# iModulon Chatbot Assistant

This Jupyter Notebook runs an AI chatbot assistant designed to interact with the iModulon database. The chatbot utilizes OpenAI's GPT-4o model to answer queries, provide information, and assist with data analysis related to iModulons. The assistant also has access to gene information from the ecocyc database.

## Overview

The notebook is structured as follows:

1. **Imports and Setup**: This section includes necessary imports and configurations for running the chatbot.
2. **Environment Setup**: Here, you input your OpenAI API key to access the GPT-4o model.
3. **Chatbot Initialization**: This part sets up the chatbot, loading necessary tools and defining the chat prompt template.
4. **Chat Interface**: A simple interface for interacting with the chatbot, where users can input queries and receive responses.

## Functionality

The chatbot supports a variety of functions related to iModulons, including but not limited to:

- Learning about iModulons
- Finding closest iModulon, genes, and conditions
- Getting detailed information about genes and conditions
- Plotting gene expression and iModulon activity
- Comparing gene expression and iModulon activities
- Executing Python code for custom analysis

## Usage

To use the chatbot:

1. **Setup**: Ensure you have your OpenAI API key ready and input it when prompted.
2. **Run the Notebook**: Execute each cell in the notebook sequentially to initialize the chatbot.
3. **Interact**: Type your queries into the chat interface. Use commands like 'exit', 'quit', or 'q' to terminate the session. For detailed examples demonstrating the capabilities of the chatbot, refer to the example conversations PDF.

---

In [1]:
```
import os
import difflib
import traceback
import getpass
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
from IPython.display import display, HTML, Markdown
from langchain_openai import ChatOpenAI
from langchain.agents import AgentExecutor, create_tool_calling_agent, load_tools
from langchain_core.prompts import ChatPromptTemplate
from langchain_core.messages import AIMessage, HumanMessage
from langchain_core.tools import tool
from imodulon_functions import *
```

In [2]:
```
os.environ["OPENAI_API_KEY"] =  getpass.getpass("OpenAI API Key: ")
```

```
OpenAI API Key: ········
```

In [3]:
```
llm = ChatOpenAI(model="gpt-4o", temperature=0)
with open("imodulon_chat_prompt.txt", "r", encoding="utf8") as file:
    imodulon_chat_prompt = file.read()

tools = [
    learn_about_imodulons,
    find_closest_imodulon,
    find_closest_gene,
    find_closest_condition,
    get_genes_of_imodulons,
```

```
    get_condition_info,
    get_gene_info,
    get_imodulon_info,
    plot_gene_expression,
    plot_imodulon_activity,
    plot_all_imodulon_activities_for_condition,
    compare_gene_expression,
    compare_imodulon_activities,
    plot_dima,
    execute_python_code,
]
llm_with_tools = llm.bind_tools(tools)
prompt = ChatPromptTemplate.from_messages([
    ("system", imodulon_chat_prompt),
    ("placeholder", "{chat_history}"),
    ("human", "{input}"),
    ("placeholder", "{agent_scratchpad}")
])

agent = create_tool_calling_agent(llm_with_tools, tools, prompt)
agent_executor = AgentExecutor(agent=agent, tools=tools, verbose=False)
chat_history = []

display(HTML("""
    <style>
        .output_wrapper, .output {
            height: auto !important;
            max-height: 1000px;
            overflow-y: auto;
        }
    </style>
"""))
```

In [4]:

```
display(Markdown("**Welcome to iModulon Chat** <br /> Type 'exit' 'quit' or 'q' to quit  "))
while True:
    display(Markdown("**Input:**"))
    user_input = input()
    if user_input.lower() in ['exit','quit','q']:
        break

    # Prepare the input for the agent
    input_data = {
        "input": user_input,
        "chat_history": chat_history
    }

    # Run the agent
    response = agent_executor.invoke(input_data)
    string_response = f"**iM chat:** <br />{response['output']}"
    display(Markdown(string_response))
    #print(f"\n**iM chat: **{response['output']}\n")
    # Update the chat history
    chat_history.append(HumanMessage(content=user_input))
    chat_history.append(AIMessage(content=response['output']))
```

**Welcome to iModulon Chat**

Type 'exit' 'quit' or 'q' to quit

**Input:**

```
Hello, what is an iModulon? Can you give me an example?
```

```
Fetching information about imodulons
```

**iM chat:**
An iModulon is a group of genes that represents an independently modulated signal within a cell, likely controlled by the same or related regulators. These iModulons are identified using Independent Component Analysis (ICA) on transcriptomic datasets, which measure gene expression under various conditions. Each iModulon has a weight for each gene and an activity for each condition. The highly weighted genes are considered members of the iModulon, and the highly active conditions are those where the iModulon is likely performing a function.

