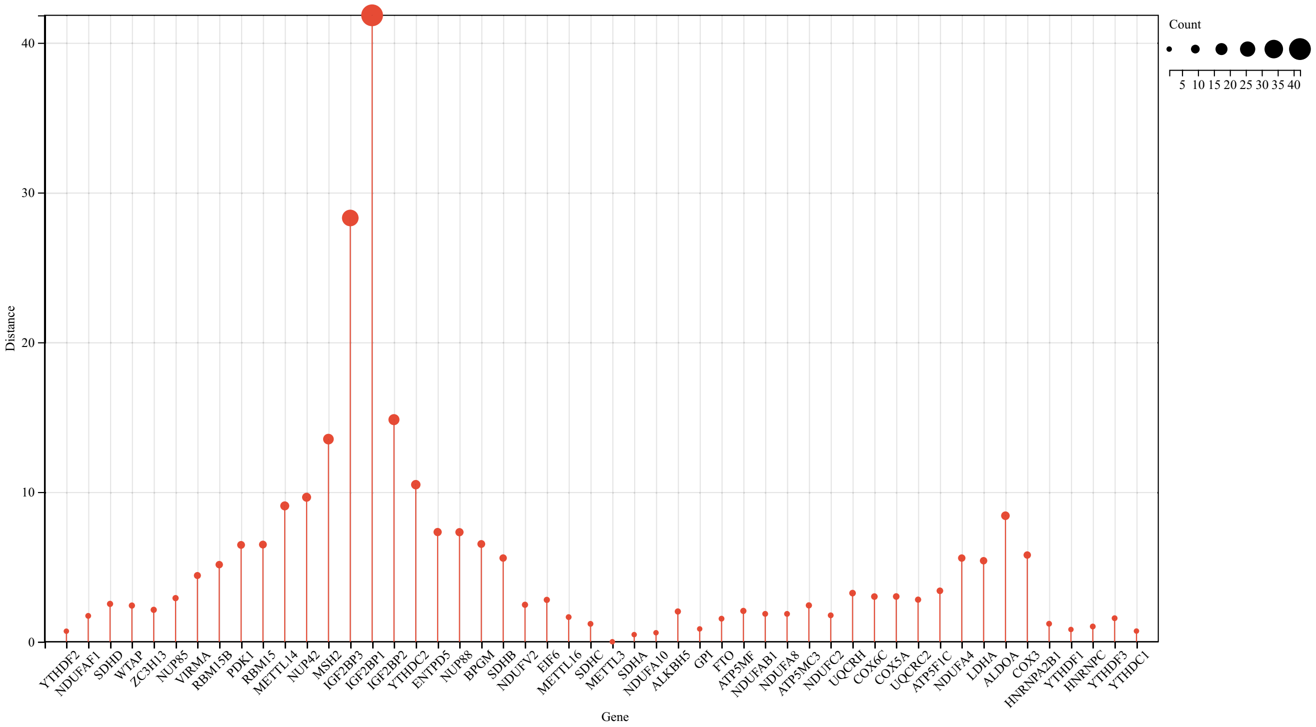

Figure 1. Identify genes with similar functionality based on co-expression levels. Smaller values correspond to reduced disparities, signifying heightened co-expression levels. The abscissa represents gene sequences computed using the ACO algorithm, while the ordinate represents the co-expression function values of each gene with respect to the METTL3 gene

as $f_{ad}(t)$. Corresponding sets of co-expressed genes can be obtained, denoted as C_{ad}. Let

$$C(METTL3) = C_{ctrl} \cap C_{ad} \tag{5}$$

The set $C(METTL3)$ includes genes that maintain a co-expressed state with the METTL3 gene consistently throughout the development of AD (from normal to disease stages). Genes included in the $C(METTL3)$ set may be associated with AD since they consistently co-express with METTL3.

2.4. ACO

In nature, ants search for food and return to their nests. While searching for paths, they release chemicals (pheromones) to mark the paths they have taken. These pheromones attract other ants to follow the same paths. This collective behavior eventually leads the entire ant colony to find the source of food.

The basic principle of the ACO algorithm is to simulate the movement of ants in a search space, where each ant selects a path based on its previous experience and the pheromone information. It utilizes the searching behavior of ants in the problem space, with ants leaving behind pheromone trails during their search, which influences the decisions of other ants.

The ACO algorithm introduces the concept of pheromones, which represent the quality of paths. Ants release pheromones on the paths they traverse, and the concentration of pheromones on a path is directly proportional to the path's quality. Pheromones are spread among the ant colony through the paths and affect ant path selection.

Ants randomly choose paths in the problem space. They tend to prefer paths with higher pheromone concentrations, as it indicates that other ants have found better solutions on those paths. However, the ACO algorithm includes exploration mechanisms to ensure that ants are not limited to known good paths.

After completing a search, pheromones are updated based on the search results. Typically, pheromones gradually evaporate over time to prevent getting trapped in local optima. Ants increase or decrease the concentration of pheromones on paths based on their search results.

The ACO algorithm achieves global search through cooperation among ants. Ants leave pheromones during the search, guiding the decisions of other ants. Gradually, the ant colony concentrates around the best solutions to the problem because paths with higher pheromone concentrations become more attractive.

Below, we simulate the movement of ants by solving the TSP to illustrate the ACO algorithm model. Table 1 summarizes the symbols required for the following discussion and provides explanations for each.

Suppose m ants are randomly placed on n cities, where d_{ij} represents the distance between city i and city j, and $\tau_{ij}(t)$ represents the strength of information pheromone on the path between city i and

Table 1. The main symbols used in ant colony optimization

Symbols	Description
d_{ij}	The Euclidean distance from city i to city j
$\tau_{ij}(t)$	The strength of information pheromone on the path (i, j) at the $t - th$ iteration
$\eta_{ij}(t)$	The heuristic function, for a given TSP problem, $\eta_{ij}(t)$ is a constant, and its expression is $\frac{1}{d_{ij}}$.
$\Delta\tau_{ij}(t)$	The increment of information along the path (i, j) during the $t - th$ iteration.
$p_{ij}^{k}(t)$	The state transition probability of ant k moving from city i to city j at the $t - th$ iteration.
$tabu_k$	Record the cities already visited by ant k.

city j at the $t-th$ iteration. Initially, all paths have equal information, and let $\tau_{ij}(0) = Const$ (where $Const$ is a positive constant). Ant $k(k = 1, 2, \ldots, m)$ determines its transition direction during movement based on the information on each path and the path length. Here, a tabu list $tabu_k(k = 1, 2, \ldots, m)$ is used to record the cities visited by ant k at the current moment, and the set $tabu_k$ is dynamically adjusted as the ant colony search process progresses. $p_{ij}^k(t)$ represents the state transition probability of ant k moving from city i to city j at the $t-th$ iteration, defined as follows:

$$p_{ij}^k(x) = \begin{cases} \frac{\tau_{ij}^{\alpha}(t)\cdot\eta_{ij}^{\beta}(t)}{\sum_{s \in \text{allowed}_k} \tau_{is}^{\alpha}(t)\cdot\eta_{is}^{\beta}(t)}, & j \in \text{allowed}_k \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

