Quantum Annealing for Enhanced Feature Selection in Single-Cell RNA Sequencing Data Analysis

Selim Romero[1,2,3], Shreyan Gupta[1,3], Victoria Gatlin[1,3], Robert S. Chapkin[2,3], James J. Cai[1,3,4*]

[1]Department of Veterinary Integrative Biosciences, Texas A&M University, College Station, TX 77843, USA.
[2]Department of Nutrition, Texas A&M University, College Station, TX 77843, USA.
[3]CPRIT Single Cell Data Science Core, Texas A&M University, College Station, TX 77843, USA.

[4]Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843, USA.

Keywords: Quantum annealing, quadratic unconstrained binary optimization (QUBO), feature selection, scRNA-seq, quantum computing

Abstract

Feature selection is vital for identifying relevant variables in classification and regression models, especially in single-cell RNA sequencing (scRNA-seq) data analysis. Traditional methods like LASSO often struggle with the nonlinearities and multicollinearities in scRNA-seq data due to complex gene expression and extensive gene interactions. Quantum annealing, a form of quantum computing, offers a promising solution. In this study, we apply quantum annealing-empowered quadratic unconstrained binary optimization (QUBO) for feature selection in scRNA-seq data. Using data from a human cell differentiation system, we show that QUBO identifies genes with nonlinear expression patterns related to differentiation time, many of which play roles in the differentiation process. In contrast, LASSO tends to select genes with more linear expression changes. Our findings suggest that the QUBO method, powered by quantum annealing, can reveal complex gene expression patterns that traditional methods might overlook, enhancing scRNA-seq data analysis and interpretation.

Introduction

Single-cell RNA sequencing (scRNA-seq) has transformed our understanding of cellular heterogeneity by providing a detailed view of gene expression at the individual cell level. This technology has enabled unprecedented exploration of gene expression programs that govern cell fate and regulate various cellular processes, such as stem cell differentiation and the epithelial-mesenchymal transition. However, dissecting the molecular mechanisms underlying these cellular processes remains a daunting task due to the complexity of gene function. A single gene may be involved in multiple cellular processes simultaneously, and genes often interact within intricate regulatory networks. In addition, functionally similar genes may compensate for one another, leading to genetic redundancy, which further complicates the identification of key genes involved in specific cellular processes.

Feature selection is a statistical technique used to identify a subset of input variables that are most relevant to a target variable. In single-cell research, feature selection is critical for identifying informative genes that capture essential biological insights while reducing data complexity, thereby enhancing the interpretability of biological questions. In this study, we focus on the feature selection problem for regression in scRNA-seq data, where the gene expression matrix holds numerical input variables, and a measurement of cell state serves as the numerical target variable. Our goal is to select a subset of genes that can accurately predict cell state. It is worth noting that the feature selection problem is more often formulated for classification tasks, such as distinguishing different cell types (YANG *et al*. 2021). When used for classification, the target variable (e.g., cell type) is discrete, whereas for regression, the target variable (e.g., cell state) is continuous. The Least Absolute Shrinkage and Selection (LASSO) is a popular method for feature selection (TIBSHIRANI 1996). LASSO can be used within embedded methods in other applications to reduce the complexity of the single-cell data, making it an efficient, interpretable, and effective method in handling high-dimensional data (YANG *et al*. 2021). However, LASSO is limited to linear models and may not capture nonlinear relationships. Thus, there is a clear need for new effective solutions given that conventional optimization methods e.g., LASSO and other linear regression methods, may struggle with the high dimensionality and nonlinearity inherent in scRNA-seq data, and may become trapped in local minima, potentially missing critical features.

Quantum computing, particularly quantum annealing, has emerged as a viable tool to tackle complex problems in data analysis (ARAI *et al*. 2023). This work explores the feasibility of using currently available quantum computer architectures to achieve quantum feature selection in single-cell data analysis. Specifically, we consider feature selection through a quadratic unconstrained binary optimization (QUBO) model, designed to identify features that are both independent and influential. Quadratic optimization is known to scale exponentially with the number of features, which typically pose a significant computational challenge. However, implementing QUBO on quantum annealers can offer a substantial speedup and an increased likelihood of finding the global minimum by exploring the solution space simultaneously through an adiabatic process. By harnessing the power of quantum annealing, QUBO-based feature selection may provide a more effective solution to the regression problems in scRNA-seq data analysis. To this end, we adapt the method proposed by (MÜCKE *et al*. 2023) and develop a quantum feature selection framework tailored to scRNA-seq data. This framework aims to address the limitations of traditional methods by capturing complex, nonlinear relationships in gene expression that are crucial for understanding cellular processes.