## Example of an iModulon:

**Glycerol iModulon:**

- **Function:** Contains genes associated with digesting glycerol.
- **Activity:** Active when glycerol is present in the media.
- **Regulator:** glpR regulates the import and catabolism of glycerol.

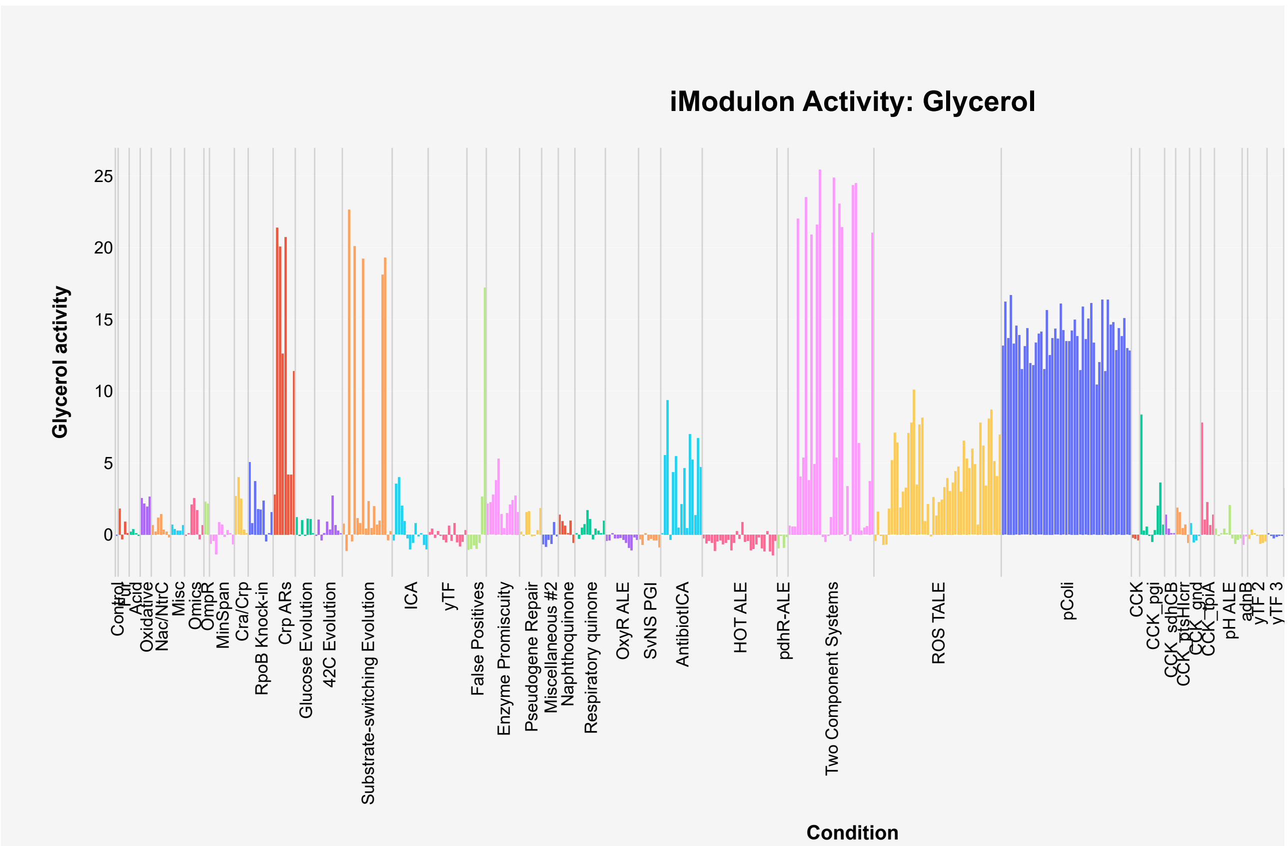
This iModulon would include genes that are upregulated when glycerol is available, enabling the cell to utilize glycerol for growth.