In this context, $allowed_k$ represents the set of cities that ant k is allowed to choose for its next step. α is the pheromone heuristic factor, which signifies the relative importance of the trail and reflects the role of pheromones in guiding the ant's movement. A higher α value makes the ant more inclined to choose paths that other ants have traveled, enhancing the cooperation within the ant colony. β is the visibility heuristic factor, representing the relative importance of visibility, and it indicates the emphasis placed on heuristic information in the ant's path selection. A higher β value makes the state transition probability closer to the greedy rule. $\eta_{ij}(t)$ is the heuristic function, which, for a given TSP problem, is a constant. Its expression is as follows:

$$\eta_{ij}(t) = \frac{1}{d_{ij}} \tag{7}$$

where d_{ij} represents the distance between city i and city j, and

$$d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \tag{8}$$

where (x_i, y_i) and (x_j, y_j) are the coordinates of cities i and j, respectively. For ant k, the smaller d_{ij}, the larger $\eta_{ij}(t)$, and thus, $p_{ij}^k(t)$ is greater. This heuristic function reflects the degree of inspiration from city i to city j for ant k.

To avoid an excessive accumulation of residual pheromones that might drown out the heuristic information, the residual pheromones need to be updated after each step of an ant or at the end of a complete cycle. As a result, the amount of information on the $path(i,j)$ at the $t-th$ iteration can be adjusted according to the following rule:

$$\begin{aligned} \tau_{ij}(t+1) &= (1-\rho)\tau_{ij}(t) + \Delta\tau_{ij}(t) \\ \Delta\tau_{ij}(t) &= \sum_{i=1}^{m} \Delta\tau_{ij}^k(t) \end{aligned} \tag{9}$$

where ρ represents the pheromone evaporation coefficient, and $1 - \rho$ represents the pheromone residual factor. To prevent the infinite accumulation of pheromones, the range of ρ is $\rho \in (0, 1)$. $\Delta\tau_{ij}(t)$ represents the increment of pheromones on the $path(i,j)$ during the current cycle. At the initial time, the increment of pheromones is 0, that is, $\Delta\tau_{ij}(0) = 0$. $\Delta\tau_{ij}^k(t)$ represents the amount of pheromones left on the path (i,j) by the $k-th$ ant during the $t-th$ cycle.

According to different pheromone update strategies, the ACO algorithm can be divided into three types, namely the Ant-Cycle model, Ant-Density model, and Ant-Quantity model.

In the Ant-Cycle model,

$$\Delta\tau_{ij}^k(t) = \begin{cases} \frac{Q}{L_k}, & \text{If the } k-th \text{ ant traverses the path } (i,j) \text{ in the } t-th \text{ iteration} \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

In the equation, L_k represents the total length of the path traversed by the $k-th$ ant in this cycle.

In the Ant-Density model,

$$\Delta\tau_{ij}^k(t) = \begin{cases} Q, & \text{If the } k-th \text{ ant traverses the path } (i,j) \text{ in the } t-t\ h \text{ iteration} \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

In the Ant-Quantity model,

$$\Delta\tau_{ij}^k(t) = \begin{cases} \frac{Q}{d_{ij}}, & \text{If the } k-th \text{ ant traverses the path } (i,j) \text{ in the } t-th \text{ iteration} \\ 0, & \text{otherwise} \end{cases} \tag{12}$$

In the equation, Q is a constant representing the strength of information pheromone. The difference between these three models lies in the fact that the Ant-Quantity and Ant-Density models utilize local information updates, meaning ants update the information on the paths they have just traversed after each step. On the other hand, the Ant-Cycle model employs global information updates, which means that after all ants complete one cycle, the information on all paths is updated. Among these models, the Ant-Cycle model performs relatively well in solving the TSP, and therefore, it is commonly used as the fundamental model for ant colony algorithms.

Algorithm 1 will use the Ant-Cycle model as an example to illustrate the specific implementation steps of the ant colony algorithm in solving the optimal gene sequence.

Algorithm 1: Ant colony optimization (ACO)

```
    Data: weighted graph and ant system parameters
    Result: optimal path
Initialization: Initialize pheromone trails for all paths among cities. Let
  $m$ ants position at different cities to travel. Pre-assign a maximum
  iteration number $t_{max}$ and let $t = 0$, where $t$ denotes the $t-th$ iteration.
  Initialize pheromone on path $(i,j)$ at the $t-th$ iteration $\tau_{ij}(0) =$
  $Const(Constisaconstant)$. The initial increment of information on path
  $(i,j)$ is 0, i.e., $\Delta\tau_{ij}(0) = 0$.
while $t \leq t_{max}$ do
    foreach ant do
        foreach city do
            Ant $k$ selects the next city based on the formula (6) and
              places city into ant $k$'s tabu list $tabu_k$.
    Update pheromone values for all edges according to formula
      (10), formula (11), and formula (12);
    Increment the iteration t by 1;
Select the path with the shortest length as the output;
```

2.5. Gene sequence calculated by ACO

In this paper, 50 genes were selected from 54675 genes for testing. To identify genes that change with METTL3 gene alterations, for the selected dataset (GSE5281), AD samples were divided into METTL3 high-expression and METTL3 low-expression subgroups, with the median expression of METTL3 chosen as the cutoff value. Differential gene expression analysis was conducted using the lmFit and eBayes methods for both AD patients, the control group as well as the METTL3 high-expression and METTL3 low-expression subgroups. The “WGCNA” package (version 1.71) was used to identify METTL3-related AD genes in the expression data from GSE5281. Finally, to associate modules with features, modules primarily associated with METTL3_High or METTL3_Low subgroups were defined as key modules and selected for further filtering. The “STRINGdb” package (version 2.4.2) was used for protein-protein interaction analysis of genes in the key modules. Network data was visualized using the “ggraph” function. Functional enrichment analysis was performed on key module genes and key AD genes, and genes from the most valuable pathways were selected as key AD genes (i.e., the 50 genes used in this paper).

We approached the issue of gene interrelations as a TSP. Specifically, the selected 50 genes were analogized to 50 cities, and the squared Euclidean distances between these “cities” (i.e., genes) were computed. Subsequently, the ACO algorithm was employed to determine the order of genes, aiming to cluster genes with similar expression levels together. Furthermore, the genes were arranged based on their correlation strength, with those exhibiting stronger correlations being placed in closer proximity to each other. This arrangement was evident in the visualizations provided by the heat maps and co-expression curves, which highlighted the strong correlations in the gene sequences post ACO algorithm arrangement. Additionally, by conducting an intersection analysis of relevant genes during the controllable and diseased stages of AD, we identified four key genes consistently related to the METTL3 gene. This identification offers a new perspective in understanding the molecular mechanisms underlying AD.