Results

Framework of QUBO feature selection for regression

We tackled a regression task using a scRNA-seq dataset X consisting of p cells and n genes, with a target variable T representing the cell state to be predicted. The feature selection problem was framed as identifying a subset of these n genes that can achieve performance comparable to the original dataset for regression tasks. This was accomplished by solving an optimization problem depicted in **Fig. 1**, where the objective was to find the optimal feature set (vector $\boldsymbol{F}^*$) that minimized the QUBO cost function (matrix $\boldsymbol{Q}(\boldsymbol{F}, \alpha; \boldsymbol{I}, \boldsymbol{R})$), counting for both the importance and redundancy of features. The solution $\boldsymbol{F}^*$ is a binary vector representing the selected features (genes), where $F_i^* = 1$ indicates that the i-th feature is selected. The implementation leverages the parameter α to balance the importance and redundancy to construct QUBO matrix $\boldsymbol{Q}$ for annealing on a quantum computer. This approach ensures that the most informative genes are captured, while minimizing redundancy and enhancing the interpretation and efficiency of the resulting gene set for regression analysis. To estimate feature importance and redundancy in order to construct a single cost function, we followed the previous study (MÜCKE *et al.* 2023) and adapted the mutual information. The framework was adapted by altering importance and redundancy measurements, such as entropy or Pearson's correlation, and the incorporation of prior knowledge.

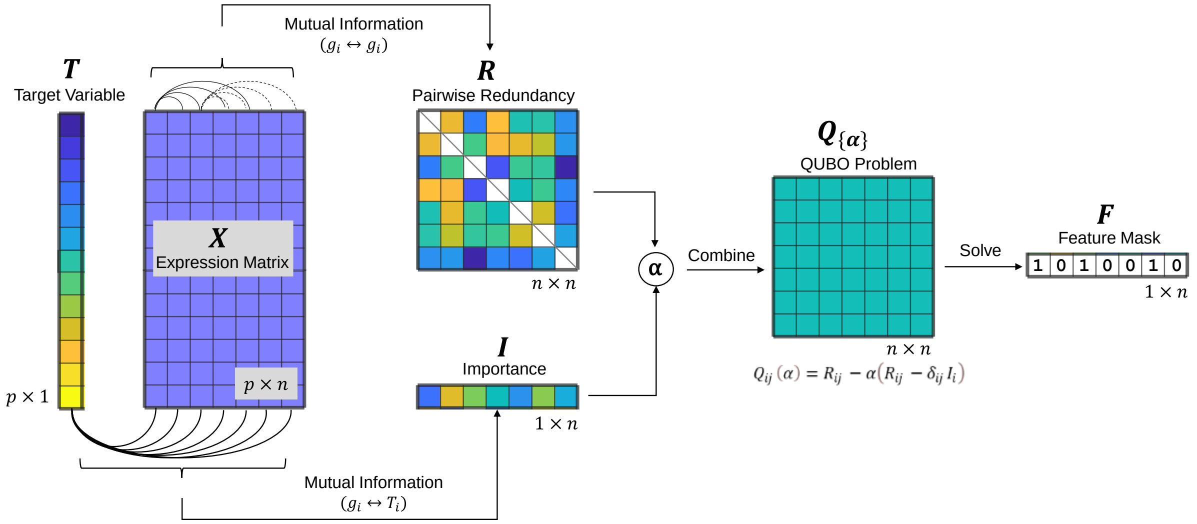

Fig. 1. Quantum annealing-based QUBO feature selection for regression in scRNA-seq data analysis. The process begins with an scRNA-seq gene expression matrix $\boldsymbol{X} \in \mathbb{R}^{p \times n}$, where p represents the number of cells and n the number of genes, along with a provided cell state vector $\boldsymbol{T}$ (the target variable). Mutual information is used to compute the redundancy matrix $\boldsymbol{R}$ and the importance vector $\boldsymbol{I}$. A balancing parameter α is then applied to weight importance against redundancy, resulting in the QUBO matrix $\boldsymbol{Q}$. The QUBO is subsequently solved using a quantum annealer, producing a binary solution vector $\boldsymbol{F}^*(\boldsymbol{x}^*)$, which acts as a bitmask to indicate the selected features.

Solving QUBOs using quantum annealing

Quantum annealing leverages quantum mechanics, specifically utilizing superposition and tunneling effects, to solve optimization problems. In this study, we employed the quantum annealer from D-Wave Systems, specifically designed to address QUBO problems, accessed via the Ocean software development kit, a Python library that interfaces with the quantum annealer. The annealer uses qubits, which can exist in a superposition state of 0 and 1, enabling the simultaneous exploration of many potential solutions. These qubits are superconducting loops, controlled by electric currents and magnetic fields, which create an "embedding landscape" on the chip to guide the optimization process. To validate the results of quantum annealing, we compared the features selected by the D-Wave quantum annealer through the Ocean software development kit with those selected using classical annealing. The latter employed an iterated tabu search algorithm (Palubeckis 2006), implemented in the MATLAB quantum computing package. In all tested cases, whether using quantum or classical annealing, the results were consistent. Overall, the results obtained from the quantum annealer were cross-validated by the independent classical QUBO solver, confirming that the quantum annealer's selections were fully corroborated by classical methods.

