frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 27 February 2023
DOI 10.3389/fgene.2023.1135260

OPEN ACCESS

EDITED BY
Suyan Tian,
Jilin University, China

REVIEWED BY
Guojun Liu,
Xi'an University of Finance and Economics, China
Changjing Zhuge,
Beijing University of Technology, China

*CORRESPONDENCE
Dan Li,
ldan@dlut.edu.cn
Jia Wang,
wangjia77@hotmail.com

SPECIALTY SECTION
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

RECEIVED 31 December 2022
ACCEPTED 16 February 2023
PUBLISHED 27 February 2023

CITATION
Wang Z, Gu H, Zhao M, Li D and Wang J (2023), MSC-CSMC: A multi-objective semi-supervised clustering algorithm based on constraints selection and multi-source constraints for gene expression data.
Front. Genet. 14:1135260.
doi: 10.3389/fgene.2023.1135260

MSC-CSMC: A multi-objective semi-supervised clustering algorithm based on constraints selection and multi-source constraints for gene expression data

Zeyuan Wang[1], Hong Gu[1], Minghui Zhao[1], Dan Li[1]* and Jia Wang[2]*

[1]Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology, Dalian, Liaoning, China, [2]Department of Breast Surgery, Second Hospital of Dalian Medical University, Dalian, Liaoning, China

Many clustering techniques have been proposed to group genes based on gene expression data. Among these methods, semi-supervised clustering techniques aim to improve clustering performance by incorporating supervisory information in the form of pairwise constraints. However, noisy constraints inevitably exist in the constraint set obtained on the practical unlabeled dataset, which degenerates the performance of semi-supervised clustering. Moreover, multiple information sources are not integrated into multi-source constraints to improve clustering quality. To this end, the research proposes a new multi-objective semi-supervised clustering algorithm based on constraints selection and multi-source constraints (MSC-CSMC) for unlabeled gene expression data. The proposed method first uses the gene expression data and the gene ontology (GO) that describes gene annotation information to form multi-source constraints. Then, the multi-source constraints are applied to the clustering by improving the constraint violation penalty weight in the semi-supervised clustering objective function. Furthermore, the constraints selection and cluster prototypes are put into the multi-objective evolutionary framework by adopting a mixed chromosome encoding strategy, which can select pairwise constraints suitable for clustering tasks through synergistic optimization to reduce the negative influence of noisy constraints. The proposed MSC-CSMC algorithm is testified using five benchmark gene expression datasets, and the results show that the proposed algorithm achieves superior performance.

KEYWORDS

semi-supervised clustering, constraint selection, multi-source constraints, gene expression data, multi-objective optimization

1 Introduction

The rapid development of microarray technology has generated a large amount of gene expression data and mining the inherent patterns in the massive gene expression data is a major challenge in the current bioinformatics field (Bandyopadhyay et al., 2007; Pirooznia et al., 2008). As an important unsupervised data mining method, clustering has become a powerful tool for gene expression data analysis. One of the main tasks of gene expression data clustering is to identify co-expressed genomes, which is a useful tool for further research on gene function (Bandyopadhyay et al., 2007; Chen et al., 2019). Compared with the unsupervised clustering methods, the semi-supervised clustering methods use prior information to guide the clustering process through data labels or pairwise constraints, which can effectively improve the performance of clustering (Wagstaff et al., 2001; Bilenko et al., 2004; Yin et al., 2010).

For semi-supervised clustering algorithms, the pairwise constraints are usually used to describe if two data belong to the same cluster. Specifically, the must-link constraint (ML) means that two data must be divided into the same cluster, and the cannot-link constraint (CL) means that two data must be divided into different clusters. The quality of the selected pairwise constraints is of vital importance, which significantly affects the performance of semi-supervised clustering algorithms (Grira et al., 2008; Vu et al., 2012; Masud et al., 2019; Abin and Vu, 2020). The pairwise constraints can be generated by directly using part of the known data labels (Lai et al., 2021) or by using an active learning method (Masud et al., 2019). In practical, most gene expression data are unlabeled, for which it is impossible to obtain pairwise constraints based on labels. Vu et al. (2012) indicated that the generation of the pairwise constraints should mainly focus on the data samples on the cluster boundaries, which are more likely to be misclassified. To this end, Basu et al. (2004) developed a farthest-first traversal scheme-based active learning method to obtain pairwise constraints. However, this method has been reported to be sensitive to noise (Davidson and Qi, 2008). Grira et al. (2008) proposed an active learning method to generate pairwise constraints by determining cluster boundary data using membership obtained by fuzzy clustering. Vu et al. (2012) identified data in sparse regions based on *k*-nearest neighbor graphs and constructed pairwise constraints. However, it was claimed that some pairwise constraints might not be generated by this method (Abin and Vu, 2020). Liu et al. (2018) proposed an entropy-based query strategy to select the most uncertain pairwise constraints. Abin (2018) proposed a random walk approach on the adjacency graph of data for querying informative constraints. Masud et al. (2019) used local density estimation to identify the most informative objects as pairwise constraints. Abin and Vu (2020) proposed a density tracking method which takes into account the density relationship between data, and uses the information about boundaries and skeleton of clusters to generate the pairwise constraints.

Although the above methods can automatically mine and learn the pairwise constraints of unlabeled datasets through different approaches, there are inevitably noisy constraints, i.e., constraints inconsistent with the ground-truth clusters, in the obtained pairwise constraints (Yin et al., 2010; Lai et al., 2021). However, the existing semi-supervised clustering algorithms are mostly based on the assumption that pairwise constraints conform to real cluster information, and usually susceptible to noisy constraints. Therefore, it is necessary to implement constraints selection, where noisy constraints are filtered out, and only pairwise constraints that are beneficial for semi-supervised clustering are retained. In addition, most of the pairwise-constraints-based semi-supervised clustering algorithms were developed for single-source constraints, i.e., the pairwise constraints are obtained only from the data itself. In real-world applications, many data also possess related domain information. For example, Gene Ontology (GO) (Ashburner et al., 2000), which describes gene products in terms of their associated biological processes, cellular components and molecular functions, can further provide gene annotation information for gene expression data. In this paper, the multi-source constraints are the pairwise constraints formed by the data itself and domain information. Apparently, compared with the single-source pairwise constraints based solely on gene expression data, the multi-source constraints formed by the fusion of gene ontology can provide more comprehensive information about the structure of gene clusters and help to guide semi-supervised clustering to obtain more accurate clustering results.

Aiming at the unlabeled gene expression data and from the perspective of reducing the negative impact of noisy constraints and integrating multi-source constraints, a method called multi-objective semi-supervised clustering algorithm based on constraints selection and multi-source constraints (MSC-CSMC) is proposed in this research. At first, the proposed algorithm uses gene expression data and GO information to generate multi-source pairwise constraints. Then, under the multi-objective optimization framework of Non-dominated Sorting Genetic Algorithm-II (NSGA-II), the constraints selection and the cluster prototypes are collaboratively optimized to realize the selection of pairwise constraints suitable for clustering with respect to the multi-source constraints and to improve the accuracy of semi-supervised clustering of gene expression data by reducing the negative impact of noisy constraints.

2 Methods

In this section, the details of our proposed MSC-CSMC algorithm are described. Our proposed method consists of two parts. Firstly, multi-source pairwise constraints are generated by integrating gene expression and gene ontology (GO) information. Then, by using the improved penalty weights as well as mixed chromosome encoding strategy of cluster prototype and constraints selection, multi-objective semi-supervised clustering based on constraints selection and multi-source constraints is performed to identify co-expressed gene groups. The workflow of MSC-CSMC is shown in Figure 1.

2.1 Generation of multi-source pairwise constraints

Gene expression data and gene ontology (GO) describe gene-related information from the abundance of mRNA of genes and gene annotation. Compared with the method only using gene expression data, the combination of these two aspects of information can help

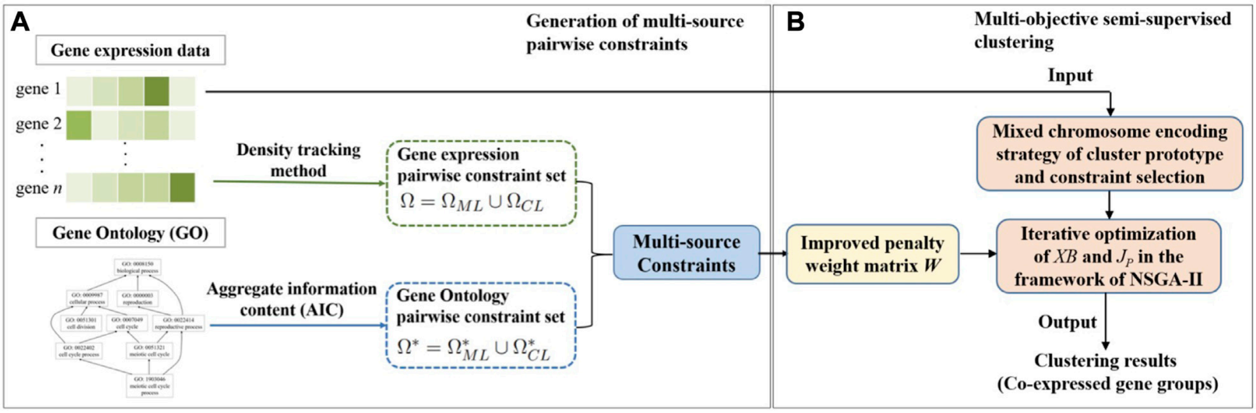
A
Generation of multi-source pairwise constraints
Gene expression data
gene 1
gene 2
gene n
Density tracking method
Gene expression pairwise constraint set
$\Omega = \Omega_{ML} \cup \Omega_{CL}$
Gene Ontology (GO)
Aggregate information content (AIC)
Gene Ontology pairwise constraint set
$\Omega^* = \Omega^*_{ML} \cup \Omega^*_{CL}$
Multi-source Constraints
B
Multi-objective semi-supervised clustering
Input
Mixed chromosome encoding strategy of cluster prototype and constraint selection
Improved penalty weight matrix W
Iterative optimization of XB and J_P in the framework of NSGA-II
Output
Clustering results (Co-expressed gene groups)