3. Results

3.1. ACO in gene order

In this research, ACO algorithm was utilized for the analysis of gene sequences. The initial phase of the study involved the application of heatmap technology to contrast the levels of gene expression before and after the implementation of ACO algorithm sorting. This approach effectively demonstrated the variations in gene expression patterns, thereby elucidating the dynamic distribution of gene expression levels under varying conditions. Subsequently, the research emphasis was shifted to the analysis of distance between the METTL3 gene, the focus of this experiment, and other genes, employing distance curves to disclose their interconnectedness at the expression level. Additionally, through repetitive experimental procedures, the research team conducted statistical analysis and validation of the gene set associated with METTL3, ensuring the repeatability and reliability of the experimental outcomes.

Further, SVM was employed in the study for predictive analysis, aiming to investigate the potential link between genes associated with METTL3, the focal point of this experiment, and AD. Ultimately, by analyzing the function curves, the research revealed changes in gene co-expression levels throughout the progression of AD. The application of these methodologies enabled the ACO algorithm to play a crucial role in this series of complex gene analyses, offering new scientific perspectives for a deeper understanding of intricate gene networks and their relationship with the disease.

Firstly, we employed heatmap technology to compare gene expression level data before and after the application of the ACO algorithm for gene sorting. The primary objective of this phase was to visually depict changes in gene expression patterns, thus providing a clear representation of the differences in gene expression levels before and after sorting. Through the observations presented in Figure 2, it becomes evident that the sorted gene sequences exhibit a higher degree of order, and the expression levels of neighboring genes become more similar. This outcome underscores the effectiveness of the ACO algorithm in gene sorting. This discovery establishes a more robust foundation, enabling us to delve deeper into the exploration of interrelationships among genes and their relevance to diseases. This process lays a solid groundwork for subsequent analyses.

Following this, we applied the results of gene sorting to construct distance curves. The horizontal axis represents the names of various genes, arranged according to the results of gene sorting, while the vertical axis represents the distance between these genes and the METTL3 gene. Visualizing the distance relationship between each gene and METTL3 in the form of distance curves, shorter distances indicate a higher degree of co-expression. As shown in Figure 3, this optimization process serves as a magnifying glass, clearly showcasing the co-expression details between any gene and METTL3. This work not only addresses the shortcomings of gene sorting and other clustering methods, allowing us to analyze the relationships between genes more finely, but also simplifies the identification of co-expressed genes for specific genes, such as METTL3 used in this experiment, leading to a more accurate understanding of the interactions between these genes.

In this study, we implemented the ACO algorithm for the computational analysis of gene sequences. By introducing a strategy of multiple repeated runs during the operation of the algorithm, we were able to achieve a stable and reliable solution on a statistical basis. This stability is reflected in the consistency of the solutions after multiple iterations, rather than relying on the output of a single run. Given the randomness and heuristic search characteristics of the ACO algorithm, it is unrealistic to expect the algorithm to produce a completely consistent and definite single optimal solution in each run. Therefore, this strategy of multiple repeated runs significantly enhances the statistical stability and reliability of the results.

In specific application scenarios, we utilized the ACO algorithm to compute gene sequences 10 times, resulting in 10 distinct gene sequences. Particularly, we focused on the interaction patterns of METTL3 gene with their neighboring genes (5 genes before and after in the gene sequence) and conducted a statistical analysis of genes that repeated more than 5 times within this proximity range. We analyzed the genes neighboring METTL3 during the controllable and onset stages of AD, identifying the gene intersections in these two conditions.

This study includes two statistical tables that record the proximity frequency of the METTL3 gene with its neighboring genes (five genes before and after in the gene sequence) during both the controllable and onset stages of AD. These tables correspond to the two different stages of AD, providing comprehensive data on the proximity frequency of METTL3 and its neighboring genes. Table 2 pertains to the controllable stage of AD, detailing the proximity frequency between METTL3 and its neighboring genes during this phase. Table 3 relates to the onset stage of AD, illustrating the proximity frequency between METTL3 and its neighboring genes during this stage.

3.2. SVM prediction analysis: Revealing gene associations with AD

Based on the intersection of data from Tables 2 and 3, we selected four significant genes, namely, SDHA, NDUFA10, SDHC, and GPI. Subsequently, we conducted SVM predictive