Simulation evaluation of QUBO feature selection

To numerically validate the effectiveness of quantum annealing for nonlinear feature selection, we conducted a simulation study using synthetic data. We began by generating a normally distributed random matrix B of size $n \times n$ with $n = 50$ features. From this, we computed the correlation matrix R derived from the covariance matrix $C = B^T B$. We then generated multivariate normal data X consisting of $p = 10{,}000$ observations and n features, with a mean $\mu = 0$ and a covariance matrix $\sigma = R$. Correlations were introduced between source features with indices $s = [5, 11, 7, 1, 14]$ and corresponding target features with indices $t = [16, 17, 18, 19, 20]$ using the relation $X_{i,t} = X_{i,s} + \rho Y_i$, where $\rho = 0.1$ is a correlation coefficient, and Y_i is a random normal noise term for the i-th observation. The target variable y was constructed using a nonlinear function that depends on the source features:

$$y_i(X_{i,s}) = 0.5\cos(7X_{i,s_4}) + \sin(X_{i,s_3}X_{i,s_2}) + 0.1\exp(X_{i,s_5})\log_2|10X_{i,s_1}| + \rho_1 Y_i',$$

where Y_i' is an additional random noise term. The function y_i was designed to be highly nonlinear, with synthetic data influencing the target function. We normalized the target variable $\hat{y}$ to the range $[0,1]$ and applied z-score standardization to the data $\hat{X}$. We applied QUBO feature selection to the standardized data $\hat{X}$ and predictor $\hat{y}$ to identify features, aiming to recover the original source features s. The results were compared with those obtained using LASSO feature selection. The QUBO model successfully identified the features $[14, 1, 11, 7\ , 5]$, recovering all source features with a 100% success rate in this simulated scenario. This result indicates the QUBO model's robustness in detecting nonlinear relationships embedded in the predictors. We attribute this performance to the combined influence of mutual information and balanced redundancy-importance setting, enhancing predictive power in complex problems. In contrast, the LASSO model identified the features $[14, 1, 2, 3, 4]$, achieving only a 40% success rate. The LASSO model was configured with cross-validation to estimate the mean squared error (MSE), using a 10-fold cross-validation approach. The largest magnitude coefficients within one standard error were selected. Although the LASSO model identified two of the source features, its performance was less impressive compared to the QUBO model. In summary, the QUBO feature selection method outperformed LASSO in this highly nonlinear scenario by producing more accurately identifying relevant features. This suggests that QUBO feature selection may be particularly effective in identifying functionally relevant genes in data derived from complex biological systems.

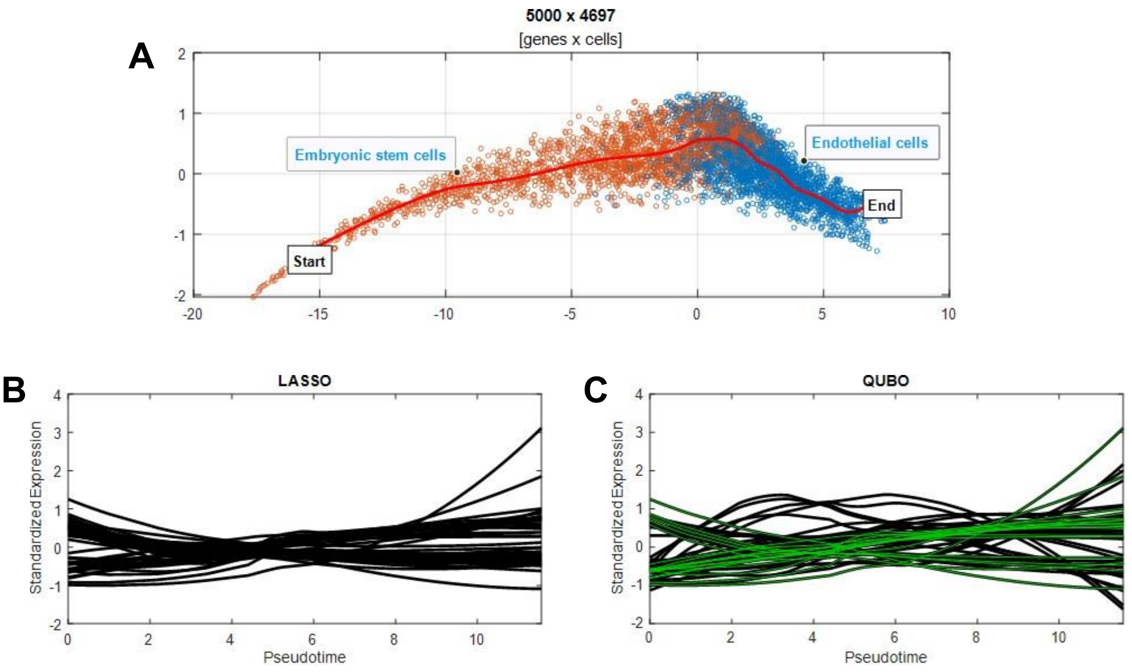

Fig. 2. QUBO-based feature selection in stem cell differentiation into endothelial cells. (**A**) PHATE projection of scRNA-seq data, with cells colored by their inferred cell type assignment. The red curve indicates the pseudotime trajectory, tracing the transition from embryonic stem cells (ESCs) to endothelial cells (ECs). (**B**) Locally weighted linear regression (LOESS)-smoothed expression profiles of the top 50 genes identified by LASSO were plotted as a function of pseudotime. (**C**) Same as (**B**) but displaying the expression profiles of the top 50 genes identified through QUBO-based feature selection. Green lines denote the 18 genes identified by both LASSO and QUBO-based methods.

