**Updated Analysis of the Manual Query Script: manual\_query.py**

A new script, manual\_query.py, was created to simulate the queries performed by Biofilter. The simulation isolated the GENE **ENSG00000000003** and conducted a query outside the full logic implemented in Biofilter. This process was simplified to focus solely on basic query operations.

**Workflow Overview**

1. **Identify the Biopolymer ID:**
   * Queried the biopolymer\_name table to identify the biopolymer\_id associated with the input gene.
2. **Retrieve Records from group\_biopolymer:**
   * Using the retrieved biopolymer\_id, all associated records in the group\_biopolymer table were queried. While the table contains a source\_id field, no filters were applied other than the biopolymer\_id.
3. **Retrieve Complementary Data from group:**
   * For each group\_id related to the input biopolymer\_id (gene), complementary data from the group table, such as the label and source\_id, were retrieved.

**Results**

* The output is documented in **Analysis\_run\_1.xlsx:ANALYSIS\_SINGLE\_GENE**, which contains **72 records**.
* All records are directly retrieved from group\_biopolymer and complemented with fields from the group table.

**Observations**

Based on the structure of the tables:

1. **Significance of source\_id in group\_biopolymer:**
   * The source\_id in group\_biopolymer indicates the linkage between group and biopolymer and is essential for defining the origin of this relationship.
2. **source\_id in the group Table:**
   * The source\_id in group reflects the origin of the label in the group table. However, it does not confirm that the same origin represents a valid linkage between the group and the biopolymer.

**Key Findings for the GENE ENSG00000000003**

1. **source\_id in group\_biopolymer:**
   * For the 72 linkages retrieved from group\_biopolymer, the source\_id is consistently 0.
2. **source\_id in the group Table:**
   * When the corresponding group\_id is queried in the group table, the source\_id values retrieved are 1, 3, 8, 9, and 10.
3. **Results via Biofilter:**
   * Using Biofilter's filtering logic with a source\_id filter for 3, 5, and 7, the query returned **6 records**.
   * These records originate from the group table with a source\_id of 3. However, the same group\_id entries have a source\_id of 0 in group\_biopolymer.

**Analysis and Possible Alternative Interpretation**

* **Issue with Biofilter Logic:**
  + The results returned by Biofilter are misleading. While the records are filtered based on the source\_id in the group table (source\_id = 3), this field reflects the origin of the group label, not the relationship between the group and the biopolymer.
  + This creates a false implication that the retrieved records represent valid group-to-biopolymer linkages originating from the specified source (3, 5, or 7).
* **Alternative Interpretation:**
  + Another possibility is that the group\_biopolymer table is not intended to track the source of the linkage, and this responsibility lies with the group table. This interpretation is supported by the observation that group\_biopolymer only contains source\_id values of 0 and 14.
  + If this interpretation is correct, the origin control resides in the group table.

**Next Steps and Future Improvements**

* **Continuation of Analysis:**
  + Based on the premise that the origin control resides in the group table, I will proceed with further analysis to validate this structure and ensure alignment with Biofilter's query logic.
* **Future Considerations for Biofilter and LOKI:**
  + In future versions of Biofilter and LOKI, this topic can be revisited to improve the structure of these tables. The current design causes confusion and consumes valuable time to infer the scenario of origin control.

**Conclusion**

The filtering logic in Biofilter needs to be revisited to address this inconsistency. Specifically:

1. If the origin control resides in the group\_biopolymer table, then the filtering logic must use the source\_id field from group\_biopolymer to ensure accurate relationships between group and biopolymer.
2. If the group table is responsible for origin control, this must be explicitly documented, and the filtering logic should align with this design.

Failing to address this issue will result in systematically incorrect results for all queries that use source\_id as a filtering argument in Biofilter. Revisiting the table design will also enhance maintainability and reduce ambiguity in future analyses.