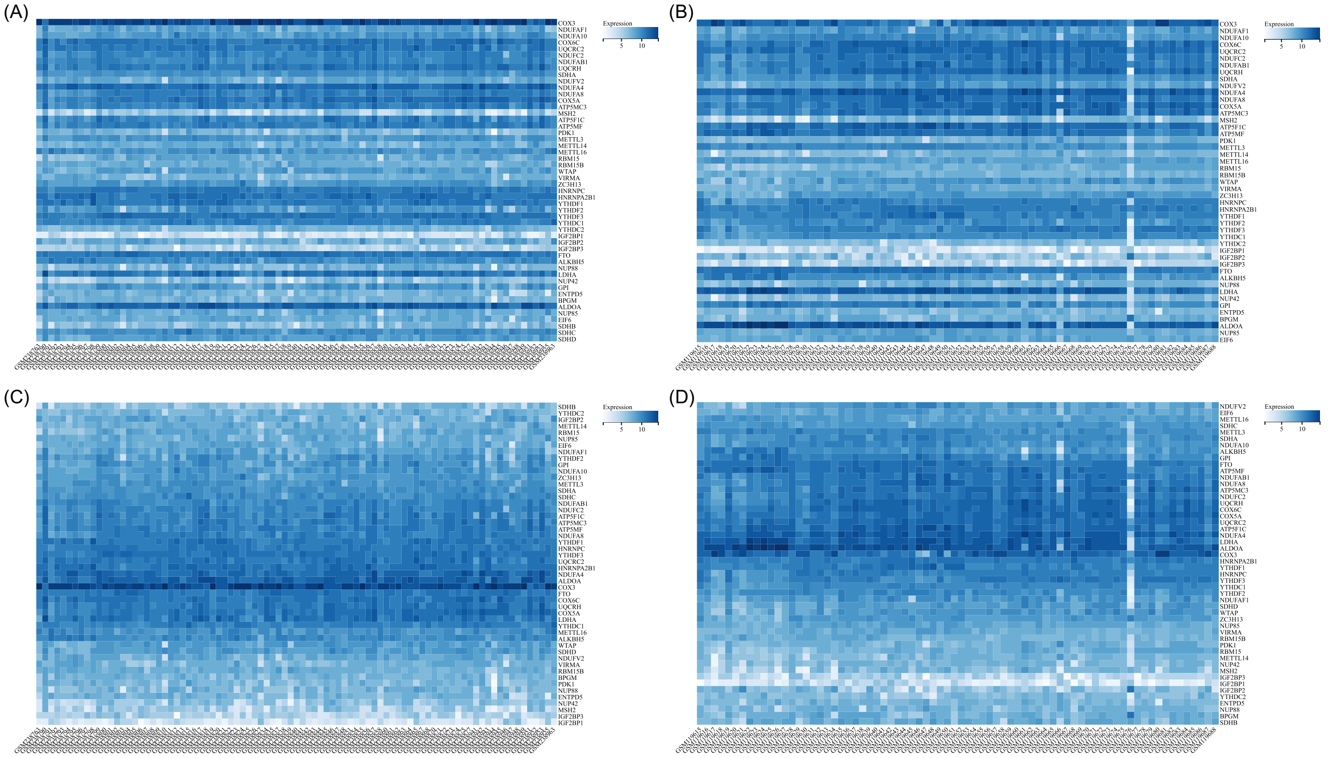

Figure 2. This set of images demonstrates the effectiveness of the ACO algorithm in gene sorting through heatmaps. (A) shows a heatmap of gene expression level data from the controllable phase of AD. (B) shows a heatmap of gene expression level data from the disease phase of Alzheimer's disease. (C) shows the gene expression level data from the controllable phase after sorting by the ACO algorithm. Finally, (D) shows the gene expression level data from the disease phase after sorting by the ACO algorithm.

analysis on these four genes. In the subsequent experiments, we employed the SVM algorithm to construct a predictive model aimed at investigating the association between the METTL3 and SDHA genes with AD. Through in-depth analysis of the SVM model, we observed its outstanding performance in classifying individuals as either AD patients or non-AD individuals. Similarly, we conducted multiple experiments in which we utilized METTL3 along with our selected related genes (NDUFA10, SDHC, and GPI) as features for SVM analysis. As depicted in Figure 4, these results demonstrate that METTL3, in combination with our selected related genes as features, exhibits exceptional classification capabilities within the SVM model. This suggests its potential utility as a valuable tool for early diagnosis of AD and related research.

We constructed an SVM model using METTL3 and SDHA as features and conducted 10-fold cross-validation to assess the model's performance. Each cross-validation iteration was accompanied by the plotting of ROC curves, which facilitated the visualization of the model's performance in AD diagnosis, as depicted in Figure 5. Furthermore, we conducted multiple experiments where we used METTL3 and, separately, NDUFA10, SDHC, and GPI as features to build distinct SVM models. The results of these experiments revealed that each model achieved average AUC values of 0.91, 0.86, and 0.92, respectively. These high average AUC values indicate that whether using METTL3 and SDHA or other genes as features, these SVM models exhibit exceptional performance in AD diagnosis.

4. Discussion

The primary objective of using DNA microarray technology in this study is to screen for candidate pathogenic genes associated with AD, providing a foundation for further biological validation [25–27]. Starting with genes that function similarly to known pathogenic genes can potentially save time in the search for causative genes [28–30]. Currently, numerous domestic and international institutions and researchers have conducted analyses on gene expression data, using machine learning methods such as the K-nearest neighbor method [28], clustering [29], neural networks [30], support vector machines (SVMs) [31], and more. Among these methods, clustering analysis is the most widely used statistical technique for gene expression data. For example, Brown et al. [32] used neural network-based clustering methods to study the evolutionary process of yeast cells. Alizadeh et al. [33] conducted research on diffuse large B-cell lymphoma using sample hierarchical clustering. Alon et al. [34] applied segmentation-based clustering algorithms to cluster genes and samples from 40 tumor and 22 normal colon tissues, comprising 6,500 genes. Sugiyama and Kotani [35] combined self-organizing maps and the k-means method, obtaining results superior to the sole use of the k-means method in clustering boundary delineation. Li et al. [28] employed GAs in conjunction with the K-nearest neighbor (KNN) method to extract features, selecting the most informative gene subset. Ryu and Cho [36] introduced the concept of an integrated classifier, combining multiple perceptrons, KNN methods, SVM, and other methods to construct classifiers. Futschik and Kasabov [37] and others applied fuzzy C-means clustering to experiment with yeast gene expression datasets in the presence of noise, demonstrating strong robustness.

While clustering analysis has many inherent advantages, it does not address the issue of cluster result ordering. Therefore, using global optimization algorithms to determine the optimal sequence of gene data to some extent surpasses clustering, as global