QUBO feature selection identifies key genes implicated in cell differentiation

To evaluate the performance of QUBO-based feature selection, we used a published scRNA-seq data set comprising human embryonic stem cells (hESCs) and derived endothelial cells (ECs) (XU *et al*. 2023). The hESC-EC cell line was generated using the “FLI1-PKC system,” a well-established, high-efficiency hESC-EC induction method for studying cellular processes involving cell differentiation (ZHAO *et al*. 2018). After preprocessing the scRNA-seq data (see **Methods** for details), we used PHATE, a visualization method that captures both local and global nonlinear structure in scRNA-seq data (MOON *et al*. 2019), to embed cells (**Fig. 2A**). Subsequently, pseudotime trajectory inference was performed using the splinefit method (CAI 2019), to construct a curve representing the path of cell transition during hESC-EC differentiation. Here, single-cell data is considered as a snapshot of a continuous process, and the trajectory is reconstructed by finding paths through cellular space that minimize transcriptional changes between neighboring cells (LUECKEN AND THEIS 2019). The ordering of cells along these paths is described by a pseudotime variable. Each cell was projected onto the curve and assigned with a pseudotime value. These estimated pseudotime values were utilized as the target variable $\boldsymbol{T}$, for which regression models were constructed. We applied both LASSO and QUBO feature selection methods to select the top 50 genes for their respective regression models. Among these top genes, 18 genes (ANP32E, APLN, ARPC3, ATF5, CAMTA1, CLDN7, EIF2S2, FABP5, FLNC, GNG3, GNG5, HMGB3, ID1, IER2, OAZ1, PDAP1, THY1 and ZNF706) were identified by both LASSO and QUBO. Many of them are known to be implicated in the cellular process of hESC-EC differentiation. For example, APLN, derived from Apelin, regulates the development of endothelial cells and promotes vascular repair (MASOUD *et al*. 2020). ID1 directly upregulates VEGF, promoting the proliferation and migration of endothelial cells (QIU *et al*. 2022). THY1 (CD90), commonly known as a stemness marker due to its role in cellular growth and development, is actively involved in angiogenesis (LEE *et al*. 1998).

Next, we focused on genes identified using QUBO-based feature selection. Among them, for example, DYNLT1 codes for Tctex-1, which directly interacts with bone morphogenetic protein receptor 2 (BMPR-II), playing a critical role in regulating angiogenesis (WILEY AND JIN 2011); YWHAB codes Protein Kinase C Inhibitor Protein 1, an inhibitor for PKC, crucial in endothelial cell specification (ZHAO *et al*. 2018). To understand the function of the top genes exclusively identified using QUBO as a whole set, we conducted gene function enrichment analysis using EnrichR (KULESHOV *et al*. 2016). Genes uniquely identified by QUBO feature selection were found to be involved in several cellular processes, including *regulation of smooth muscle cell proliferation* (GO:0048660), *regulation of vascular associated smooth muscle cell proliferation* (GO:1904705), and *blood vessel morphogenesis* (GO:0048514). Additionally, these genes were associated with pathways such as *tight junctions* (a KEGG term for protein complexes forming semi-permeable connections between endothelial cells), *EPH-ephrin signaling pathway* (R-HSA-2682334), *VEGF-mediated signaling*, and *signaling by GPCR* (R-HSA-372790).

Finally, we plotted gene expression levels as a function of pseudotime to visualize the changes in individual gene expression during hESC-EC differentiation. The top 50 genes selected by LASSO exhibited patterns of monotonic increases or decreases over time (**Fig. 2B**). In contrast, the top 50 genes identified by QUBO displayed more nonlinear changes as time progressed (**Fig. 2C**).

Discussion

Most people envision our solar system with planets in neat circles around the sun, this heliocentric model is a simplification. Our solar system travels through space at a whopping 70,000 kilometers per hour. The sun, like any star at the center of its own gravitational system, influences the planets to revolve around it. This motion is more accurately described as a helical path, rather than a simple flat orbit. In this work, we show how a similar perspective is key to effectively exploiting cellular processes. In single-cell data analysis, cellular processes such as stem cell differentiation and epithelial-mesenchymal transition are often simplified into linear processes, which is far from reality. In complex biological systems, differentiating and transitioning cells may also undergo or experience cell growth, division, and death. Inside cells, cell proliferation signaling pathways modulate various gene expression programs, making the study of internal controls responsible for regulating any cellular processes challenging.