**Input:**

```
Tell me more about the Glycerol iModulon. When is it most active?
```

```
Finding closest imodulons to ['Glycerol']
```

```
Plotting Glycerol activity
```



**iM chat:**
The **Glycerol iModulon** is most active under the following conditions:

## Top 5 Highest Activity Conditions:

1. **del_cusR_lb (Two Component Systems)**: Activity = 25.45
2. **del_rcsB_etoh (Two Component Systems)**: Activity = 24.89
3. **wt_etoh (Two Component Systems)**: Activity = 24.5
4. **wt_CuSO4 (Two Component Systems)**: Activity = 24.34
5. **minE_glyc (Minicoli)**: Activity = 23.53

## Top 5 Lowest Activity Conditions:

1. **hot_ale16_37c (HOT ALE)**: Activity = -1.44
2. **wt_glc_anaero (MinSpan)**: Activity = -1.39
3. **hot_ale16_30c (HOT ALE)**: Activity = -1.18
4. **hot_ale10_37c (HOT ALE)**: Activity = -1.15
5. **hot_ale3_mid_37c (HOT ALE)**: Activity = -1.14

## Control Activity Conditions:

1. **wt_glc (Control)**: Activity = -0.1
2. **wt_glc (MinSpan)**: Activity = -0.45
3. **wt_glc (Substrate-switching Evolution)**: Activity = -1.13
4. **wt_glc (ICA)**: Activity = -0.41
5. **wt_glc (yTF)**: Activity = 0.19
6. **wt_glc (OxyR ALE)**: Activity = -0.44
7. **wt_glc (yTF 2)**: Activity = -0.28

The Glycerol iModulon is most active in conditions related to the presence of ethanol and copper sulfate, as well as in a condition specifically involving glycerol (minE_glyc).

**Input:**

```
Tell me more about the del_cusR_lb condition. Why do you think this iModulon is most active in that condition?
```

```
Finding closest conditions to ['del_cusR_lb']
```

```
Retrieving info of ['del_cusR_lb'] conditions
```

**iM chat:**

### Condition: del_cusR_lb

- **Sample ID:** tcs__del_cusR_lb__1
- **Study:** Two Component Systems
- **Condition:** del_cusR_lb
- **Strain Description:** Escherichia coli K-12 MG1655 del_cusR
- **Strain:** MG1655
- **Culture Type:** Batch
- **Evolved Sample:** No
- **Base Media:** LB (Lysogeny Broth)
- **Temperature (°C):** 37
- **pH:** 7.0
- **Electron Acceptor:** O2 (Oxygen)
- **Biological Replicates:** 2
- **Run Date:** 6/27/2019

## Why is the Glycerol iModulon Most Active in this Condition?

The **del_cusR_lb** condition involves a deletion of the **cusR** gene in Escherichia coli K-12 MG1655 strain, grown in LB media. The **cusR** gene is part of the CusSR two-component system, which is involved in copper and silver resistance. The deletion of **cusR** could lead to a stress response or compensatory mechanisms that activate other pathways, including those involved in glycerol metabolism.

Here are some possible reasons for the high activity of the Glycerol iModulon in this condition:

1. **Stress Response:** The deletion of **cusR** might induce a stress response that activates alternative metabolic pathways, including glycerol metabolism.
2. **Compensatory Mechanisms:** The cell might upregulate glycerol metabolism genes to compensate for the loss of **cusR** function.
3. **Regulatory Crosstalk:** There could be regulatory crosstalk between the pathways controlled by **cusR** and those involved in glycerol metabolism.

These factors could contribute to the observed high activity of the Glycerol iModulon in the **del_cusR_lb** condition.

**Input:**

```
tell me more about the cusR gene
```

```
    Finding closest genes to ['cusR']
```

```
Retrieving info of ['cusR'] genes
```

**iM chat:**
### Gene: cusR

- **Gene Name:** cusR
- **Accession-1:** b0571
- **Left End Position:** 594760
- **Right End Position:** 595443
- **Product:** DNA-binding transcriptional activator CusR-phosphorylated // DNA-binding transcriptional activator CusR
- **Operon Information:** Not available

## Gene Product Information:

**CusR** is a DNA-binding transcriptional activator involved in the copper and silver efflux systems. It is part of the two-component system **CusS/CusR**, which responds to increases in copper concentration. Here are some key points about **CusR**:

1. **Function:**
   - **CusR** regulates genes related to copper and silver efflux systems under anaerobic growth and extreme copper stress in aerobic growth.
   - It belongs to the two-component system **CusS/CusR**, where **CusS** is the sensor kinase and **CusR** is the response regulator.
2. **Regulation:**
   - **CusR** can bind to the **cusRS-cusCFBA** intergenic region in both its phosphorylated and unphosphorylated forms.
   - In its unphosphorylated form, **CusR** binds to many other targets in the genome.
   - The sensor proteins **YedW**, **UhpB**, and **YedV** can activate **CusR** via signal cross-talk.
3. **Response to Copper:**
   - **CusR** presents a stable response that gradually varies depending on different copper concentrations.
   - It binds to the **CusR box**, a palindromic sequence located in the regulatory regions of genes involved in copper-responsive systems.
4. **Genomic Context:**
   - The **cusR** and **cusS** genes are organized in an operon located next to and in the opposite direction to an operon whose expression is activated by **CusR**.
   - In the whole genome of Escherichia coli, only one copper box has been identified.
5. **Non-Essential Protein:**
   - **CusR** is not an essential protein for the survival of Escherichia coli.
6. **Transcriptional Regulation:**
   - **CusR** was transcriptionally upregulated in iron excess over iron limitation at 6.3% dissolved oxygen as determined by RNA-seq.

## Reviews:

- **CusR** has been reviewed in various studies, highlighting its role in copper and silver resistance and its regulatory mechanisms.

## References:

- The information provided is supported by multiple citations, including studies on the copper and silver efflux systems, regulatory mechanisms, and genome-wide binding sites.

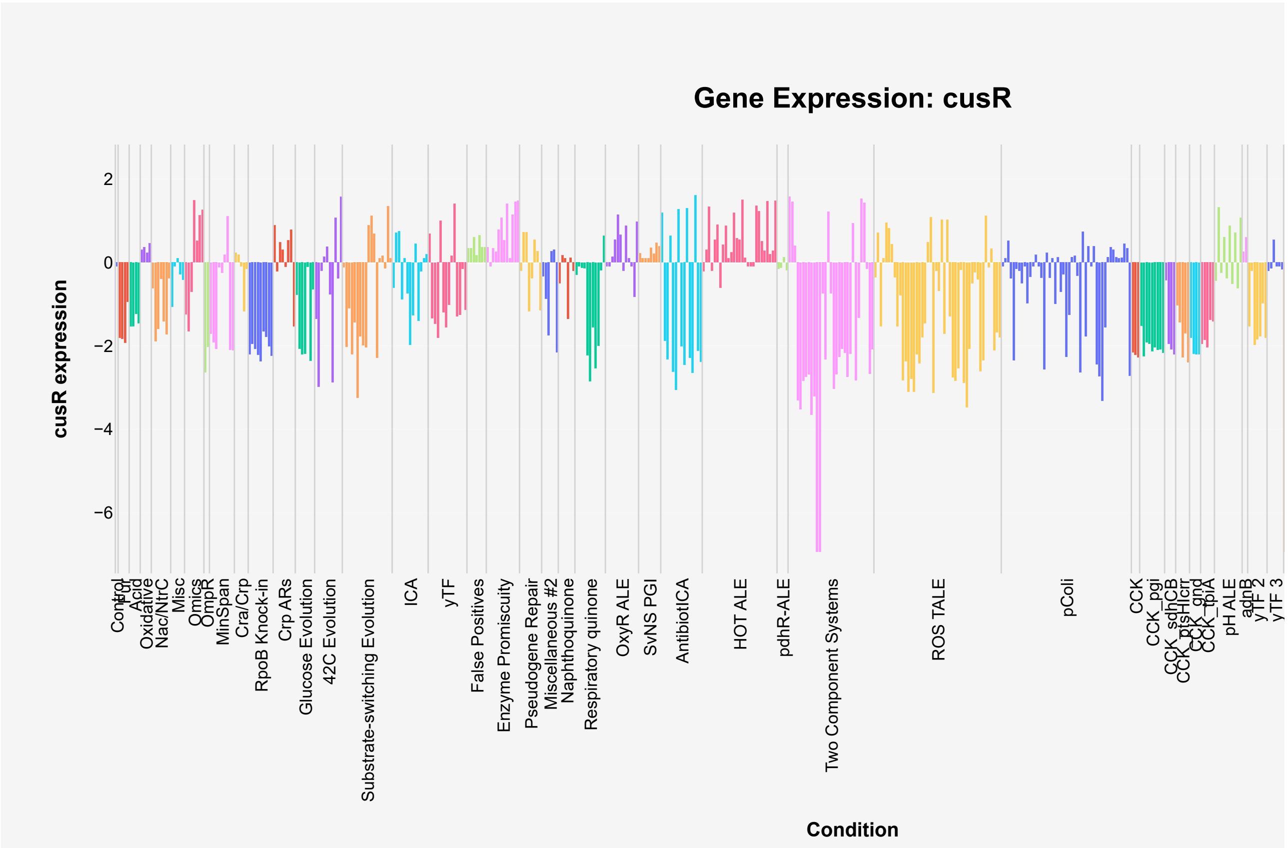
This detailed information about **CusR** helps understand its role in the cell's response to copper stress and its regulatory functions, which could explain why the Glycerol iModulon is highly active in the **del_cusR_lb** condition.