FIGURE 1
Workflow of MSC-CSMC. **(A)** Generation of multi-source pairwise constraints. **(B)** Multi-objective semi-supervised clustering.

to further improve the clustering accuracy of gene expression data (Giri and Saha, 2020; Li et al., 2022). In this paper, we use gene expression data and gene ontology information to generate multi-source pairwise constraints for semi-supervised clustering.

In view of the superior performance of the density tracking method (Abin and Vu, 2020), we use this method to generate the initial gene expression constraint set. The method consists of three steps: density estimation, density following, and constraints generation. Let $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots \boldsymbol{x}_n\}$, $\boldsymbol{x}_i \in \mathbb{R}^d$ denote a d-dimensional gene expression dataset with n genes. Gene $\boldsymbol{x}_i$'s density is obtained by

$$Density(\boldsymbol{x}_i) = \frac{1}{\max\limits_{\boldsymbol{x}_j \in N_b(\boldsymbol{x}_i)} \left\| \boldsymbol{x}_i - \boldsymbol{x}_j \right\|_2}, \tag{1}$$

where $N_b(\boldsymbol{x}_i)$ is the set of b nearest genes of gene $\boldsymbol{x}_i$; $\|\cdot\|_2$ is the Euclidean distance. Based on the density in Formula 1, the density tracking method constructs density chains according to the density relationship between data. Specifically, starting from each gene $\boldsymbol{x}_i$, the closest gene $\boldsymbol{x}_j \in N_b(\boldsymbol{x}_i)$ whose density is greater than that of $\boldsymbol{x}_i$ is selected, and the relation between them is recorded as density chain $\boldsymbol{x}_i \rightarrow \boldsymbol{x}_j$. Then start from gene $\boldsymbol{x}_j$ and continue the above density tracking until there exists no gene whose density is greater than that of the gene at the end of the chain. Consequently, the density chain $Chains(\boldsymbol{x}_i)$ can be denoted as $\boldsymbol{x}_i \rightarrow \boldsymbol{x}_j \rightarrow \cdots \rightarrow \boldsymbol{x}_e$. After constructing all the density chains, the total times of gene $\boldsymbol{x}_i$ appearing in all the chains is referred to as centrality and denoted by $Centrality(\boldsymbol{x}_i)$. The sum of centrality with respect to all genes in a density chain is used as the centrality of the density chain. All density chains with a common endpoint are considered connected density chains and the points belonging to them are considered to be in the same density group. Besides, the impurity of gene $\boldsymbol{x}_i$ is defined as follows:

$$Impurity(\boldsymbol{x}_i) = \left[1 - \sum_{g=1}^{|Groups|} \left(\frac{\sum_{\boldsymbol{x}_j \in S} \mathbb{I}\left(Group\left(\boldsymbol{x}_j\right) = g\right)}{b+1} \right)^2 \right] \times \left[1 - \frac{Density(\boldsymbol{x}_i)}{Density(\boldsymbol{x}_e)}\right] \tag{2}$$

with $|Groups|$ being the total number of groups, $S = \{\boldsymbol{x}_i \cup N_b(\boldsymbol{x}_i)\}$, $Group(\boldsymbol{x}_j)$ being the group index of $\boldsymbol{x}_j$, $\mathbb{I}$ being the indictor function.

According to the density, impurity, density chain, and density group of the data, the density tracking method proposes three assumptions for mining informative pairwise constraints. Let Ω denote the pairwise constraint set, whose elements satisfy the following key assumptions: (1) providing feasible information about the boundary data of clusters; (2) providing feasible information about the boundary between various clusters; (3) providing feasible information about the skeleton of clusters. Among them, assumptions (1) and (3) are used to generate the must-link constraint set Ω_{ML}, assumption (2) is used to generate the cannot-link constraint set Ω_{CL}. With the subsets Ω_{ML} and Ω_{CL}, the penalization can be constructed for the cost function of the clustering. The workflow of density tracking method is given in Figure 2. The initial gene expression constraint set $\Omega = \Omega_{ML} \cup \Omega_{CL}$ is generated as follows.

1. For each gene $\boldsymbol{x}_i$, calculate its $Density(\boldsymbol{x}_i)$ and $Impurity(\boldsymbol{x}_i)$. Construct density chain $Chains(\boldsymbol{x}_i)$ and density group $Group(\boldsymbol{x}_i)$, get the centrality of density chain. Initialize $\Omega_{ML} = \varnothing$, $\Omega_{CL} = \varnothing$;
2. Select gene $\boldsymbol{x}_i$ in descending order of $Impurity(\boldsymbol{x}_i)$, query the nearest neighbor gene $\boldsymbol{x}_j$ that is not in its density group $Group(\boldsymbol{x}_i)$, and add the pairwise constraint $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ into the cannot-link constraint set, i.e., $\Omega_{CL} = \Omega_{CL} \cup \{(\boldsymbol{x}_i, \boldsymbol{x}_j)\}$.
3. Select gene $\boldsymbol{x}_i$ in descending order of $Impurity(\boldsymbol{x}_i)$, and find the next gene $\boldsymbol{x}_j$ along its density chain $Chains(\boldsymbol{x}_i)$. Let $\varepsilon > 0$ denote the density drop rate. If $Density(\boldsymbol{x}_j) \geq \varepsilon \times Density(\boldsymbol{x}_e)$, then add the pairwise constraint $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ to the must-link constraint set, i.e., $\Omega_{ML} = \Omega_{ML} \cup \{(\boldsymbol{x}_i, \boldsymbol{x}_j)\}$;
4. Select the density chain $Chains(\boldsymbol{x}_i)$ in descending order of the centrality of the density chain, start from the starting gene $\boldsymbol{x}_i$, select the gene $\boldsymbol{x}_j$ with an interval, and add the pairwise constraint $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ to the must-link constraint set, i.e., $\Omega_{ML} = \Omega_{ML} \cup \{(\boldsymbol{x}_i, \boldsymbol{x}_j)\}$.

For a set of genes to be analyzed, each gene can be annotated with several GO terms. Thus, the functional similarity between genes can be deduced based on the term similarity. In the proposed MSC-CSMC algorithm, we adopt the aggregate information content (AIC)

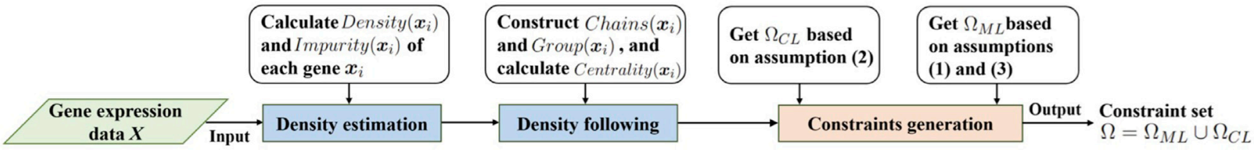
Calculate $Density(\boldsymbol{x}_i)$ and $Impurity(\boldsymbol{x}_i)$ of each gene $\boldsymbol{x}_i$
Construct $Chains(\boldsymbol{x}_i)$ and $Group(\boldsymbol{x}_i)$, and calculate $Centrality(\boldsymbol{x}_i)$
Get Ω_{CL} based on assumption (2)
Get Ω_{ML} based on assumptions (1) and (3)
Gene expression data X
Input
Density estimation
Density following
Constraints generation
Output
Constraint set $\Omega = \Omega_{ML} \cup \Omega_{CL}$

FIGURE 2
Workflow of density tracking method.