Classical feature selection methods like LASSO primarily focus on individual target variable importance, which might lead to the inclusion of redundant features that holds similar information, thereby increasing model complexity. In contrast, QUBO feature selection, with its ability to explore a broader solution space, can potentially identify and remove these redundant features while selecting the most informative ones, resulting in a more concise and efficient feature set. Additionally, feature selection in high-dimensional datasets becomes computationally expensive with classical methods due to iterative processes. Thus, it is noteworthy that QUBO feature selection translates the selection process into a form suitable for its hardware, enabling it to explore a vast number of possibilities simultaneously and potentially address the computational complexity inherent in high-dimensional feature selection. Traditional methods also struggle with complex relationships between features, often overlooking valuable combinations. However, the inherent nature of quantum mechanics allows QUBO feature selection to explore intricate connections between features, which can lead to identifying feature subsets that capture complex interactions, thus improving model performance.

One of the most significant findings of our study is QUBO's superior ability to identify genes with nonlinear expression patterns during cell differentiation. QUBO successfully detected a broader range of expression dynamics. This capability is particularly crucial in the context of scRNA-seq data, where complex gene interactions and regulatory networks often result in nonlinear expression profiles. The identification of these nonlinear patterns provides a more comprehensive view of the gene regulatory landscape during cell differentiation. It allows researchers to capture subtle yet potentially critical changes in gene expression that might be overlooked by traditional linear methods. Importantly, many of the genes identified by QUBO as having nonlinear expression patterns are functionally implicated in the cell differentiation process. This finding suggests that our quantum annealing-based approach is not just detecting mathematical nonlinearities but is uncovering biologically meaningful gene expression changes. The functional relevance of these genes underscores the potential of QUBO to provide insights that are both statistically robust and biologically significant.

Pseudotime is just an example of a target variable. Many other cellular continuous variables can be used in the QUBO feature selection for regression problems. This represents a significant step forward in the analysis and interpretation of scRNA-seq data along with other cell state measurements, offering a complementary approach to traditional statistical methods. Furthermore, the implications of this study extend beyond scRNA-seq data analysis. The successful application of quantum computing with respect to complex biomedical data analysis suggests broader possibilities for integrating quantum techniques into other areas of biological research. As quantum computing continues to evolve, its potential to address computational challenges in big data and high-dimensional datasets could lead to groundbreaking advancements in various fields.

Methods

Processing of scRNA-seq data

The scRNA-seq data was obtained from a published study on a human embryonic stem cell (hESC) to endothelial cell (EC) differentiation system (XU *et al.* 2023). The induction system employed overexpression of the transcription factor FLI1 to induce hESC differentiation into ECs. This induction approach has been shown to be more efficient than cytokine stimulation (XU *et al.* 2023). Data generated from the cell sample after 24 hour FLI1 induction was selected. The scRNA-seq count matrix was imported into scGEAToolbox (CAI 2019) for data processing, which included quality control filtering, cell type annotation, and selection of highly variable genes. In all of these steps, the default parameter setting was used. The processed data contained 5,000 genes and 4,697 cells (consisting of about equal number of hESCs and ECs). PHATE (MOON *et al.* 2019) was used to embed and visualize cells. The splinefit algorithm (CAI 2019) was used for trajectory inference to estimate the pseudotime of cells. The estimated pseudotime was used as the target variable $\boldsymbol{T}$ for constructing the cross-entropy terms in the mutual information calculation in Equations (5). For regression analysis, the gene-by-cell count matrix was transformed using Pearson residual transformation (LAUSE *et al.* 2021). The QUBO formulation, detailed in the next section, was then applied to identify $k = 50$ informative features (genes) from the processed data. The computation was conducted using quantum annealing via the Ocean SDK hybrid solver (STOGIANNOS *et al.* 2022) in D-Wave's quantum annealer, as well as using the iterated tabu search algorithm (PALUBECKIS 2006) in the MATLAB quantum computing package.

Transition from Ising model to QUBO problem

The Ising model, derived from statistical mechanics and inspired by ferromagnetism, employs a Hamiltonian operator to define an energetic state or cost function as follows,

$$\hat{H}(\boldsymbol{\sigma}) = -\sum_{i=1}^{N} h_i \sigma_i - \sum_{i=1}^{N} \sum_{i<j}^{N} J_{ij} \sigma_i \sigma_j \,. \tag{1}$$

Here $\boldsymbol{\sigma}$ represents the spin state operator, indicating the "up" or "down" magnetic states (BROOKE *et al.* 1999). This can naturally be translated to classical binary information, making it suitable for translating classical information to quantum information. The Ising model's Hamiltonian in Equation (1) defines an energy state, analogous to a cost function in optimization problems, and uses the spin state operator and spin-spin coupling constants J_{ij} to capture interaction between spins. An external magnetic field h_i tunes the ground states for a particular problem. The expected value of the Hamiltonian determines the energy or cost function ($\langle H \rangle = E$). The parametric dependence on J_{ij} and h_i defines the energy/cost function value, while the expectation value of the Hamiltonian carries implicitly a probabilistic quantum behavior. Interestingly, the Ising model's mathematical formulation resembles that of QUBO problem, which is a combinatorial optimization problem commonly encountered in various fields (MÜCKE *et al.* 2023). QUBO feature selection has become noticeable, since there are some applications for the quantum computing area such as ranking and classifying QA-FS (DACREMA *et al.* 2022) and feature selection applied to recommender system (NEMBRINI *et al.* 2021).