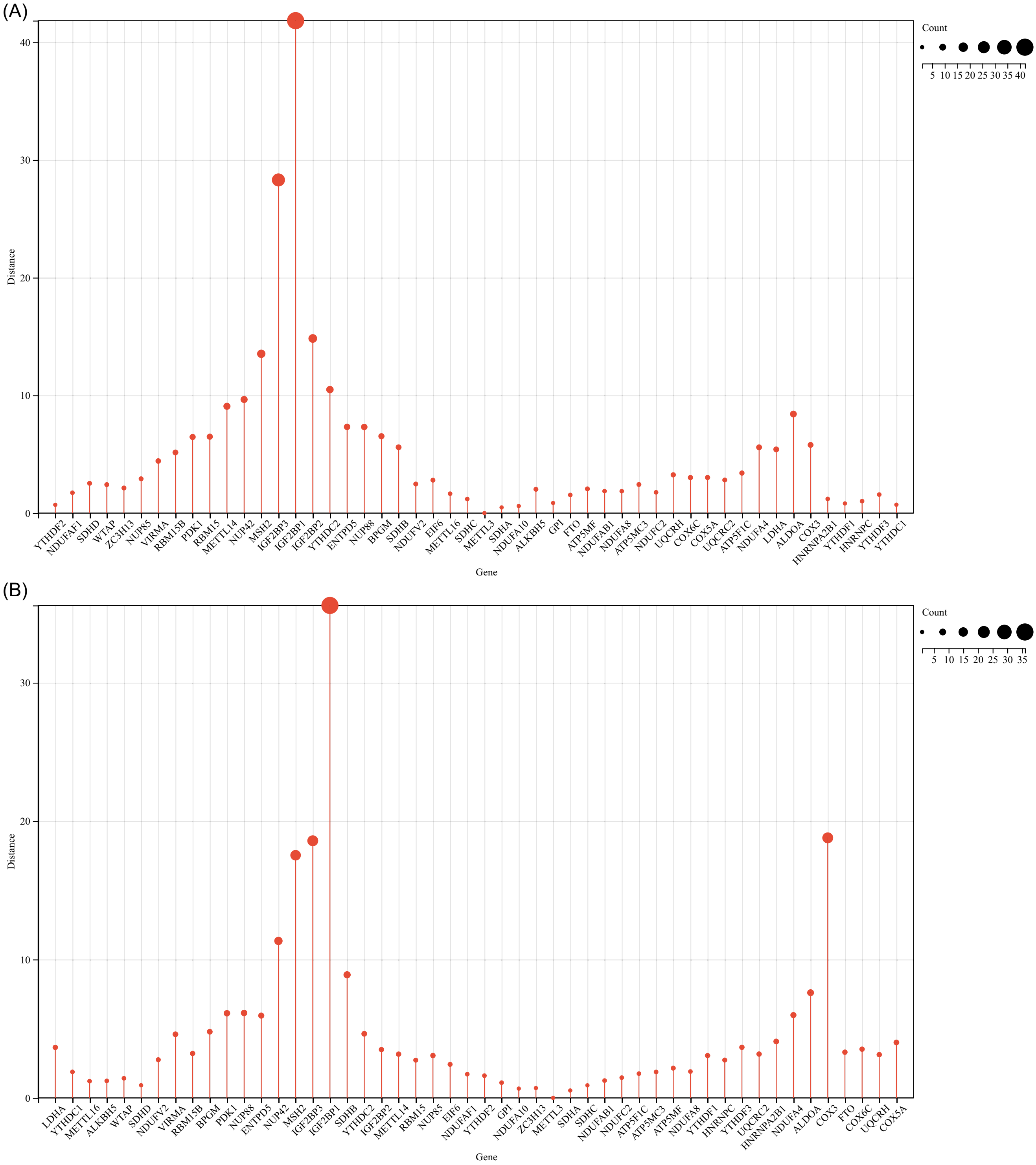

Figure 3. To provide a detailed representation of the co-expression patterns between other genes and METTL3, (A) displays the distance curves during the controllable phase of AD, where the x-axis represents gene names, and the y-axis represents the distance from the METTL3 gene. (B) illustrates the distance curves during the diseased phase of AD, with the x-axis denoting gene names and the y-axis indicating the distance from the METTL3 gene.

optimization algorithms the overall arrangement of genes, making it easier to analyze and study gene data. The current global optimization algorithms used to find the optimal gene sequence primarily include GAs and ant colony algorithms. Tsai et al. [38] transformed this problem into the TSP and applied an improved GA (FCGA) to solve the TSP problem. Lee et al. [39] applied a hybrid GA to solve this problem, while Chao-Yang Pang and others used the basic ant colony algorithm and an improved ant colony algorithm to solve the TSP problem.

Table 2. List of genes near METTL3 in 10 gene sequences during the controllable phase of AD

Gene	Frequency
SDHA	9
NDUFA10	8
SDHC	5
GPI	5

Table 3. List of genes near METTL3 in 10 gene sequences during the phase of AD

Gene	Frequency
SDHA	8
SDHC	8
YTHDF2	7
NDUFAF1	6
NDUFAB1	6
GPI	5
NDUFA10	5
ZC3H13	5

In this study, we identified genes co-expressed with METTL3 as novel features associated with AD. Utilizing DNA microarray data as the foundation for gene expression analysis, we represented each gene by a vector comprising expression levels under various experimental conditions, with each component corresponding to a specific condition. This vector encapsulates the intrinsic information of the gene. Genes with vectors in proximity are likely to share functional similarities and exhibit co-expression patterns. In our approach, each gene is conceptualized as a virtual city, with its position determined by the corresponding vector. By applying the ACO algorithm, we derived the optimal gene sequence—a linear arrangement where similar genes are ordered together. From this sequence, we defined a function value for each gene, representing its distance from METTL3 and quantifying the degree of co-expression. These function values collectively form a curve that maps the co-expression levels of all genes, providing a comprehensive overview of the relationships between all genes and METTL3.The function curve provides a precise and insightful depiction of gene co-expression levels, enabling a clearer observation of changes in co-expression patterns between the control and disease stages. This study identifies a key expression feature: The high co-expression levels