(Song et al., 2014) to measure the semantic similarity of GO terms t_1 and t_2:

$$sim_{AIC}(t_1, t_2) = \frac{\sum_{t \in T_{t_1} \cap T_{t_2}} 2 \times SW(t)}{SV(t_1) + SV(t_2)} \tag{3}$$

with

$$SW(t) = \frac{1}{1 + \exp(-1/IC(t))}, \quad SV(t) = \sum_{t' \in T_t} SW(t')$$

Here, T_t is the set of ancestors of term t in the GO graph, $p(t)$ is the frequency of the term appearing in the GO database, $IC(t) = -\log p(t)$ is the information content of term t. The higher the annotation frequency, the more general the information contained and the smaller the corresponding IC value. $SW(t)$ normalizes the knowledge reflected by $1/IC(t)$, describing the semantic weight of term t. Consequently, the functional similarity of genes $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ can be obtained as follows:

$$sim_{GO}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \frac{\sum_{t_2 \in ann(\boldsymbol{x}_j)} sim(\boldsymbol{x}_i, t_2) + \sum_{t_1 \in ann(\boldsymbol{x}_i)} sim(\boldsymbol{x}_j, t_1)}{|ann(\boldsymbol{x}_i)| + |ann(\boldsymbol{x}_j)|} \tag{4}$$

where

$$sim(\boldsymbol{x}_i, t_2) = \max_{t_1 \in ann(\boldsymbol{x}_i)} sim_{AIC}(t_1, t_2)$$

is the similarity of gene $\boldsymbol{x}_i$ and term t_2. $ann(\boldsymbol{x}_i)$ and $ann(\boldsymbol{x}_j)$ represent the sets of GO terms that annotate the two genes, respectively. The cardinalities of $ann(\boldsymbol{x}_i)$ and $ann(\boldsymbol{x}_j)$ are denoted by $|ann(\boldsymbol{x}_i)|$ and $|ann(\boldsymbol{x}_j)|$, respectively.

The gene function similarity obtained through GO can also reflect the pairwise constraint relationship between genes to a certain extent. In the proposed MSC-CSMC algorithm, gene pairs with a similarity of more than 0.9 constitute the GO must-link constraint set $\Omega^{\star}_{ML}$, gene pairs with a similarity less than 0.1 constitute the GO cannot-link constraint set $\Omega^{\star}_{CL}$, and then generate the GO pairwise constraint set $\Omega^{\star} = \Omega^{\star}_{ML} \cup \Omega^{\star}_{CL}$. Finally, the gene expression pairwise constraint set Ω and the gene ontology pairwise constraint set $\Omega^{\star}$ together constitute multi-source constraints for gene clustering.

2.2 Semi-supervised clustering objective functions based on multi-source constraints

At present, multi-objective optimization has gradually become a mainstream method for solving gene expression data clustering problems, which can achieve better clustering results on gene expression data compared with single-objective optimization methods. In the unsupervised multi-objective clustering problem of gene expression data, the cluster validity indices J_{FCM} (Bezdek et al., 1981) and XB (Xie and Beni, 1991), which measure the intra-cluster compactness and inter-cluster separation respectively, are commonly used as objective functions to realize the evolution of decision variables based on two conflicting objectives (Bandyopadhyay et al., 2007; Maulik et al., 2009; Mukhopadhyay et al., 2013; Li et al., 2022). In this paper, the proposed MSC-CSMC algorithm uses XB and the function based on quadratic-regularized fuzzy c-means with constraint violation penalty, namely, J_P (Mei, 2019), as the objective functions. Furthermore, the constraint violation penalty weights in J_P are improved to achieve semi-supervised clustering of gene expression data based on the multi-source constraints in the NSGA-II framework. The objective functions of XB and J_P are as follows:

$$XB = \frac{\sum_{c=1}^{k} \sum_{i=1}^{n} u_{ic}^2 \|\boldsymbol{x}_i - \boldsymbol{v}_c\|_2^2}{n \times \min_{f \neq c} \|\boldsymbol{v}_f - \boldsymbol{v}_c\|_2^2} \tag{5}$$

$$J_P = \sum_{c=1}^{k} \sum_{i=1}^{n} u_{ic} \|\boldsymbol{x}_i - \boldsymbol{v}_c\|_2^2 + \frac{\eta}{2} \sum_{c=1}^{k} \sum_{i=1}^{n} u_{ic}^2 - \frac{\beta}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \boldsymbol{u}_i^{\top} \boldsymbol{u}_j \tag{6}$$

Here,

$$\boldsymbol{v}_c = \frac{\sum_{i=1}^{n} u_{ic} \boldsymbol{x}_i}{\sum_{i=1}^{n} u_{ic}}$$

is the cth cluster prototype. k is the number of clusters, parameters η and β control the level of fuzziness and the contribution of the penalty term during clustering, respectively. u_{ic} is the membership degree of the datum $\boldsymbol{x}_i$ belonging to the cth cluster, obtained by

$$u_{ic} = \frac{1}{k} + \frac{1}{\eta}\left(u_{ic}^{FCM_q} + \beta u_{ic}^{P}\right) \tag{7}$$

$$u_{ic}^{FCM_q} = \frac{1}{k} \sum_{f=1}^{k} \|\boldsymbol{x}_i - \boldsymbol{v}_f\|_2^2 - \|\boldsymbol{x}_i - \boldsymbol{v}_c\|_2^2 \tag{8}$$

$$u_{ic}^{P} = \sum_{j=1}^{n} w_{ij} u_{jc} - \frac{1}{k} \sum_{f=1}^{k} \sum_{j=1}^{n} w_{ij} u_{jf} \tag{9}$$

where $w_{ij} \in \boldsymbol{W}$ is the penalty weight for violating pairwise constraint $(\boldsymbol{x}_i, \boldsymbol{x}_j)$. In order to simultaneously consider both the gene expression constraint set $\Omega = \Omega_{ML} \cup \Omega_{CL}$ and gene ontology constraint set $\Omega^{\star} = \Omega^{\star}_{ML} \cup \Omega^{\star}_{CL}$, that is, the multi-source constraints proposed in this paper, we improve the constraint violation penalty weights through the following analysis: (1) if

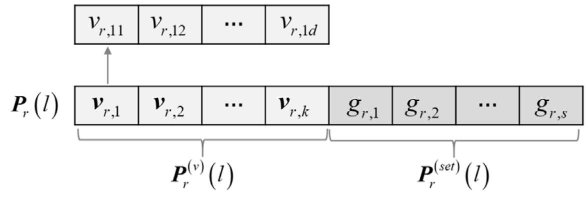

FIGURE 3
The mixed chromosome encoding strategy used in MSC-CSMC.

pairwise constraint $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ exists in both Ω_{ML} and $\Omega_{ML}^{\star}$, or in both Ω_{CL} and $\Omega_{CL}^{\star}$, it means that the same category information of gene pair $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ can be obtained from gene expression and gene annotation, so the weight of violating this constraint should be increased; (2) if pairwise constraint $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ exists in Ω_{ML} but not in $\Omega_{ML}^{\star}$, or exists in Ω_{CL} but not in $\Omega_{CL}^{\star}$, it indicates that the category information of gene pair $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is not clear enough, thus the penalty weight w_{ij} should be decreased; (3) if pairwise constraint $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ exists in both Ω_{ML} and $\Omega_{CL}^{\star}$, or in both Ω_{CL} and $\Omega_{ML}^{\star}$, it should be regarded as a contradictory constraint and removed from the constraint sets Ω and $\Omega^{\star}$. Based on the above idea, the MSC-CSMC algorithm proposed in this paper improves the constraint violation penalty weight as follows:

$$w_{ij} = \begin{cases} 1-\theta, & (\boldsymbol{x}_i, \boldsymbol{x}_j) \in \boldsymbol{\Omega}_{ML} \text{ and } (\boldsymbol{x}_i, \boldsymbol{x}_j) \notin \boldsymbol{\Omega}_{ML}^{\star} \\ -1+\theta, & (\boldsymbol{x}_i, \boldsymbol{x}_j) \in \boldsymbol{\Omega}_{CL} \text{ and } (\boldsymbol{x}_i, \boldsymbol{x}_j) \notin \boldsymbol{\Omega}_{CL}^{\star} \\ 1+\theta, & (\boldsymbol{x}_i, \boldsymbol{x}_j) \in \boldsymbol{\Omega}_{ML} \text{ and } (\boldsymbol{x}_i, \boldsymbol{x}_j) \in \boldsymbol{\Omega}_{ML}^{\star} \\ -1-\theta, & (\boldsymbol{x}_i, \boldsymbol{x}_j) \in \boldsymbol{\Omega}_{CL} \text{ and } (\boldsymbol{x}_i, \boldsymbol{x}_j) \in \boldsymbol{\Omega}_{CL}^{\star} \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

with $\theta > 0$ being the GO action parameter. It can be seen that the improved penalty weights can effectively integrate the gene expression and Gene Ontology information, and provide reasonable violation penalty for pairwise constraints in semi-supervised clustering.