The QUBO problem aims to minimize a cost function $f(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{Q} \boldsymbol{x}$, where $\boldsymbol{Q}$ is an upper triangular matrix, and $\boldsymbol{x} = (x_1, \dots, x_N)^T$ contains x_i binary elements. The QUBO matrix $\boldsymbol{Q}$ encodes the problem's constraints and objectives, while the column vector $\boldsymbol{x}$ represents the variables to be optimized. Usually, $f(\boldsymbol{x})$ diagonal elements $(q_i Q_{ii} q_i)$ can be simplified due to binary representation $(q_i q_i = q_i)$ as follows,

$$f(\boldsymbol{x}) = \sum_{i=1}^{N} Q_{ii} x_i + \sum_{i=1}^{N} \sum_{i<j}^{N} Q_{ij} x_i x_j. \tag{2}$$

the annealing process aims to minimize the energetic configuration, akin to finding the ground spin state for the encoded problem. As pointed out before, QA can accomplish the cost function minimization in a better way than classical computing. Interestingly, the objective minimization is similar to the least squares problem, where matrices $\boldsymbol{A}$, $\boldsymbol{x}$, and $\boldsymbol{b}$ are used to minimize $||\boldsymbol{Ax}-\boldsymbol{b}||^2$ for an optimal $\boldsymbol{x}$, emphasizing that $||\mathbf{Ax}-\mathbf{b}||$ and Q despite being described in a quadratic form are described in a set of linear equations (JUN 2024).

QUBO feature selection model construction

The feature selection problem aims to identify a subset $S \subset [n]$ of informative features from the original set $[n] = \{1,2, \dots, n\}$, where n is the total number of features. Suppose a dataset $\boldsymbol{D} \coloneqq \{(\boldsymbol{x}^i, \boldsymbol{\tau^i})\}_{i\in[n]}$ containing n features and p observations, where $\boldsymbol{x}^i \in \boldsymbol{X} \subseteq \mathbb{R}^p$ represents the feature vectors and $\boldsymbol{\tau}^i \in \boldsymbol{T} \subseteq \mathbb{R}^p$ represents the target variable. The goal is to preserve the nature of the data while reducing its size for better interpretation and reduced complexity by obtaining a subset of the original dataset $D_S = \{(\boldsymbol{x}_S^i, \boldsymbol{\tau^i})\}_{i\in[n]}$ with $\boldsymbol{x}_S^i \in \boldsymbol{X}_S \subseteq \mathbb{R}^p$. In our implementation, we aim to accurately describe the underlying biological processes with a representative set of features/genes.

Interestingly, this task can be formulated as an optimization problem with a cost function. We propose a novel QUBO-based feature selection approach (MÜCKE *et al.* 2023) for scRNA-seq data. The QUBO formulation seeks the optimal feature vector $\boldsymbol{F}$ that minimizes the cost function $\boldsymbol{Q}$ as follows,

$$\boldsymbol{F}^* \coloneqq arg_{\boldsymbol{F}\in\{0,1\}^n}\, min\, \boldsymbol{Q}(\boldsymbol{F}, \alpha; \boldsymbol{I}, \boldsymbol{R}), \tag{3}$$

where $\boldsymbol{F}^*$ is a binary vector containing $F_i^* = 1$ for selected features. Here, $\boldsymbol{I}$ and $\boldsymbol{R}$ are parameters defining the cost function. $\boldsymbol{I}$ represents the MI within feature $\boldsymbol{x}^i$ and target $\boldsymbol{\tau}^i$, such that $I_i \in \boldsymbol{I} \subseteq \mathbb{R}^n$. $\boldsymbol{R}$ represents the MI within features $\boldsymbol{x}^i$ and $\boldsymbol{x}^j$, such that $R_{ij} \in \boldsymbol{R} \subseteq \mathbb{R}^{n\times n}$. α is a parameter between 0 and 1 used to balance the solution. Our QUBO cost function $\boldsymbol{Q}$ is defined as follows,

$$\boldsymbol{Q}(\boldsymbol{F}, \alpha; \boldsymbol{I}, \boldsymbol{R}) \coloneqq -\alpha \sum_{i=1}^{n} I_i F_i + (1-\alpha) \sum_{i,j=1}^{n} R_{ij} F_i F_j. \tag{4}$$

The $I_i = I(\boldsymbol{x}^i; \boldsymbol{\tau^i})$ (Equation (5)) is the i-th importance element. The redundancy element $R_{ij} = I(\boldsymbol{x}^i; \boldsymbol{x}^j)$ (Equation (6)) is the MI between the i-th and j-th features. Since a self-feature is never redundant, we set $R_{ii} = 0$. MI values are positive, indicating strong interactions with higher values and no interaction near zero. This framework evaluates the importance of the target $\boldsymbol{\tau}^i$ in relation to individual features, where crosstalk between features reduces the importance of some variables. This determines each feature's overall relevance to the response variable $\boldsymbol{\tau}^i$, is balanced by the parameter α $(0 \le \alpha \le 1)$.

