問都給你問_ENG

1. Did you consider using other clustering methods besides complete linkage? If so, can you discuss why you ultimately chose complete linkage as your final clustering method and why this method was chosen over others?
    - In the early stages of our research, we used the single linkage hierarchical clustering method, and later, we also tried K-means as a clustering approach. However, we ultimately chose complete linkage as our final clustering method because it provides more coherent results, which aids in gaining a more comprehensive understanding of the interrelationships between diseases.
2. Why did you choose to use complete linkage as the clustering method? Are there specific reasons or prior research supporting this decision?
    - Relative to single linkage, complete linkage and average linkage methods are better at considering the similarity between diseases, which helps to gain deeper insights into disease correlations. We compared complete linkage and average linkage and found that complete linkage provided results that better represented disease relationships.
3. Why did you opt for a clustering approach to classify diseases instead of other methods, such as supervised classification models? What are the advantages of this approach in this study?
    - We chose a clustering approach to classify diseases primarily to explore comorbidity relationships between different diseases. We considered the disease occurrence

profiles as features and calculated their similarity to understand disease relationships. Using a distance matrix transformed from these relationships for hierarchical clustering helps group similar diseases together, providing a deeper understanding of the interplay between different diseases.

4. You mentioned that you chose to divide diseases into five clusters. Are there specific reasons for this choice? Did you consider other numbers of clusters, and why did you ultimately choose five?
    - The decision to divide diseases into five clusters was primarily based on the initial clustering results. When we used the single linkage clustering method, we initially had seven clusters. Therefore, when we switched to the current approach, we started with the basis of seven clusters and adjusted it according to the actual distribution and results. Dividing into six or seven clusters would have been based on the five-cluster result, and it wouldn't have provided substantial insights for our analysis.
5. You mentioned that you tried K-means. Why did you not ultimately adopt this method? Are there specific reasons or advantages that led you to choose complete linkage clustering?
    - We did indeed attempt to use K-means as a clustering method, but the reason for not selecting it lies in the fact that K-means requires specifying the number of clusters during initialization. Hierarchical clustering, on the other hand, does not require such specification and can automatically adjust the number of clusters based on the results. Our goal was to find diseases with related themes

or high similarities, and that's why we chose hierarchical clustering.

6. Is there a contradiction between the low variation required for differential methylation analysis and the need for low variation in internal control gene loci? Are the differences the same?
    - No, they are not the same.
    - The primary purpose of differential methylation analysis is to compare the methylation differences at specific gene loci between different conditions or groups. **Through statistical analysis, we filter out gene loci that exhibit significant statistical differences.**
    - On the other hand, internal control gene loci are characterized by their expected stability regardless of the time, space, or experimental method. This means that the methylation state at these loci should be relatively consistent under various conditions. These loci serve as a baseline or reference.
7. Filtered DMPs already show significant differences. Why do you consider them suitable as candidate internal control gene loci references?
    - Firstly, we followed a **standard differential methylation analysis procedure**. This means that our analysis method is based on **established and widely used standard practices**. This ensures the reliability of our results and that our results can be compared with other studies using the same methods.
    - We observed that these DMPs exhibit **high stability across multiple repeated experiments**. This consistency provides confidence that these DMPs can be used as internal controls in DNA methylation experiments. The key

characteristic of internal controls is their **expected stability under different experiments or conditions**.

- Moreover, we noticed that the methylation levels of these DMPs remain relatively stable **across different experiments**. This ensures that these DMPs are not significantly affected by minor variations in the experiments, providing a stable baseline for comparing other potential variations or differences.
- Combining these points, we believe they possess the ideal characteristics for internal controls.

8. Why did you choose to use Accuracy as the evaluation criterion?
    - In our dataset, the occurrence rates of the diseases we selected are nearly normally distributed, indicating a relatively balanced distribution between suffering and not suffering instances. In this context, Accuracy is a reasonable and representative metric since the model's random guessing is less likely to achieve high accuracy.
    - **If they ask why other metrics like Precision, Recall, and F1 Score were not used as evaluation criteria:**
        - Although Precision, Recall, and F1 Score are valuable metrics, especially when considering the model's performance in specific scenarios or focusing on specific types of errors, in our context, our primary concern was the overall predictive capability of diseases using the selected internal control gene features. We were mainly interested in how well the model could correctly predict diseases overall and not delving into specific prediction errors. Thus, Accuracy became an intuitive and suitable metric that best reflects the core objective of our study.

    - We wanted the results to be communicated in a simple and straightforward manner. Accuracy is an easily understandable metric that directly tells us the percentage of correctly predicted cases, which is intuitive for a non-specialized audience.

9. Why did you choose to use SVM (Support Vector Machine) instead of Random Forest to build the disease prediction model? Are there specific reasons or advantages for this choice? Can you explain the considerations behind your decision?
    - We chose SVM primarily because it exhibits a maximal-margin property, making it well-suited for smaller datasets. Random Forest, on the other hand, typically relies on larger sample sizes to ensure diversity among its constituent trees. Particularly when dealing with a smaller number of features, SVM can be more effective at preventing overfitting by tuning regularization parameters.
    - While Random Forest generally has good generalization performance, it can be challenging to construct multiple diverse trees when the number of features is very limited. SVM, by adjusting regularization parameters, can more easily avoid overfitting.
    - Interpretability: Additionally, SVM provides good interpretability, offering an intuitive display of the importance of each feature. Although Random Forest also provides measures of feature importance, SVM's weights are typically more understandable, especially when dealing with a small number of features.
    - In summary, our decision was based on considerations specific to our situation. While Random Forest is a powerful algorithm, SVM can provide a simpler and more efficient

solution, especially when dealing with a small number of features. However, the best approach remains testing different models on your specific dataset and making decisions based on their actual performance.