2.3 Mixed chromosome encoding strategy used in MSC-CSMC

For the purpose of co-optimizing the constraints selection and clustering in the process of multi-objective evolution, a mixed encoding strategy combining the constraints selection and cluster prototype is adopted, as shown in Figure 3. Let $\boldsymbol{P}$ denote the genetic population, N be the population size, and s be the number of pairwise constraints to be selected. Considering the existence of noisy constraints in the initial pairwise constraint set and to improve the search efficiency of the algorithm, $2s$ constraints are randomly selected from the initial pairwise constraint set to generate the candidate constraint set Ω_p, and a serial number is assigned for each pairwise constraint. For a gene expression dataset with k clusters $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots \boldsymbol{x}_n\}$, $\boldsymbol{x}_i \in \mathbb{R}^d$, the rth individual in the lth generation $\boldsymbol{P}_r(l)$ consists of two parts: the cluster prototype $\boldsymbol{P}_r^{(v)}(l)$ and the constraints selection $\boldsymbol{P}_r^{(set)}(l)$. Among them, $\boldsymbol{P}_r^{(v)}(l) = [\boldsymbol{v}_{r,1}, \boldsymbol{v}_{r,2}, \ldots, \boldsymbol{v}_{r,k}]$ encode k cluster prototypes $\boldsymbol{v}_{r,c} = [v_{r,c1}, v_{r,c2}, \ldots, v_{r,cd}] (1 \le c \le k)$ with real numbers, $\boldsymbol{P}_r^{(set)}(l) = [g_{r,1}, g_{r,2}, \ldots, g_{r,s}]$ encode the serial numbers of s pairwise constraints $g_{r,j} (1 \le g_{r,j} \le 2s, 1 \le j \le s)$ selected from Ω_p with integers.

In the proposed algorithm, the two parts of the chromosomes are initialized separately. For the cluster prototype part, in order to ensure initialization quality and population diversity, half of the individuals are encoded as the k cluster prototypes obtained by the density peak method (Rodriguez and Laio, 2014), and the other half are encoded from the randomly generated cluster prototypes. For the constraints selection part of each individual, the components are initialized with non-repeated random integers in $[1, 2s]$.

2.4 Genetic operations

In the genetic evolution process of the MSC-CSMC algorithm, the roulette wheel strategy is first used to implement the selection. Since the NSGA-II algorithm tends to select individuals with lower non-domination ranks, for the rth individual $\boldsymbol{P}_r(l)$ of the lth generation, the selection probability (Zhou and Zhu, 2018) is calculated as follows:

$$p_s(\boldsymbol{P}_r(l)) = \alpha(1-\alpha)^{f_{rank}-1} \tag{11}$$

Here, $\alpha \in (0,1)$ is the selection parameter, f_{rank} is the non-domination rank of individual $\boldsymbol{P}_r(l)$.

For the parent individuals $\boldsymbol{P}_{r_1}(l)$ and $\boldsymbol{P}_{r_2}(l)$, let the crossover probability be p_c, different crossover operators are used for the cluster prototypes and constraints selection. Among them, $\boldsymbol{P}_{r_1}^{(v)}(l)$ and $\boldsymbol{P}_{r_2}^{(v)}(l)$ generate offspring through the normal distribution crossover operator (Zhang and Luo, 2009), and the offspring cluster prototypes are:

$$\boldsymbol{offsp}_1^{(v)} = \frac{\boldsymbol{P}_{r_1}^{(v)}(l) + \boldsymbol{P}_{r_2}^{(v)}(l)}{2} + 1.481 \times \frac{\boldsymbol{P}_{r_1}^{(v)}(l) - \boldsymbol{P}_{r_2}^{(v)}(l)}{2} \times |N(0,1)| \tag{12}$$

$$\boldsymbol{offsp}_2^{(v)} = \frac{\boldsymbol{P}_{r_1}^{(v)}(l) + \boldsymbol{P}_{r_2}^{(v)}(l)}{2} - 1.481 \times \frac{\boldsymbol{P}_{r_1}^{(v)}(l) - \boldsymbol{P}_{r_2}^{(v)}(l)}{2} \times |N(0,1)| \tag{13}$$

where $N(0,1)$ is a random variable of normal distribution. The constraints selection $\boldsymbol{P}_{r_1}^{(set)}(l)$ and $\boldsymbol{P}_{r_2}^{(set)}(l)$ adopts the single-point crossover operator, for a random integer $rand_c$ in $[1, s]$, the offspring constraints selections are:

$$\boldsymbol{offsp}_1^{(set)} = [g_{r_1,1}, \ldots, g_{r_1,rand_c}, g_{r_2,rand_c+1}, \ldots, g_{r_2,s}] \tag{14}$$

$$\boldsymbol{offsp}_2^{(set)} = [g_{r_2,1}, \ldots, g_{r_2,rand_c}, g_{r_1,rand_c+1}, \ldots, g_{r_1,s}] \tag{15}$$

If repeated pairwise constraints appear after crossover, non-repeated pairwise constraints are randomly selected from the candidate constraint set Ω_p as a replacement. For individual $\boldsymbol{P}_r(l)$, different mutation operators are adopted for the two parts. The polynomial mutation operator (Rousseeuw, 1987) is applied for $\boldsymbol{P}_r^{(v)}(l)$, where site $v_{r,ci}$ mutates with probability p_m:

$$v'_{r,ci} = v_{r,ci} + \delta \times (v_u - v_l), 1 \le c \le k, 1 \le i \le d \tag{16}$$

where, v_u and v_l are the upper and lower bounds of the cluster prototype, respectively. For normalized gene expression data, the bounds are set to 1 and 0. δ is determined as follows (Deb and Tiwari, 2008):

$$\delta = \begin{cases} \left(2 \times rand_m + (1 - 2 \times rand_m)(1 - v_{r,ci})^{\eta_m + 1}\right)^{\frac{1}{\eta_m + 1}} - 1, rand_m < 0.5 \\ 1 - \left[2 \times (1 - rand_m) + 2 \times (rand_m - 0.5) v_{r,ci}^{\eta_m + 1}\right]^{\frac{1}{\eta_m + 1}}, rand_m \geq 0.5 \end{cases} \tag{17}$$

Here, η_m is the distribution index, $rand_m$ is a random number in $[0, 1]$. For $\boldsymbol{P}_r^{(set)}(l)$, random mutation is used, that is, first randomly select a position in $\boldsymbol{P}_r^{(set)}(l)$, and then replace its value with a random integer in $[1, 2s]$ that is not repeated with others. In summary, the procedure of the MSC-CSMC algorithm is shown as follows:

Input: Gene expression dataset $\boldsymbol{X}$, number of neighbors b, density drop rate ε, population size N, maximal number of generations L_{max}, number of clusters k, fuzzy parameter η, penalty parameter β, constraint number s, GO action parameter θ, selection parameter α, crossover probability p_c, mutation probability p_m, and distribution index η_m.

Step 1: Generate gene expression pairwise constraint sets Ω based on density tracking method.

Step 2: Calculate the functional similarity of genes based on AIC, and generate the gene ontology pairwise constraint set Ω^*. Then delete the contradictory constraints, and determine the penalty weight matrix $\boldsymbol{W}$ corresponding to the multi-source constraints based on Formula 10.

Step 3: Randomly select $2s$ pairwise constraints from the initial constraint set to construct the candidate constraint set Ω_p, and initialize the population.

Step 4: When the genetic generation index is $l\,(l = 1, 2, \ldots, L_{max})$, for each individual $\boldsymbol{P}_r(l)$ $(1 \leq r \leq N)$, decode to obtain the cluster prototypes and the selected pairwise constraints. Update the membership degree according to Formulas 7-9, and calculate the individual fitness values based on Formulas 5-6.

Step 5: According to the individual fitness values, calculate the non-domination rank and crowding distance of each individual.

Step 6: Apply selection, crossover, and mutation based on Formulas 11-17, and update the individual fitness values according to Formulas 5-6.

Step 7: Merge the parent and offspring populations, and select the next-generation according to the elite retention strategy.

Step 8: If $l = \lfloor 0.\ 5 \times L_{max} \rfloor$ or $l = \lfloor 0.\ 8 \times L_{max} \rfloor$, update the penalty parameter $\beta = 2 \times \beta$ to increase the penalty for violating the currently selected constraints.

Step 9: Set $l = l + 1$, repeat Steps 4-8 until the maximal number of generations L_{max} is reached.

Output: The Pareto optimal solutions.

3 Results

3.1 Datasets

In this study, five benchmark gene expression datasets, namely, Yeast Galactose Metabolism, Yeast Cell Cycle, Yeast Sporulation, Serum, and Arabidopsis are used for the experiment.

The Yeast Galactose Metabolism dataset (Ideker et al., 2001) is composed of 205 genes whose expression patterns reflect four functional categories. The gene expression profiles were measured with four replicate assays across 20 time points. The Yeast Cell Cycle dataset (Cho et al., 1998) contains the expression levels of 384 genes involved in yeast cell cycle regulation at 17 time points, and these data are related with five phases of cell cycle. The Yeast sporulation dataset (Chu et al., 1998) contains the expression levels of more than 6,000 genes measured during the sporulation process of budding yeast across seven time points. The genes that showed no significant changes in expression during the harvesting were excluded, and the resulting set consists of 474 genes. The Serum dataset (Iyer et al., 1999) contains the expression levels of 517 human genes. The dataset has 13 dimensions corresponding to 12 time points and 1 unsynchronized sample. The Arabidopsis dataset (Reymond et al., 2000) consists of 138 *Arabidopsis Thaliana* genes. Each gene has eight expression values that correspond to eight time points. The details of the datasets are shown in Table 1.