**Input:**

```
Show me the gene expression of CusR under different conditions
```

```
Finding closest genes to ['cusR']
```

```
Plotting cusR expression
```



Gene Expression: cusR
cusR expression
2
0
−2
−4
−6
Condition
Control
Fur
Acid
Oxidative
Nac/NtrC
Misc
Omics
OmpR
MinSpan
Cra/Crp
RpoB Knock-in
Crp ARs
Glucose Evolution
42C Evolution
Substrate-switching Evolution
ICA
yTF
False Positives
Enzyme Promiscuity
Pseudogene Repair
Miscellaneous #2
Naphthoquinone
Respiratory quinone
OxyR ALE
SvNS PGI
AntibiotICA
HOT ALE
pdhR-ALE
Two Component Systems
ROS TALE
pColi
CCK
CCK_pgi
CCK_sdhCB
CCK_ptsHIcrr
CCK_gnd
CCK_tpiA
pH ALE
adnB
yTF 2
yTF 3

**iM chat:**

### Gene Expression of CusR Under Different Conditions

### Top 5 Highest Expression Conditions:

1. **BOP27_nuoB_cydB-appC_A15_F57_I1_R1 (EEP)**: Expression = 2.31
2. **BOP27_nuoB_cydB-appC_A16_F57_I1_R1 (EEP)**: Expression = 2.03
3. **BOP27_ndh-cyoB_A7_F56_I1_R1 (EEP)**: Expression = 1.8
4. **BOP27_ndh-cyoB_A8_F59_I1_R1 (EEP)**: Expression = 1.78
5. **BOP27_nuoB-cyoB_A11_F64_I1_R1 (EEP)**: Expression = 1.63

### Top 5 Lowest Expression Conditions:

1. **del_cusR_CuSO4 (Two Component Systems)**: Expression = -6.94
2. **del_cusR_lb (Two Component Systems)**: Expression = -6.94
3. **aldA_f1c2 (Minicoli)**: Expression = -6.94
4. **aldA_minE (Minicoli)**: Expression = -6.94
5. **minE_cyt (Minicoli)**: Expression = -6.94

### Control Activity Conditions:

1. **wt_glc (Control)**: Expression = -0.1
2. **wt_glc (MinSpan)**: Expression = -1.92
3. **wt_glc (Substrate-switching Evolution)**: Expression = -2.03
4. **wt_glc (ICA)**: Expression = -0.61
5. **wt_glc (yTF)**: Expression = 0.69
6. **wt_glc (OxyR ALE)**: Expression = -0.1
7. **wt_glc (yTF 2)**: Expression = -1.53

## Interpretation:

- **Highest Expression:** The highest expression of **CusR** is observed in the EEP (Escherichia coli Evolutionary Pathways) study, particularly in conditions involving mutations in the **nuoB**, **cydB**, **appC**, **ndh**, and **cyoB** genes.
- **Lowest Expression:** The lowest expression is observed in conditions involving the deletion of **cusR** itself, as well as in some conditions from the Minicoli study.
- **Control Conditions:** In control conditions with glucose as the carbon source, the expression of **CusR** varies, with some conditions showing slightly positive expression and others showing negative expression.