One challenge is that directly estimating mutual information from real-world data is difficult due to the requirement for the joint probability mass function of the features and target variables. To address this challenge, we employed a binned discretization approach. Here, each feature was divided into B bins using quantiles, ensuring a fair distribution of the data across the bins. Similarly, we applied discretization to our target variable since it is a continuous variable. The binning process is as follows:

- Quantile calculation: For each feature $\boldsymbol{x}^i$, we calculate the $(B+1)$ quantiles q_i^L for $L \in \{0,1,\dots,B\}$. These quantiles create the boundaries of the bins.
- Binning: Each bin β_i^L is defined as the interval $\left[q_i^{L-1}, q_i^L\right)$ for $L \in \{1,2,\dots,B-1\}$, and the final bin $\beta_i^B$is $\left[q_i^{B-1}, q_i^B\right]$.
- Assigning bins: Each feature value x_j^i is assigned to a bin based on which interval it falls into. For instance, if x_j^i falls between q_i^{L-1} and q_i^L, it is assigned to bin L $\left(x_j^i \in \beta_i^L\right)$.

The discretized features $\hat{\boldsymbol{x}}^i$ and target variable $\hat{\boldsymbol{\tau}}^i$ are integrated into the discretized data $\widehat{D} = \left\{\left(\hat{\boldsymbol{x}}^i, \hat{\boldsymbol{\tau}}^i\right)\right\}_{i\in[n]}$, where $\hat{\boldsymbol{x}}^i \in [\beta_{i,x}] \subseteq \mathbb{R}^B$ and $\hat{\boldsymbol{\tau}}^i \in [\beta_{i,\tau}] \subseteq \mathbb{R}^B$ with $\beta_{i,\boldsymbol{k}} = \left\{\beta_{i,\boldsymbol{k}}^1, \dots, \beta_{i,\boldsymbol{k}}^B\right\}$ for all $i \in [n]$. This binning approach allows us to approximate the information entropy across features and target variables. Thus, importance $(\boldsymbol{I})$ and redundancy $(\boldsymbol{R})$ are described as follows,

$$I_i \coloneqq I\left(\boldsymbol{x}^i; \boldsymbol{\tau}^i\right) \approx \sum_{\hat{\boldsymbol{x}}^i \in [\beta_{i,x}]} \sum_{\hat{\boldsymbol{\tau}}^i \in [\beta_{i,\tau}]} \hat{p}\left(\hat{\boldsymbol{x}}^i, \hat{\boldsymbol{\tau}}^i\right) log\left(\frac{\hat{p}\left(\hat{\boldsymbol{x}}^i, \hat{\boldsymbol{\tau}}^i\right)}{\hat{p}\left(\hat{\boldsymbol{x}}^i\right)\hat{p}\left(\hat{\boldsymbol{\tau}}^i\right)}\right), \tag{5}$$

$$R_{ij} \coloneqq I\left(\boldsymbol{x}^i; \boldsymbol{x}^j\right) \approx \sum_{\hat{\boldsymbol{x}}^i \in [\beta_{i,x}]} \sum_{\hat{\boldsymbol{x}}^j \in [\beta_{j,x}]} \hat{p}\left(\hat{\boldsymbol{x}}^i, \hat{\boldsymbol{x}}^j\right) log\left(\frac{\hat{p}\left(\hat{\boldsymbol{x}}^i, \hat{\boldsymbol{x}}^j\right)}{\hat{p}\left(\hat{\boldsymbol{x}}^i\right)\hat{p}\left(\hat{\boldsymbol{x}}^j\right)}\right). \tag{6}$$

Equations (5) and (6) involve summation over discretized bins and calculate the log-ratio between joint and marginal probabilities estimated from discretized data $\widehat{D}$. The discretized empirical probabilities mass function is defined as,

$$\hat{p}\left(\hat{\boldsymbol{x}}^i, \hat{\boldsymbol{\tau}}^i\right) \coloneqq \frac{1}{n} \sum_{(\hat{\boldsymbol{x}}', \hat{\boldsymbol{\tau}}') \in \widehat{D}} \mathbb{I}_{\{\hat{\boldsymbol{x}}^i = \hat{\boldsymbol{x}}' \wedge \hat{\boldsymbol{\tau}}^i = \hat{\boldsymbol{\tau}}'\}}\,, \tag{7}$$

with an indicator function,

$$\mathbb{I}_{\{P\}} \coloneqq \begin{cases} 1 \text{ if P is true} \\ 0 \text{ otherwise} \end{cases}, \tag{8}$$

defined for logical statements P. A semi-last remark, Equation (4) can be simplified as $Q_{ij}(\alpha) = R_{ij} - \alpha\left(R_{ij} - \delta_{ij} I_i\right)$, where the Kronecker delta δ_{ij} is unity if $i = j$ and 0 otherwise. Interestingly, this QUBO matrix $\boldsymbol{Q}$ can find an optimal or a quasi-optimal solution to the feature selection problem. It also avoids gauge problems commonly encountered in QUBO formulations. In our simulations, we consider one predictor for all features $\boldsymbol{\tau}^i \equiv \boldsymbol{T}$. Importantly, while the QUBO model is unconstrained, additional penalty terms could be introduced to customize the energy landscape for specific requirements ($Q' = Q^{problem} + MQ^{constraint}$).