3.2 Model evaluation criteria and parameter assignment

In order to evaluate the effectiveness of the model, the silhouette index (Rousseeuw, 1987) is chosen as the evaluation criterion for the clustering results. For gene $\boldsymbol{x}_i$, the silhouette width is calculated as follows:

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, 1 \leq i \leq n \tag{18}$$

Here, $a(i)$ is the average distance from gene $\boldsymbol{x}_i$ to other genes in the same cluster, $b(i)$ is the minimum average distance between gene $\boldsymbol{x}_i$ and genes in the other clusters. The silhouette index SI of dataset $\boldsymbol{X}$ is the mean value of the silhouette widths of all genes, with $SI \in [-1, 1]$. A greater SI value represents the algorithm with better clustering quality. Besides, as suggested by (Saha and Bandyopadhyay, 2013), the final solution of MSC-CSMS is selected from Pareto optimal solutions by using the silhouette index.

According to (Mei, 2019) and (Abin and Vu, 2020), the parameters of MSC-CSMC are assigned as follows: $\varepsilon = 0.8$, $b = 10$, $\eta = 0.001$, $\beta = 0.1$, $N = 100$, $L_{max} = 300$, $\alpha = 0.3$, $\eta_m = 5$, $p_c = 0.8$, $p_m = 0.1$. The number of pairwise constraints s is chosen as 0, 5, 10, 15, 20, and 25. In gene expression data analysis, the determination of the number of clusters k is an open problem. Generally, there are two approaches to determine the value of k; one is to directly set it as the true number of clusters (Yu et al., 2018; Zhao et al., 2021; Li et al., 2022; Liu et al., 2022; Wu and Ma, 2022); The other approach is applicable to the case where the true number of clusters is unknown, in which the variation range of k is determined firstly, and the k corresponding to the optimal value of an index (Silhouette index, Dunn index, Davies–Bouldin index, *etc.*) can be chosen as the optimal number of clusters (Gao et al., 2019; Acharya et al., 2020; López-Cortés et al., 2020; Zhang et al., 2022). In this paper, we adopt the first approach, and the number of clusters k is selected according to Table 1. In order to analyze the impact of the GO action parameter θ, we set θ from 0.1 to 0.9 at intervals of 0.1 under the condition that the number of the pairwise constraints is 15. The results are shown in Figure 4. It can be seen that the value of SI barely changes as θ increases, which means that the algorithm is not very

TABLE 1 Description of datasets.

Dataset	Number of genes	Number of features	Number of clusters
Yeast Galactose Metabolism	205	80	4
Yeast Cell Cycle	384	17	5
Yeast Sporulation	474	7	6
Serum	517	13	6
Arabidopsis	138	8	4

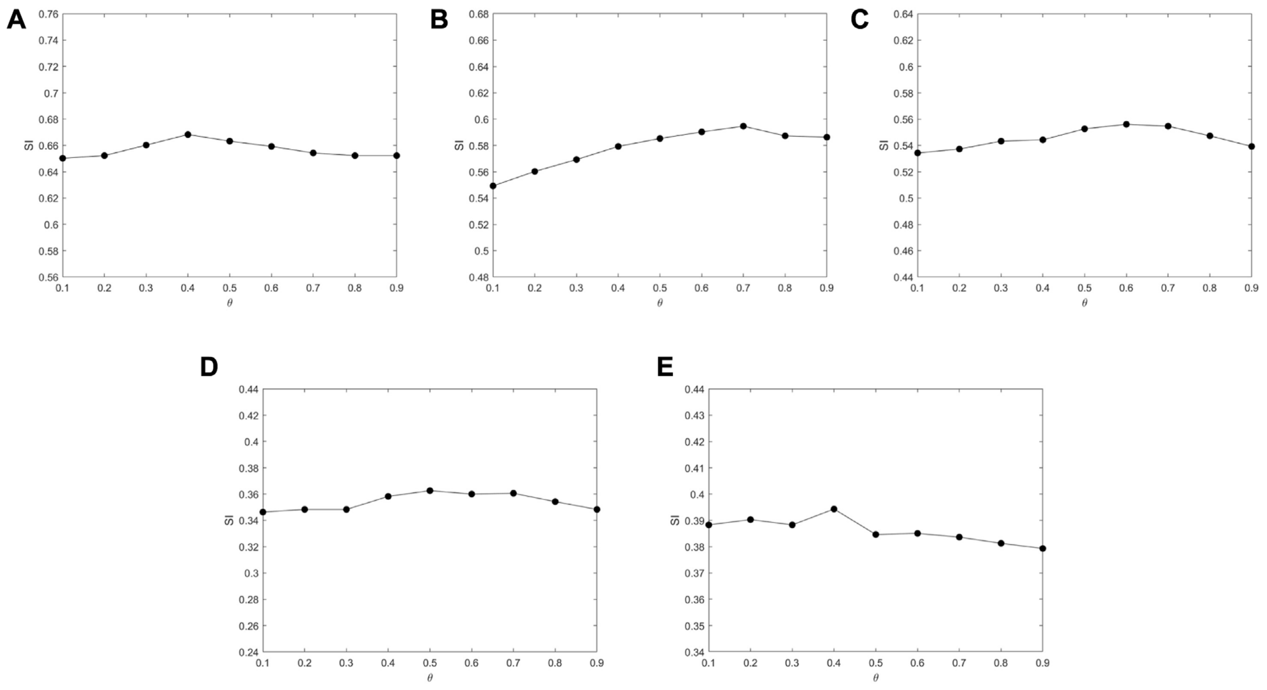

FIGURE 4
The impact of parameter *θ* on *SI* tested on different datasets. **(A)** Yeast Galactose Metabolism **(B)** Yeast Cell Cycle **(C)** Yeast Sporulation **(D)** Serum **(E)** Arabidopsis.

sensitive to the value of *θ*. For Yeast Galactose Metabolism, Yeast Cell Cycle, Yeast Sporulation, Serum, and Arabidopsis, the *θ* values are respectively set to 0.4, 0.7, 0.6, 0.5, and 0.4, which lead to the optimal clustering performances.

3.3 Result analysis and model comparison

For the purpose of inspecting the performance of the proposed MSC-CSMC algorithm, several advanced semi-supervised clustering algorithms based on single-source constraints, including COP-Kmeans (Wagstaff et al., 2001), PCKMeans (Basu et al., 2004), MPCKMeans (Bilenko et al., 2004), PCCA (Grira et al., 2008), PCFCMq (Mei, 2019) and MSC-CS (Zhao and Li, 2022), are used for comparison. Among them, the MSC-CS algorithm is the single-source constrained version of MSC-CSMC, which does not consider the annotation information provided by GO. In the above algorithms, the pairwise constraints are randomly selected from the initial gene expression constraint set Ω. To avoid the influence of randomness, each method is run for ten times under the same number of pairwise constraints, and the mean value of the clustering results is taken as the final result. The *SI* values of all seven algorithms applied to five datasets are shown in Tables 2–6, the optimal solutions in each row are highlighted in bold.

According to Tables 2–6, it can be seen that the proposed MSC-CSMS algorithm and its single-source constraint version MSC-CS can always achieve optimal and suboptimal clustering results on five gene expression datasets, demonstrating the effectiveness of the constraints selection. The mixed chromosome encoding strategy combining the constraint selection and cluster prototype can find the pairwise constraints suitable for clustering in the co-evolution process and improve clustering accuracy, and the highly accurate clustering

TABLE 2 *SI* **values on Yeast Galactose Metabolism with different number of constraints.**

s	COP-Kmeans	PCKMeans	MPCKMeans	PCCA	PCFCMq	MSC-CS	MSC-CSMC
0	0. 384	0. 254	0. 305	0.525	0. 465	**0. 566**	**0. 566**
5	0. 423	0. 479	0. 258	0.348	0. 254	0. 583	**0. 628**
10	0. 460	0. 484	0. 471	0.144	0. 274	0. 592	**0. 631**
15	0. 458	0. 484	0. 463	0.198	0. 402	0. 645	**0. 668**
20	0. 459	0. 457	0. 370	0.383	0. 351	0. 645	**0. 668**
25	0. 445	0. 433	0. 413	0.351	0. 290	0. 645	**0. 668**

The bold values indicate the optimal solutions in each row.

TABLE 3 *SI* **values on Yeast Cell Cycle with different number of constraints.**

s	COP-Kmeans	PCKMeans	MPCKMeans	PCCA	PCFCMq	MSC-CS	MSC-CSMC
0	0. 256	0. 252	0. 281	0.350	0. 408	**0. 436**	**0. 436**
5	0. 264	0. 250	0. 251	0.115	0. 385	0. 456	**0. 497**
10	0. 273	0. 227	0. 203	0.208	0. 409	0. 519	**0. 542**
15	0. 258	0. 275	0. 202	0.133	0. 408	0. 528	**0. 594**
20	0. 282	0. 263	0. 322	0.229	0. 408	0. 530	**0. 606**
25	0. 264	0. 261	0. 318	0.267	0. 409	0. 584	**0. 607**

The bold values indicate the optimal solutions in each row.