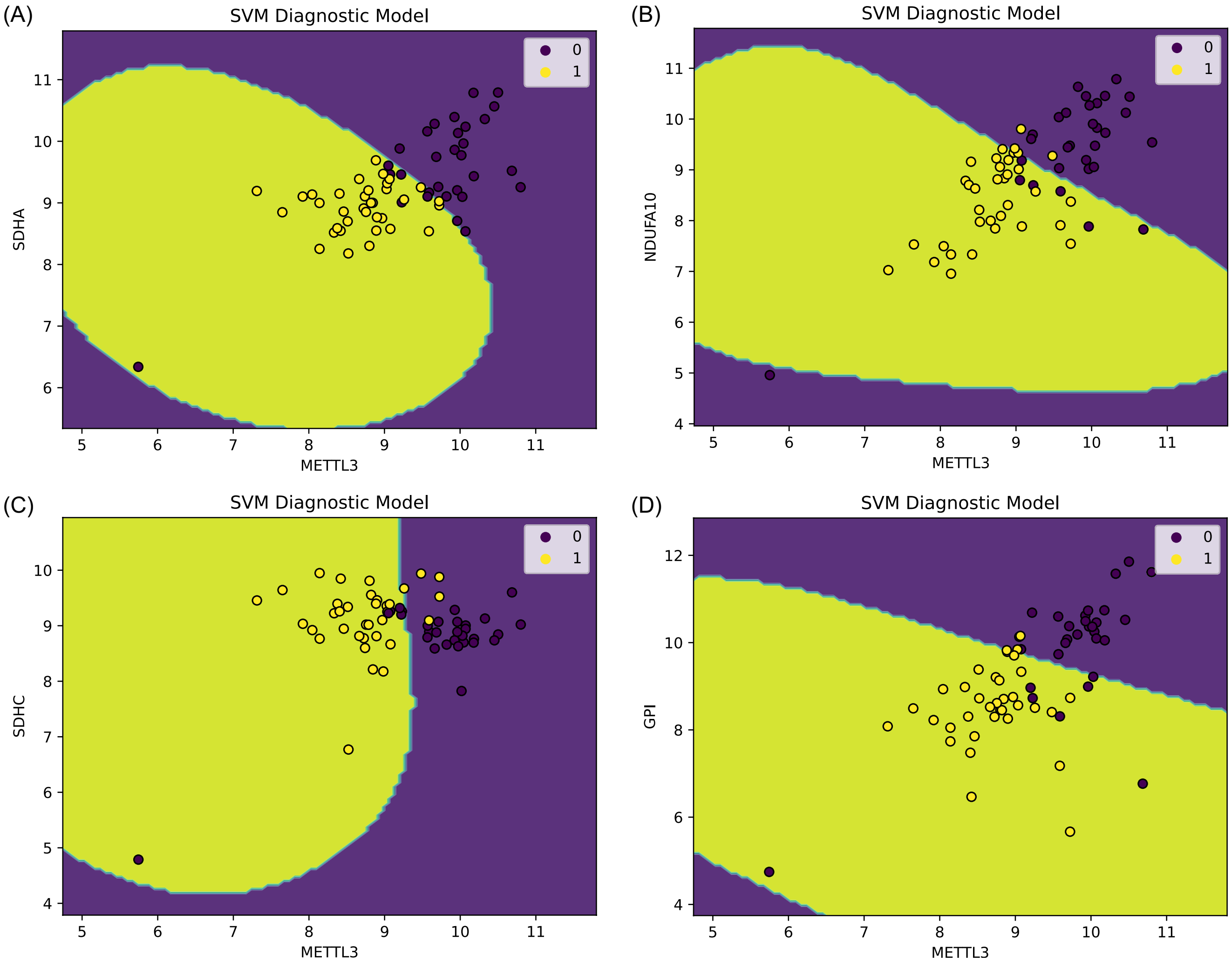

Figure 4. In the figure, AD patients are represented in yellow, while the control group is represented in purple. (A–D) Constructing diagnostic models of METTL3 and SDHA (NDUFA10, SDHC, and GPI) in AD. The horizontal and vertical coordinates represent the expression levels of METTL3 and SDHA (NDUFA10, SDHC, and GPI), respectively

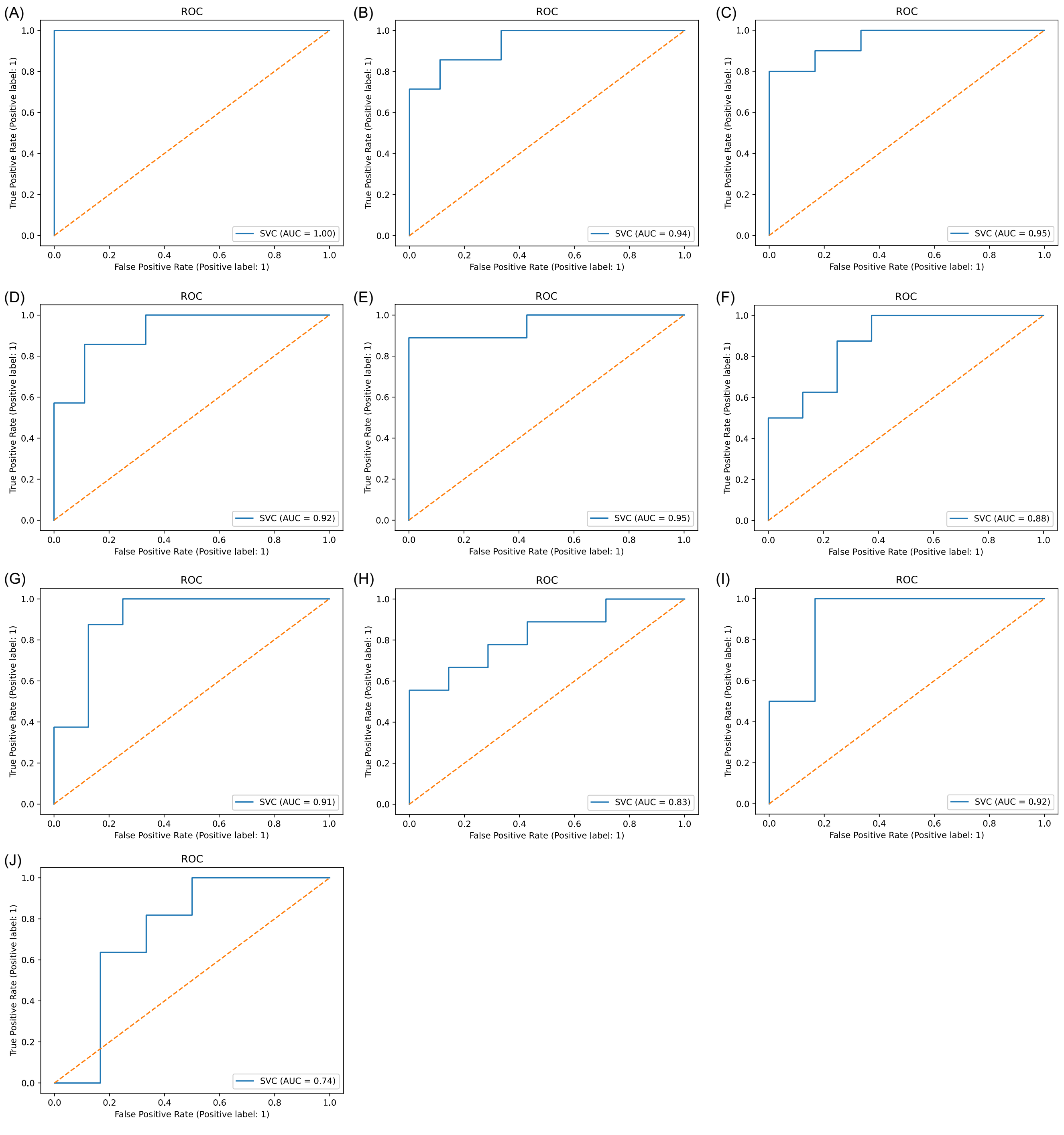

Figure 5. Validation of the model classification performance. (A–J) ROC curves. AUC represents the area under the ROC curve. When 0.5 < AUC < 1, the model demonstrates excellent classification performance and predictive value. The training and test sets were randomly divided ten times using cross-validation. The predictive performance of the SVM diagnostic model in AD was assessed by calculating the average of ten AUC values.

observed in both control and disease stages suggest that co-expressed genes, along with key marker genes, play a critical role in the progression of AD.

However, a significant limitation of this method is the extended computational time required when handling large-scale gene data. This challenge necessitated focusing on a subset of genes potentially associated with METTL3 in this study, which could result in incomplete conclusions and potential limitations of the findings. To address this issue, our research team is actively developing an efficient algorithm designed to handle datasets comprising tens of thousands of genes. We anticipate that this rapid algorithm will significantly reduce computation time and enable a comprehensive and in-depth analysis of gene expression patterns in AD in future studies. By overcoming the current computational constraints, we aim to uncover a broader spectrum of genetic information, thereby advancing our understanding of the complex pathological mechanisms underlying AD.

5. Conclusions

In this study, we leveraged an innovative gene analysis method grounded in the ACO algorithm to unravel the intricate etiological factors contributing to AD. By rigorously comparing gene co-expression curves and sequences between healthy and diseased states, we systematically identified a set of pivotal genes, namely METTL3, SDHA, NDUFA10, SDHC, and GPI. These genes were distinguished by their significant co-expression patterns, which strongly imply a potential collective involvement in the pathogenesis of AD.