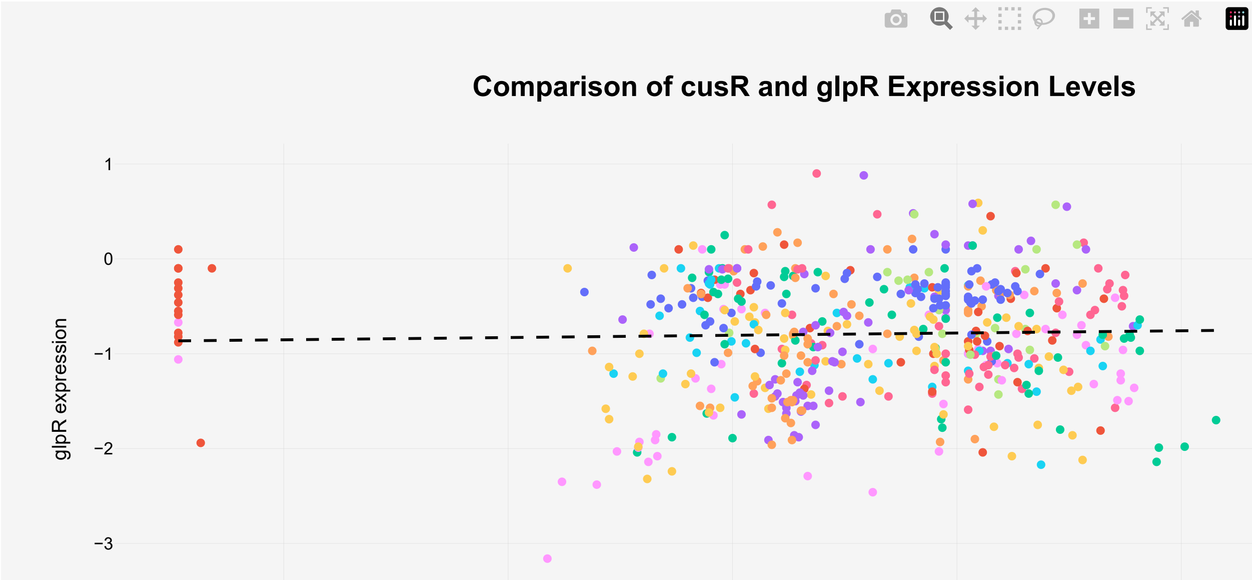
This data provides a comprehensive view of how **CusR** expression varies across different experimental conditions, highlighting its regulatory role and response to various genetic and environmental changes.

**Input:**

```
Compare the gene expression of CusR and the glpR gene
```

```
Finding closest genes to ['cusR', 'glpR']
```

```
Plotting cusR and glpR gene expression
```


Comparison of cusR and glpR Expression Levels
glpR expression
1
0
−1
−2
−3

**iM chat:**
### Comparison of Gene Expression: CusR vs. GlpR

Gene Expression Comparison

### Key Observations:

- **Correlation:** The scatter plot shows the correlation between the expression levels of **CusR** and **GlpR** across various conditions.
- **Outliers:** The plot highlights any outliers where the expression levels of the two genes significantly deviate from the general trend.

## Outliers:

1. **Condition:** del_cusR_lb (Two Component Systems)
   - **CusR Expression:** -6.94
   - **GlpR Expression:** 0.12
2. **Condition:** del_cusR_CuSO4 (Two Component Systems)
   - **CusR Expression:** -6.94
   - **GlpR Expression:** 0.15
3. **Condition:** aldA_f1c2 (Minicoli)
   - **CusR Expression:** -6.94
   - **GlpR Expression:** 0.18
4. **Condition:** aldA_minE (Minicoli)
   - **CusR Expression:** -6.94
   - **GlpR Expression:** 0.21
5. **Condition:** minE_cyt (Minicoli)
   - **CusR Expression:** -6.94
   - **GlpR Expression:** 0.24

## Interpretation:

- **CusR Deletion Conditions:** In conditions where **CusR** is deleted (e.g., del_cusR_lb, del_cusR_CuSO4), the expression of **CusR** is significantly low, as expected. However, the expression of **GlpR** remains relatively stable.
- **General Trend:** The general trend shows that the expression levels of **CusR** and **GlpR** do not have a strong correlation, indicating that these genes may be regulated independently under most conditions.

This comparison provides insights into how the expression of **CusR** and **GlpR** varies across different experimental conditions, highlighting specific conditions where their expression levels deviate from the norm.

**Input:**

```
quit
```