TABLE 4 *SI* **values on Yeast Sporulation with different number of constraints.**

s	COP-Kmeans	PCKMeans	MPCKMeans	PCCA	PCFCMq	MSC-CS	MSC-CSMC
0	0. 329	0. 328	0. 345	0.400	0.364	**0. 491**	**0. 491**
5	0. 331	0. 354	0. 411	0.067	0.463	0. 520	**0. 528**
10	0. 324	0. 429	0. 404	0.164	0.420	0. 525	**0. 531**
15	0. 300	0. 404	0. 409	0.325	0.434	**0. 565**	0. 556
20	0. 324	0. 403	0. 405	0.235	0.416	0. 571	**0. 592**
25	0. 346	0. 396	0. 394	0.286	0.413	0. 592	**0. 594**

The bold values indicate the optimal solutions in each row.

TABLE 5 *SI* **values on Serum with different number of constraints.**

s	COP-Kmeans	PCKMeans	MPCKMeans	PCCA	PCFCMq	MSC-CS	MSC-CSMC
0	0. 212	0. 208	0. 186	0.290	0.270	**0. 312**	**0. 312**
5	0. 210	0. 205	0. 211	0.080	0.264	**0. 327**	0. 325
10	0. 200	0. 202	0. 197	0.146	0.271	**0. 341**	0. 340
15	0. 200	0. 181	0. 184	0.235	0.264	0. 354	**0. 362**
20	0. 198	0. 206	0. 217	0.144	0.262	0. 368	**0. 385**
25	0. 193	0. 202	0. 185	0.238	0.269	0. 379	**0. 403**

The bold values indicate the optimal solutions in each row.

results can further improve the constraint selection ability of the algorithm in turn. Conversely, the algorithms for comparison are based on the assumption that the pairwise constraints conform to the real cluster information and are easily affected by noisy constraints. This is consistent with the analysis of the negative effects of noisy constraints by (Yin et al., 2010) and (Lai et al., 2021). In addition, the

TABLE 6 ***SI*** **values on Arabidopsis with different number of constraints.**

s	COP-Kmeans	PCKMeans	MPCKMeans	PCCA	PCFCMq	MSC-CS	MSC-CSMC
0	0. 220	0. 223	0. 197	0.314	0.353	**0. 358**	**0. 358**
5	0. 207	0. 216	0. 192	-0.151	0.353	0. 368	**0. 373**
10	0. 212	0. 210	0. 206	0.046	0.353	0. 373	**0. 387**
15	0. 200	0. 201	0. 185	0.106	0.354	0. 375	**0. 394**
20	0. 197	0. 189	0. 185	0.308	0.344	0. 381	**0. 396**
25	0. 187	0. 187	0. 181	0.335	0.352	0. 389	**0. 397**

The bold values indicate the optimal solutions in each row.

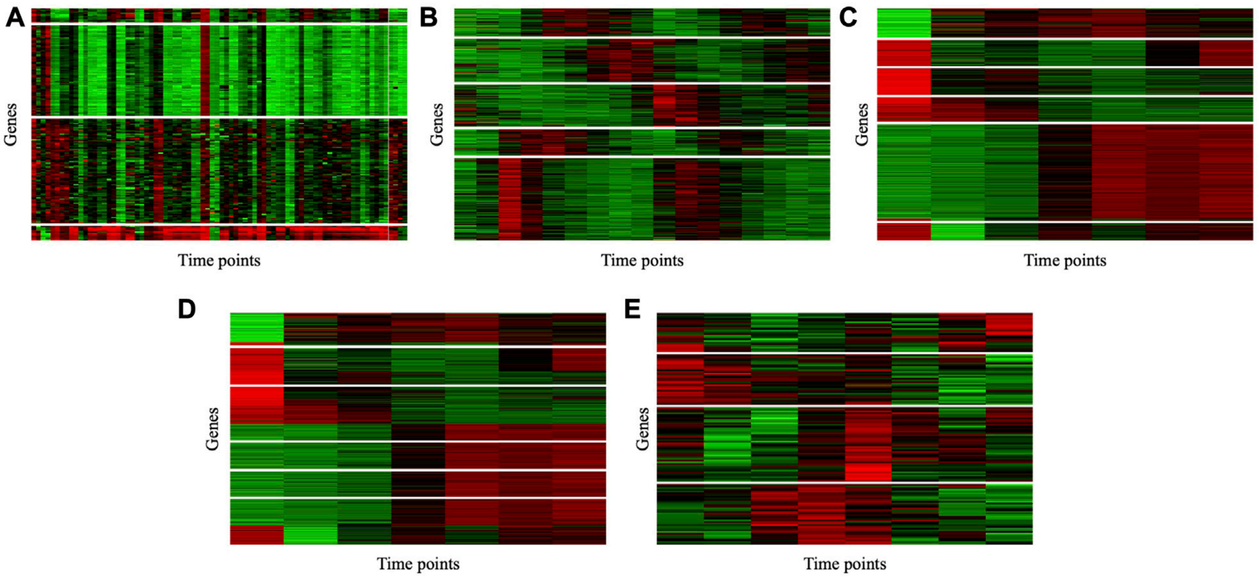

FIGURE 5
Eisen plots of the gene clusters obtianed by MSC-CSMC. **(A)** Yeast Galactose Metabolism with the number of constraints *s* =5 **(B)** Yeast Cell Cycle with the number of constraints *s* =10 **(C)** Yeast Sporulation with the number of constraints *s* =15 **(D)** Serum with the number of constraints *s* =20 **(E)** Arabidopsis with the number of constraints *s* =25.

MSC-CSMC algorithm is better than MSC-CS in most cases, indicating that using multi-source constraints can improve the performance of semi-supervised clustering. The gene ontology used to generate multi-source pairwise constraints in our MSC-CSMC algorithm can explain gene expression profiles from the perspective of gene function. By effectively integrating the gene expression and Gene Ontology information, the proposed penalty weights can provide reasonable violation penalty for pairwise constraints.

In the case of $s = 0$, that is, there is no pairwise constraint, both MSC-CSMC and MSC-CS degenerate into unsupervised multi-objective clustering methods, turning out the same result. Compared with PCFCMq, which uses J_P as the single objective function, the better performance of MSC-CSMC and MSC-CS shows the advantages of using multi-objective optimization in clustering gene expression data.

Among the comparison algorithms, the performance of the PCFCMq algorithm, which is based on fuzzy clustering, is generally better than the hard clustering-based COP-Kmeans, PCKMeans, and MPCKMeans algorithms. According to (Gasch and Eisen, 2002), genes may be co-expressed with different genomes under different measurement conditions, and there is usually overlap between gene clusters. Therefore, compared with hard clustering algorithms, fuzzy clustering algorithms are more suitable for analyzing gene expression data. Furthermore, due to the proposed constraints selection and multi-source constraint fusion strategy, the MSC-CSMC algorithm achieves better clustering results than the PCFCMq algorithm. In terms of the robustness of the clustering results, the performances of semi-supervised clustering algorithms for comparison fluctuate with the increase of pairwise constraints, which is mainly due to the quality of randomly selected pairwise constraints. As stated by Lai et al. (2021), even non-noisy constraints that conform to the real cluster information may have a negative impact on the clustering results, which further illustrates the necessity of constraints selection in semi-supervised clustering algorithms. The proposed MSC-CSMC algorithm can select pairwise constraints suitable for clustering based on the co-evolution of the cluster prototype and constraints selection, which guarantees both accuracy and stability of the clustering results.

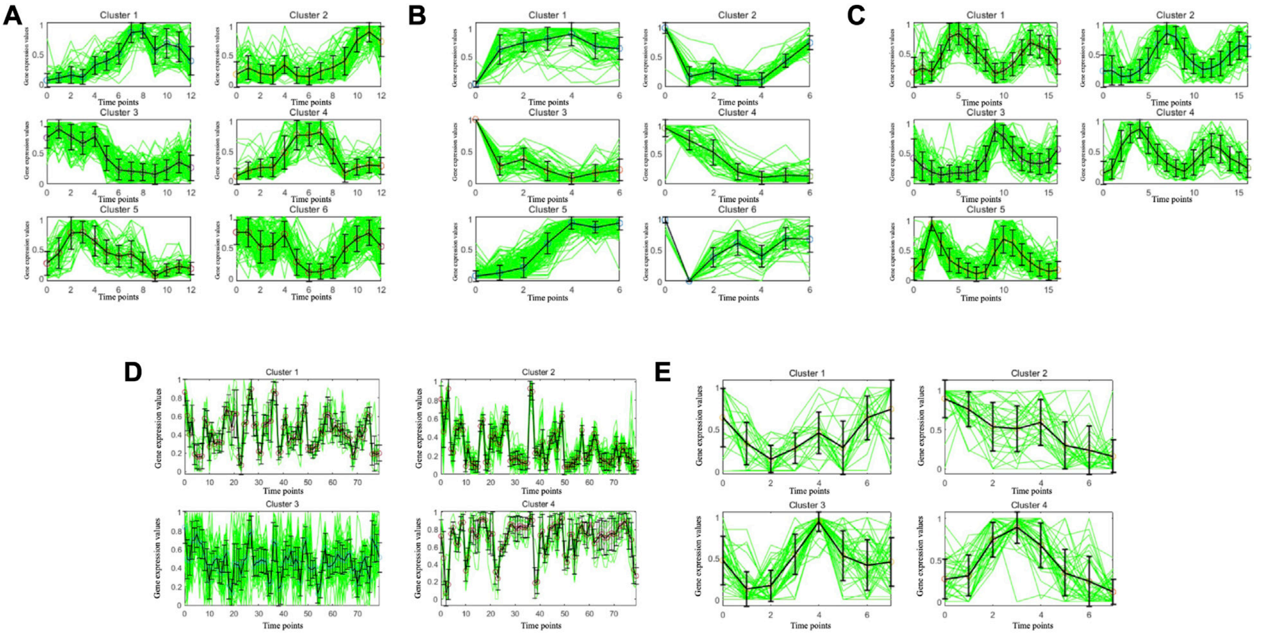

FIGURE 6
Cluster profile plots of the gene clusters obtianed by MSC-CSMC. **(A)** Yeast Galactose Metabolism with the number of constraints *s* =5 **(B)** Yeast Cell Cycle with the number of constraints *s* =10 **(C)** Yeast Sporulation with the number of constraints *s* =15 **(D)** Serum with the number of constraints *s* =20 **(E)** Arabidopsis with the number of constraints *s* =25.