Code Availability

Code implementing the method described in this paper is available at https://github.com/cailab-tamu/QUBO_feature_selection along with example data.

Acknowledgment

We are grateful to Dr. Liang Hu for sharing the hESC-EC scRNA-seq data. We acknowledge the use of advanced computing resources provided by Texas A&M High Performance Research Computing in conducting parts of this research. This study was funded by the U.S. Department of Defense (DoD, GW200026) for J.J.C, Allen Endowed Chair in Nutrition & Chronic Disease Prevention for R.S.C., and the Cancer Prevention & Research Institute of Texas (CPRIT, RP230204) for J.J.C. and R.S.C.

References

Arai, S., H. Oshiyama and H. Nishimori, 2023 Effectiveness of quantum annealing for continuous-variable optimization. Physical Review A 108: 042403.

Brooke, J., D. Bitko, T. F. Rosenbaum and G. Aeppli, 1999 Quantum annealing of a disordered magnet. Science 284: 779-781.

Cai, J. J., 2019 scGEAToolbox: a Matlab toolbox for single-cell RNA sequencing data analysis. Bioinformatics.

Dacrema, M. F., F. Moroni, R. Nembrini, N. Ferro, G. Faggioli *et al.*, 2022 Towards Feature Selection for Ranking and Classification Exploiting Quantum Annealers. Proceedings of the 45th International Acm Sigir Conference on Research and Development in Information Retrieval (Sigir '22): 2814-2824.

Jun, K., 2024 QUBO formulations for a system of linear equations. Results in Control and Optimization 14: 100380.

Kuleshov, M. V., M. R. Jones, A. D. Rouillard, N. F. Fernandez, Q. Duan *et al.*, 2016 Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res 44: W90-97.

Lause, J., P. Berens and D. Kobak, 2021 Analytic Pearson residuals for normalization of single-cell RNA-seq UMI data. Genome Biology 22.

Lee, W. S., M. K. Jain, B. M. Arkonac, D. Zhang, S. Y. Shaw *et al.*, 1998 Thy-1, a novel marker for angiogenesis upregulated by inflammatory cytokines. Circulation Research 82: 845-851.

Luecken, M. D., and F. J. Theis, 2019 Current best practices in single-cell RNA-seq analysis: a tutorial. Mol Syst Biol 15: e8746.

Masoud, A. G., J. X. Lin, A. K. Azad, M. A. Farhan, C. Fischer *et al.*, 2020 Apelin directs endothelial cell differentiation and vascular repair following immune-mediated injury. Journal of Clinical Investigation 130: 94-107.

Moon, K. R., D. van Dijk, Z. Wang, S. Gigante, D. B. Burkhardt *et al.*, 2019 Visualizing structure and transitions in high-dimensional biological data. Nat Biotechnol 37: 1482-1492.

Mücke, S., R. Heese, S. Müller, M. Wolter and N. Piatkowski, 2023 Feature selection on quantum computers. Quantum Machine Intelligence 5.

Nembrini, R., M. F. Dacrema and P. Cremonesi, 2021 Feature Selection for Recommender Systems with Quantum Computing. Entropy 23.

Palubeckis, G., 2006 Iterated Tabu Search for the Unconstrained Binary Quadratic Optimization Problem. Informatica 17: 279-296.

Qiu, J., Y. H. Li, B. Y. Wang, X. Y. Sun, D. D. Qian *et al.*, 2022 The Role and Research Progress of Inhibitor of Differentiation 1 in Atherosclerosis. DNA and Cell Biology 41: 71-79.

Stogiannos, E., C. Papalitsas and T. Andronikos, 2022 Experimental Analysis of Quantum Annealers and Hybrid Solvers Using Benchmark Optimization Problems. Mathematics 10.

Tibshirani, R., 1996 Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society Series B-Statistical Methodology 58: 267-288.

Wiley, D. M., and S. W. Jin, 2011 Bone Morphogenetic Protein functions as a context-dependent angiogenic cue in vertebrates. Seminars in Cell & Developmental Biology 22: 1012-1018.

Xu, X., J. Chen, H. Zhao, Y. Pi, G. Lin *et al.*, 2023 Single-Cell RNA-seq Analysis of a Human Embryonic Stem Cell to Endothelial Cell System Based on Transcription Factor Overexpression. Stem Cell Rev Rep 19: 2497-2509.

Yang, P., H. Huang and C. Liu, 2021 Feature selection revisited in the single-cell era. Genome Biol 22: 321.

Zhao, H., Y. Zhao, Z. L. Li, Q. Ouyang, Y. Sun *et al.*, 2018 FLI1 and PKC co-activation promote highly efficient differentiation of human embryonic stem cells into endothelial-like cells. Cell Death & Disease 9.