The identification of these key genes was further substantiated through the deployment of a SVM model, which incorporated the identified genes to evaluate their diagnostic utility. The SVM model demonstrated robust performance in differentiating between AD patients and healthy controls, underscoring the potential of these genes as biomarkers for AD. This approach not only illuminates new avenues for the early diagnosis of AD but also provides a foundation for subsequent research into the molecular underpinnings of the disease.

Despite the promising implications of our findings, we recognize that the specific biological roles of these key genes in AD remain to be fully elucidated. The complex interactions and pathways through which these genes influence the progression of AD require comprehensive investigation through targeted molecular experiments. Future research should focus on disentangling the precise mechanisms by which these genes contribute to AD pathogenesis, thereby advancing our understanding and opening new doors for therapeutic development.

Our study, therefore, offers a novel perspective on the molecular mechanisms underlying AD, combining advanced computational methods with robust statistical validation. The integration of ACO-based gene analysis with machine learning approaches such as SVM not only highlights the power of interdisciplinary strategies in biomedical research but also provides a promising framework for the discovery of novel biomarkers and therapeutic targets in complex diseases like Alzheimer's.

Funding Support

This work was supported in part by the National Natural Science Foundation of China under Grant 61988102.

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

The data that support the findings of this study are openly available in Gene Expression Omnibus at https://www.ncbi.nlm.nih.gov/geo/.

Author Contribution Statement

Chengjiang Zhu: Conceptualization, Methodology, Software, Formal analysis, Data curation, Writing – original draft, Writing – review & editing. **Lin Yang:** Conceptualization, Methodology, Validation, Data curation, Writing – original draft. **Xudong Huang:** Supervision. **Kunpei Jin:** Investigation, Resources. **Xinping Pang:** Investigation. **Xianghu Song:** Validation. **Yue Sun:** Formal analysis, Investigation, Visualization. **Chonghao Gao:** Software. **Yanyu Wei:** Supervision, Project administration. **Chaoyang Pang:** Writing – review & editing, Supervision, Project administration, Funding acquisition.

References

[1] Knopman, D. S., Amieva, H., Petersen, R. C., Chételat, G., Holtzman, D. M., Hyman, B. T., . . . , & Jones, D. T. (2021). Alzheimer disease. *Nature Reviews Disease Primers*, *7*(1), 33. https://doi.org/10.1038/s41572-021-00269-y

[2] Khan, S., Barve, K. H., & Kumar, M. S. (2020). Recent advancements in pathogenesis, diagnostics and treatment of Alzheimer's disease. *Current Neuropharmacology*, *18*(11), 1106–1125. https://doi.org/10.2174/1570159X18666200528142429

[3] Lopez, J. A. S., González, H. M., & Léger, G. C. (2019). Alzheimer's disease. In S. T. Dekosky & S. Asthana (Eds.), *Handbook of clinical neurology* (vol. 167) (pp. 231–255). Elsevier. https://doi.org/10.1016/B978-0-12-804766-8.00013-3

[4] Porsteinsson, A. P., Isaacson, R. S., Knox, S., Sabbagh, M. N., & Rubino, I. (2021). Diagnosis of early Alzheimer's disease: Clinical practice in 2021. *The Journal of Prevention of Alzheimer's Disease*, *8*, 371–386. https://doi.org/10.14283/jpad.2021.23

[5] Lorenzi, M., Filippone, M., Frisoni, G. B., Alexander, D. C., & Ourselin, S. (2019). Probabilistic disease progression modeling to characterize diagnostic uncertainty: Application to staging and prediction in Alzheimer's disease. *NeuroImage*, *190*, 56–68. https://doi.org/10.1016/j.neuroimage.2017.08.059

[6] Sperling, R. A., Aisen, P. S., Beckett, L. A., Bennett, D. A., Craft, S., Fagan, A. M., . . . , & Phelps, C. H. (2011). Toward defining the preclinical stages of Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & Dementia*, *7*(3), 280–292. https://doi.org/10.1016/j.jalz.2011.03.003

[7] Alzheimer's Association. (2022). 2022 Alzheimer's disease facts and figures. *Alzheimer's & Dementia*, *18*(4), 700–789. https://doi.org/10.1002/alz.12638

[8] Leys, D., Hénon, H., Mackowiak-Cordoliani, M. A., & Pasquier, F. (2005). Poststroke dementia. *The Lancet Neurology*, *4*(11), 752–759. https://doi.org/10.1016/S1474-4422(05)70221-0

[9] Alzheimer's Association. (2019). 2019 Alzheimer's disease facts and figures. *Alzheimer's & Dementia*, *15*(3), 321–387. https://doi.org/10.1016/j.jalz.2019.01.010

[10] Ju, Y., & Tam, K. Y. (2022). Pathological mechanisms and therapeutic strategies for Alzheimer's disease. *Neural Regeneration Research*, *17*(3), 543–549. https://doi.org/10.4103/1673-5374.320970

[11] d'Errico, P., & Meyer-Luehmann, M. (2020). Mechanisms of pathogenic Tau and Aβ protein spreading in Alzheimer's disease. *Frontiers in Aging Neuroscience*, *12*, 265. https://doi.org/10.3389/fnagi.2020.00265

[12] Tiwari, S., Atluri, V., Kaushik, A., Yndart, A., & Nair, M. (2019). Alzheimer's disease: Pathogenesis, diagnostics, and therapeutics. *International Journal of Nanomedicine*, *14*, 5541–5554. https://doi.org/10.2147/IJN.S200490

[13] Ganguly, U., Kaur, U., Chakrabarti, S. S., Sharma, P., Agrawal, B. K., Saso, L., & Chakrabarti, S. (2021). Oxidative stress, neuroinflammation, and NADPH oxidase: Implications in the pathogenesis and treatment of Alzheimer's disease. *Oxidative Medicine and Cellular Longevity*, *2021*(1), 7086512. https://doi.org/10.1155/2021/7086512

[14] Henstridge, C. M., Hyman, B. T., & Spires-Jones, T. L. (2019). Beyond the neuron–cellular interactions early in Alzheimer disease pathogenesis. *Nature Reviews Neuroscience*, *20*(2), 94–108. https://doi.org/10.1038/s41583-018-0113-1

[15] Bai, R., Guo, J., Ye, X. Y., Xie, Y., & Xie, T. (2022). Oxidative stress: The core pathogenesis and mechanism of Alzheimer's disease. *Ageing Research Reviews*, *77*, 101619. https://doi.org/10.1016/j.arr.2022.101619

[16] Abdelwahab, M. M., Al-Karawi, K. A., & Semary, H. E. (2023). Deep learning-based prediction of Alzheimer's disease using microarray gene expression data. *Biomedicines*, *11*(12), 3304. https://doi.org/10.3390/biomedicines11123304