TABLE 7 The three most significant GO terms and the corresponding *p*-values for each of the six clusters obtained by MSC-CSMC on Yeast Sporulation.

Gene cluster	GO term	*p*-value
1	meiotic cell cycle (GO:0051321)	1.42E-53
	meiotic cell cycle process (GO:1903046)	4.33E-51
	peptide biosynthetic process (GO:004304)	2.07E-48
2	sporulation (GO:0043934)	2.17E-45
	translation (GO:0006412)	4.08E-44
	sporulation resulting in formation of a cellular spore (GO:0030435)	1.02E-40
3	meiotic cell cycle (GO:0051321)	2.50E-30
	meiotic nuclear division (GO:0140013)	3.85E-28
	nuclear division (GO:0000280)	1.16E-26
4	cell cycle process (GO: 0022402)	7.37E-23
	cell cycle (GO: 0007049)	3.67E-22
	cell wall organization (GO: 0071555)	3.46E-22
5	cell development (GO: 0048468)	6.15E-20
	ascospore formation (GO: 0030437)	1.45E-19
	anatomical structure development (GO: 0048856)	3.53E-19
6	small molecule metabolic process (GO: 0044281)	2.51E-11
	amino-acid betaine metabolic process (GO: 0006577)	3.16E-09
	carnitine metabolic process (GO: 0009437)	3.15E-09

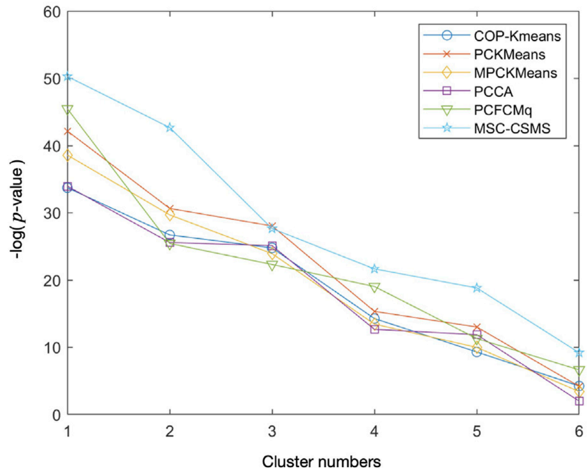

FIGURE 7
Average negative logarithmic p-values of the three most significant GO terms for each of the six clusters on Yeast Sporulation with the number of constraints s =15.

To illustrate the consistency of the gene clusters obtained by the MSC-CSMC algorithm, the Eisen plots and cluster profile plots corresponding to the clustering results of five datasets are shown in Figure 5 and Figure 6. In the Eisen plots, each row corresponds to a gene, each column to a time point (sample), and each entry of the plot represents the expression level of a gene at a specific time point by coloring the corresponding cell. To illustrate more clearly the gene clusters obtained by MSC-CSMC, the genes partitioned into the same cluster are placed together. In the cluster profile plots, the X- and Y-axis represent the time points and gene expression values, respectively. The expression values of genes partitioned into the same cluster are plotted in the same subplot. In the subplots, each green line indicates the normalized expression values of a gene over all time points, and the black line represents the mean expression level of the genes in the corresponding cluster. It can be seen in the Eisen plots that the color patterns (expression levels) of genes in the same cluster are similar to each other, while genes in different clusters show different color patterns. According to Figure 6, the cluster profiles of different clusters are different from each other, and the cluster profiles within a cluster reveal consistency.

In order to inspect the biological significance of the gene clusters obtained by the MSC-CSMC algorithm, enrichment analysis is carried out using the GO annotation database, which results in the significant GO terms shared by genes in each cluster and their corresponding p-values. Taking the case where the number of pairwise constraints in the Yeast Sporulation dataset is 15 as an example, we focus on the three most significant GO terms (corresponding to the three lowest p-values) in each of the six clusters obtained by each algorithm. Figure 7 shows the plot of the average p-values. To illustrate the difference significantly, the p-values are negative log-transformed and the clusters are sorted in descending order according to the transformed values. Table 7 reports the three most significant GO terms and the corresponding p-values in each cluster obtained by MSC-CSMC.

From Figure 7, it can be seen that the curve corresponding to MSC-CSMC is higher than those of the other algorithms, indicating that MSC-CSMC gains the result with the highest biological significance. Moreover, all the p-values of the significant GO terms listed in Table 7 are far less than 0.01, indicating that the MSC-CSMC algorithm can identify biologically relevant gene clusters.

4 Conclusion

Aiming at the problem that current semi-supervised clustering methods based on pairwise constraints are easily affected by noisy constraints and do not take the fusion of multi-source constraints into account, in this paper, we propose a multi-objective semi-supervised clustering algorithm based on constraints selection and multi-source constraints (MSC-CSMC). The proposed algorithm uses gene expression data and GO information to generate multi-source pairwise constraints and applies the multi-source constraints to the semi-supervised clustering process through improved constraint violation penalty weights. On this basis, a collaborative multi-objective optimization framework for constraints selection and cluster prototypes is constructed, and the negative impact of the noisy constraints is reduced by selecting pairwise constraints suitable for clustering. Experimental results on multiple gene expression datasets show that the MSC-CSMC algorithm effectively improves the performance of semi-supervised clustering. The validity of the proposed method proposed is not limited to the cluster analysis of gene expression data. Other semi-supervised clustering studies with multi-source information or constrained selection requirements can also be enlightened.

The effectiveness of the algorithm in this paper has been verified in small and medium-sized gene expression datasets. With the increase in the data size, the augment in the number of decision variables in the process of multi-objective evolution will lead to a decrease in algorithm efficiency and optimization performance. Therefore, the next step is to use decision variable analysis and other methods to design a multi-objective evolution strategy of the algorithm so as to further improve the applicability of the algorithm in practical clustering problems. In addition, we will also try to use various evaluation indices and design a multi-objective optimization framework with variable coding length (Rodríguez-Méndez et al., 2019) to optimize the number of clusters for gene expression data.

Data availability statement

The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding authors.

Author contributions

DL proposed the idea. ZW and MZ did the experiment. ZW, JW, and DL summarized the results and finished the manuscript. All authors proofread the manuscript.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Abin, A. A. (2018). A random walk approach to query informative constraints for clustering. *IEEE Trans. Cybern.* 48, 2272–2283. doi:10.1109/TCYB.2017.2731868

Abin, A. A., and Vu, V. (2020). A density-based approach for querying informative constraints for clustering. *Expert Syst. Appl.* 161, 113690. doi:10.1016/j.eswa.2020.113690

Acharya, S., Saha, S., and Pradhan, P. (2020). Multi-factored gene-gene proximity measures exploiting biological knowledge extracted from gene ontology: Application in gene clustering. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* 17, 207–219. doi:10.1109/TCBB.2018.2849362

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: Tool for the unification of biology. The gene ontology consortium. *Nat. Genet.* 25, 25–29. doi:10.1038/75556

Bandyopadhyay, S., Mukhopadhyay, A., and Maulik, U. (2007). An improved algorithm for clustering gene expression data. *Bioinformatics* 23, 2859–2865. doi:10.1093/bioinformatics/btm418

Basu, S., Banerjee, A., and Mooney, R. J. (2004). "Active semi-supervision for pairwise constrained clustering," in Proceedings of the 2004 SIAM International Conference on Data Mining (Philadelphia, Pennsylvania: SIAM), 333–344. doi:10.1137/1.9781611972740.31

Bezdek, J. C., Coray, C., Gunderson, R., and Watson, J. (1981). Detection and characterization of cluster substructure i. linear structure: Fuzzy c-lines. *SIAM J. Appl. Math.* 40, 339–357. doi:10.1137/0140029

Bilenko, M., Basu, S., and Mooney, R. J. (2004). "Integrating constraints and metric learning in semi-supervised clustering," in Proceedings of the Twenty-First International Conference on Machine Learning (New York, NY, USA: Association for Computing Machinery), 11. doi:10.1145/1015330.1015360

Chen, X., Huang, J. Z., Wu, Q., and Yang, M. (2019). Subspace weighting co-clustering of gene expression data. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* 16, 352–364. doi:10.1109/TCBB.2017.2705686

Cho, R. J., Campbell, M. J., Winzeler, E. A., Steinmetz, L., Conway, A., Wodicka, L., et al. (1998). A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell* 2, 65–73. doi:10.1016/s1097-2765(00)80114-8

Chu, S., DeRisi, J., Eisen, M., Mulholland, J., Botstein, D., Brown, P. O., et al. (1998). The transcriptional program of sporulation in budding yeast. *Science* 282, 699–705. doi:10.1126/science.282.5389.699

Davidson, I., and Qi, Z. (2008). "Finding alternative clusterings using constraints," in 2008 Eighth IEEE International Conference on Data Mining (Pisa, Italy: IEEE), 773–778. doi:10.1109/ICDM.2008.141

Deb, K., and Tiwari, S. (2008). Omni-optimizer: A generic evolutionary algorithm for single and multi-objective optimization. *Eur. J. Operational Res.* 185, 1062–1087. doi:10.1016/j.ejor.2006.06.042

Gao, Y., Zhou, X., and Zhang, W. (2019). An ensemble strategy to predict prognosis in ovarian cancer based on gene modules. *Front. Genet.* 10, 366. doi:10.3389/fgene.2019.00366

Gasch, A. P., and Eisen, M. B. (2002). Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering. *Genome Biol.* 3, RESEARCH0059–22. doi:10.1186/gb-2002-3-11-research0059

Giri, S. J., and Saha, S. (2020). "Multi-view gene clustering using gene ontology and expression-based similarities," in 2020 IEEE Congress on Evolutionary Computation (CEC) (Glasgow, UK: IEEE), 1–8. doi:10.1109/CEC48606.2020.9185885

Grira, N., Crucianu, M., and Boujemaa, N. (2008). Active semi-supervised fuzzy clustering. *Pattern Recognit.* 41, 1834–1844. doi:10.1016/j.patcog.2007.10.004

Ideker, T., Thorsson, V., Ranish, J. A., Christmas, R., Buhler, J., Eng, J. K., et al. (2001). Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* 292, 929–934. doi:10.1126/science.292.5518.929

Iyer, V. R., Eisen, M. B., Ross, D. T., Schuler, G., Moore, T., Lee, J. C., et al. (1999). The transcriptional program in the response of human fibroblasts to serum. *Science* 283, 83–87. doi:10.1126/science.283.5398.83

Lai, Y., He, S., Lin, Z., Yang, F., Zhou, Q., and Zhou, X. (2021). An adaptive robust semi-supervised clustering framework using weighted consensus of random *k* k-means ensemble. *IEEE Trans. Knowl. Data Eng.* 33, 1877–1890.

Li, D., Gu, H., Chang, Q., Wang, J., and Qin, P. (2022). A joint optimization framework integrated with biological knowledge for clustering incomplete gene expression data. *Soft Comput.* 2022, 1–18. doi:10.1007/s00500-022-07180-y

Liu, Y., Li, H., Xu, Y., Liu, Y., Peng, X., and Wang, J. (2022). Isocell: An approach to enhance single cell clustering by integrating isoform-level expression through orthogonal projection. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* 20, 1–475. doi:10.1109/TCBB.2022.3147193

Liu, Y., Liu, K., Zhang, C., Wang, X., Wang, S., and Xiao, Z. (2018). Entropy-based active sparse subspace clustering. *Multimedia Tools Appl.* 77, 22281–22297. doi:10.1007/s11042-018-5945-1

López-Cortés, X. A., Matamala, F., Maldonado, C., Mora-Poblete, F., and Scapim, C. A. (2020). A deep learning approach to population structure inference in inbred lines of maize. *Front. Genet.* 11, 543459. doi:10.3389/fgene.2020.543459

Masud, M. A., Huang, J. Z., Zhong, M., and Fu, X. (2019). Generate pairwise constraints from unlabeled data for semi-supervised clustering. *Data and Knowl. Eng.* 123, 101715. doi:10.1016/j.datak.2019.101715

Maulik, U., Mukhopadhyay, A., and Bandyopadhyay, S. (2009). Combining pareto-optimal clusters using supervised learning for identifying co-expressed genes. *BMC Bioinforma.* 10, 27–16. doi:10.1186/1471-2105-10-27

Mei, J. (2019). Semisupervised fuzzy clustering with partition information of subsets. *IEEE Trans. Fuzzy Syst.* 27, 1726–1737. doi:10.1109/tfuzz.2018.2889010

Mukhopadhyay, A., Maulik, U., and Bandyopadhyay, S. (2013). An interactive approach to multiobjective clustering of gene expression patterns. *IEEE Trans. Biomed. Eng.* 60, 35–41. doi:10.1109/TBME.2012.2220765

Pirooznia, M., Yang, J. Y., Yang, M. Q., and Deng, Y. (2008). A comparative study of different machine learning methods on microarray gene expression data. *BMC Genomics* 9, S13–S13. doi:10.1186/1471-2164-9-S1-S13

Reymond, P., Weber, H., Damond, M., and Farmer, E. E. (2000). Differential gene expression in response to mechanical wounding and insect feeding in arabidopsis. *Plant Cell* 12, 707–720. doi:10.1105/tpc.12.5.707

Rodriguez, A., and Laio, A. (2014). Machine learning. Clustering by fast search and find of density peaks. *Science* 344, 1492–1496. doi:10.1126/science.1242072

Rodríguez-Méndez, I. A., Ureña, R., and Herrera-Viedma, E. (2019). Fuzzy clustering approach for brain tumor tissue segmentation in magnetic resonance images. *Soft Comput.* 23, 10105–10117. doi:10.1007/s00500-018-3565-3

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* 20, 53–65. doi:10.1016/0377-0427(87)90125-7

Saha, S., and Bandyopadhyay, S. (2013). A generalized automatic clustering algorithm in a multiobjective framework. *Appl. Soft Comput.* 13, 89–108. doi:10.1016/j.asoc.2012.08.005

Song, X., Li, L., Srimani, P. K., Philip, S. Y., and Wang, J. Z. (2014). Measure the semantic similarity of go terms using aggregate information content. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* 11, 468–476. doi:10.1109/TCBB.2013.176

Vu, V., Labroche, N., and Bouchon-Meunier, B. (2012). Improving constrained clustering with active query selection. *Pattern Recognit.* 45, 1749–1758. doi:10.1016/j.patcog.2011.10.016

Wagstaff, K., Cardie, C., Rogers, S., and Schrödl, S. (2001). "Constrained k-means clustering with background knowledge," in Proceedings of the Eighteenth International Conference on Machine Learning (Burlington, MA, USA: Morgan Kaufmann Publishers Inc.), 577–584. doi:10.5555/645530.655669

Wu, W., and Ma, X. (2022). Network-based structural learning nonnegative matrix factorization algorithm for clustering of scrna-seq data. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* 20, 1–575. doi:10.1109/TCBB.2022.3161131

Xie, X. L., and Beni, G. (1991). A validity measure for fuzzy clustering. *IEEE Trans. Pattern Analysis Mach. Intell.* 13, 841–847. doi:10.1109/34.85677

Yin, X., Chen, S., Hu, E., and Zhang, D. (2010). Semi-supervised clustering with metric learning: An adaptive kernel method. *Pattern Recognit.* 43, 1320–1333. doi:10.1016/j.patcog.2009.11.005

Yu, Z., Luo, P., Liu, J., Wong, H., You, J., Han, G., et al. (2018). Semi-supervised ensemble clustering based on selected constraint projection. *IEEE Trans. Knowl. Data Eng.* 30, 2394–2407. doi:10.1109/tkde.2018.2818729

Zhang, G., Peng, Z., Yan, C., Wang, J., Luo, J., and Luo, H. (2022). Multigatae: A novel cancer subtype identification method based on multi-omics and attention mechanism. *Front. Genet.* 13, 855629. doi:10.3389/fgene.2022.855629

Zhang, M., Luo, W. J., and Wang, X. F. (2009). A normal distribution crossover for epsilon-moea. *J. Softw.* 20, 305–314. doi:10.3724/sp.j.1001.2009.00305

Zhao, M., and Li, D. (2022). "Multi-objective semi-supervised clustering algorithm based on constraint set optimization for gene expression data," in 2022 41st Chinese Control Conference (CCC) (Hefei, China: IEEE), 6570–6575. doi:10.23919/CCC55666.2022.9902131

Zhao, Y., Fang, Z., Lin, C., Deng, C., Xu, Y., and Li, H. (2021). Rfcell: A gene selection approach for scrna-seq clustering based on permutation and random forest. *Front. Genet.* 27, 665843. doi:10.3389/fgene.2021.665843

Zhou, Z., and Zhu, S. (2018). Kernel-based multiobjective clustering algorithm with automatic attribute weighting. *Soft Comput.* 22, 3685–3709. doi:10.1007/s00500-017-2590-y