[17] Pasinetti, G. M. (2001). Use of cDNA microarray in the search for molecular markers involved in the onset of Alzheimer's disease dementia. *Journal of Neuroscience Research*, *65*(6), 471–476. https://doi.org/10.1002/jnr.1176

[18] Wang, G., Zhang, Y., Chen, B., & Cheng, J. (2003). Preliminary studies on Alzheimer's disease using cDNA microarrays. *Mechanisms of Ageing and Development*, *124*(1), 115–124. https://doi.org/10.1016/S0047-6374(02)00188-4

[19] Yu, H., Gao, L., Tu, K., & Guo, Z. (2005). Broadly predicting specific gene functions with expression similarity and taxonomy similarity. *Gene*, *352*, 75–81. https://doi.org/10.1016/j.gene.2005.03.033

[20] Xiong, J., Pang, X., Song, X., Yang, L., & Pang, C. (2024). The coherence between PSMC6 and α-ring in the 26S proteasome is associated with Alzheimer's disease. *Frontiers in Molecular Neuroscience*, *16*, 1330853. https://doi.org/10.3389/fnmol.2023.1330853

[21] Yang, L., Pang, X., Guo, W., Zhu, C., Yu, L., Song, X., . . . , & Pang, C. (2023). An exploration of the coherent effects between METTL3 and NDUFA10 on Alzheimer's disease. *International Journal of Molecular Sciences*, *24*(12), 10111. https://doi.org/10.3390/ijms241210111

[22] Guo, W., Gou, X., Yu, L., Zhang, Q., Yang, P., Pang, M., . . . , & Zhang, X. (2023). Exploring the interaction between T-cell antigen receptor-related genes and MAPT or ACHE using integrated bioinformatics analysis. *Frontiers in Neurology*, *14*, 1129470. https://doi.org/10.3389/fneur.2023.1129470

[23] Tanzi, R. E., & Bertram, L. (2005). Twenty years of the Alzheimer's disease amyloid hypothesis: A genetic perspective. *Cell*, *120*(4), 545–555. https://doi.org/10.1016/j.cell.2005.02.008

[24] Dahiya, C., & Sangwan, S. (2018). Literature review on travelling salesman problem. *International Journal of Research*, *5*(16), 1152–1155.

[25] Zhang, Q., Chen, B., Yang, P., Wu, J., Pang, X., & Pang, C. (2022). Bioinformatics-based study reveals that AP2M1 is regulated by the circRNA-miRNA-mRNA interaction network and affects Alzheimer's disease. *Frontiers in Genetics*, *13*, 1049786. https://doi.org/10.3389/fgene.2022.1049786

[26] Pan, J. S., Zhang, L. G., Wang, R. B., Snášel, V., & Chu, S. C. (2022). Gannet optimization algorithm: A new metaheuristic algorithm for solving engineering optimization problems. *Mathematics and Computers in Simulation*, *202*, 343–373. https://doi.org/10.1016/j.matcom.2022.06.007

[27] Song, P. C., Chu, S. C., Pan, J. S., & Yang, H. (2022). Simplified Phasmatodea population evolution algorithm for optimization. *Complex & Intelligent Systems*, *8*(4), 2749–2767. https://doi.org/10.1007/s40747-021-00402-0

[28] Li, L., Darden, T. A., Weingberg, C. R., Levine, A. J., & Pedersen, L. G. (2001). Gene assessment and sample classification for gene expression data using a genetic algorithm/k-nearest neighbor method. *Combinatorial Chemistry & High Throughput Screening*, *4*(8), 727–739. https://doi.org/10.2174/1386207013330733

[29] Tabus, I., & Astola, J. (2003). Clustering the non-uniformly sampled time series of gene expression data. In *Proceedings of the Seventh International Symposium on Signal Processing and Its Applications*, *2*, 61–64. https://doi.org/10.1109/ISSPA.2003.1224815

[30] Toure, A., & Basu, M. (2001). Application of neural network to gene expression data for cancer classification. In *Proceedings of the International Joint Conference on Neural Networks*, *1*, 583–587. https://doi.org/10.1109/IJCNN.2001.939087

[31] Brown, M. P., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C., Ares, M., & Haussler, D. (1999). *Support vector machine classification of microarray gene expression data*. Retrieved from: https://noble.gs.washington.edu/papers/brown_knowledge_tr.pdf

[32] Brown, M. P. S., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C. W., Furey, T. S., . . . , & Haussler, D. (2000). Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proceedings of the National Academy of Sciences*, *97*(1), 262–267. https://doi.org/10.1073/pnas.97.1.262

[33] Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., . . . , & Staudt, L. M. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, *403*(6769), 503–511. https://doi.org/10.1038/35000501

[34] Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., & Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences*, *96*(12), 6745–6750. https://doi.org/10.1073/pnas.96.12.6745

[35] Sugiyama, A., & Kotani, M. (2002). Analysis of gene expression data by using self-organizing maps and K-means clustering. In *Proceedings of the 2002 International Joint Conference on Neural Networks*, *2*, 1342–1345. https://doi.org/10.1109/IJCNN.2002.1007690

[36] Ryu, J., & Cho, S. B. (2002). Gene expression classification using optimal feature/classifier ensemble with negative correlation. In *Proceedings of the 2002 International Joint Conference on Neural Networks*, *1*, 198–203. https://doi.org/10.1109/IJCNN.2002.1005469

[37] Futschik, M. E., & Kasabov, N. K. (2002). Fuzzy clustering of gene expression data. In *Proceedings of the 2002 IEEE World Congress on Computational Intelligence and 2002 IEEE International Conference on Fuzzy Systems*, *1*, 414–419. https://doi.org/10.1109/FUZZ.2002.1005026.

[38] Tsai, H. K., Yang, J. M., & Kao, C. Y. (2002). Applying genetic algorithms to finding the optimal gene order in displaying the microarray data. In *Proceedings of the 4th Annual Conference on Genetic and Evolutionary Computation*, 610–617. https://dl.acm.org/doi/abs/10.5555/2955491.2955583#bibliography

[39] Lee, S. K., Kim, Y. H., & Moon, B. R. (2003). Finding the optimal gene order in displaying microarray data. In *Genetic and Evolutionary Computation Conference*, 2215–2226. https://doi.org/10.1007/3-540-45110-2_116

How to Cite: Zhu, C., Yang, L., Huang, X., Jin, K., Pang, X., Song, X., . . . , & Pang, C. (2024). ACO-Guided Genetic Analysis: Exploring the Role of Genes in Alzheimer's Disease. *Medinformatics*, *1*(4), 211–221. https://doi.org/10.47852/bonviewMEDIN42023658