Hypothesis Agent

How did I go about building this

As this is a new domain for me, I started with looking to understand small building blocks of data, what they mean and how they play together. I wanted to build the “skills” of the system before assembling an agent because I had built agents before but I hadn’t worked with this data. This is why you may find this agent architecture perhaps more complicated than what could have been built for this particular purpose. I wanted to build a sketch and then flesh out, polish, trim as necessary.

Examples

---

you: Explain the impact of the MLH1 gene on colorectal cancer risk.

ha:

Hypothesis

The MLH1 gene is likely to have a significant impact on colorectal cancer risk due to its role in DNA mismatch repair and its interactions with other genes and pathways implicated in colorectal cancer. This hypothesis is based on the following analysis results:

1. **KEGG Pathway Information:**
   The initial KEGG query did not return a complete pathway for colorectal cancer. However, the specific query targeting the MLH1 gene within the colorectal cancer pathway (hsa05210) indicated the absence of detailed interaction types within this pathway.

   ```
   {
     "gene_name": "MLH1",
     "pathway_id": "hsa05210",
     "relationships": {
       "activation": [],
       "inhibition": [],
       "binding_or_association": [],
   ```

```
    "expression": [],
    "phosphorylation": [],
    "indirect_effect": [],
    "missing_interaction": []
  }
}
```

2. **GO Terms and Annotations:**
   Several GO terms and annotations for MLH1 related to its function in DNA mismatch repair and protein binding activities were retrieved. Significant annotations include:

   - GO:0006298 (mismatch repair)
   - GO:0003697 (single-stranded DNA binding)
   - GO:0005515 (protein binding)

   This suggests MLH1's primary role involves critical processes related to genome stability.

3. **Network Analysis:**
   The network analysis revealed a forest subarea ratio of 0.014705882352941176 for MLH1 in the context of the colorectal cancer pathway. This metric indicates that MLH1 has a moderate potential impact on the colorectal cancer pathway.

   Additional key genes and pathways connected to MLH1 were not explicitly identified in this analysis, suggesting the necessity of further investigation, possibly including cross-referencing other pathway and gene interaction databases.

4. **Inquiry into Related Genes and Pathways:**
   The analysis places emphasis on gathering comprehensive data, recommending cross-referencing with other databases and literature. This step would likely reveal more about MLH1's interactions and regulatory mechanisms involving genes such as MSH2, MSH6, and PMS2—genes known to function alongside MLH1 in mismatch repair and associated with increased colorectal cancer risk when mutated.

5. **Forest Subarea Ratio Insight:**

   - The forest subarea ratio was utilized to infer the influence of MLH1 on colorectal cancer pathways. Although moderate, the direct influence suggests potential cascades impacting gene regulation and cancer progression.

Conclusion

The MLH1 gene likely plays a pivotal role in colorectal cancer risk due to its involvement in DNA mismatch repair and its interaction with other cancer-related pathways. Despite the initial lack of direct pathway data from KEGG, GO term analysis and network metrics support the hypothesis of MLH1's significant influence on colorectal cancer. Further investigation and cross-referencing additional databases and literature are recommended to fully elucidate MLH1's impact and network of interactions in colorectal cancer.

---

me: Tell me about INSR gene and what it impacts in the diabetes pathway

ha: I have a plan. Here's a peek:

Task 1:
Objective: Identify the exact name of the diabetes pathway in the KEGG database.
Tool: kegg_query

Task 2:
Objective: Retrieve information about the role of the INSR gene in the diabetes pathway from the KEGG database.
Tool: kegg_query

Task 3:
Objective: Identify the GO terms and annotations related to the INSR gene in the context of diabetes using the GAF database.
Tool: gaf_query

Task 4:
Objective: Perform a network analysis on the INSR gene within the diabetes pathway.
Tool: graph_analysis

ha:

Hypothesis:

The **INSR (Insulin Receptor) gene** plays a crucial role in the **Type II diabetes mellitus** pathway by facilitating insulin signaling. Based on pathway and gene interaction analysis from KEGG and GAF databases, the following can be concluded:

1. **INSR Gene Role in Diabetes Pathway**:
   - **Exact Pathway Name**: Type II diabetes mellitus (from KEGG).
   - The INSR gene is pivotal in the insulin signaling pathway, which is dysregulated in Type II diabetes mellitus.
2. **Gene-Protein Interactions**:
   - **Directly Impacted Nodes**: INSR directly influences the genes **MAPK1** and **IRS1** within the pathway.
   - **Relationships Identified**:
     - **Activation**: INSR activates pathways leading to the phosphorylation of other proteins.
     - **Inhibition**: No direct inhibitive interactions identified within the dataset for INSR.
     - **Binding or Association**: INSR is known to bind insulin, initiating a cascade of downstream effects.
     - **Expression**: Changes in INSR expression can significantly impact the pathway.
     - **Phosphorylation**: INSR’s role frequently involves phosphorylation events.
3. **Pathway Information**:
   - **Pathway ID and Title**: Specific ID not provided, but it is categorically under Type II diabetes mellitus.
4. **Network Analysis**:
   - **Network Position**:
     - **Forest Subarea Ratio**: 0.0303 indicating a moderate influence within the pathway.
     - **Direct Connections**: Key nodes directly connected to INSR include **MAPK1** and **IRS1**.
     - **Influence Score**: Indicated through network positions and connections indicating INSR's critical role in insulin signaling.

5. **GO Terms and Annotations**:
    - The GAF query failed to return relevant GO terms specific to the INSR gene in diabetes, suggesting a potential gap in available annotations or a need for specific GO term mapping resources.

Additional Notes

- The pathway's structure highlights INSR's critical role in mediating downstream signaling events crucial for glucose uptake and metabolism.
- The retrieval and analysis steps should focus on filling in data gaps, particularly ensuring accurate and complete identification of pathway IDs and GO terms related to INSR.

This hypothesis underscores the need for meticulous data extraction and suggests that while INSR is a significant player in Type II diabetes mellitus, further detailed data acquisition (correct pathway identifiers, comprehensive gene role details) is essential for deeper insights.

---

me: Explain the impact of the MLH1 gene on colorectal cancer risk.

ha: I have a plan. Here's a peek:

Task 1:
Objective: Identify the colon cancer pathway in the KEGG database.
Tool: kegg_query

Task 2:
Objective: Retrieve MLH1 gene interactions within the colon cancer pathway from KEGG.
Tool: kegg_query

Task 3:
Objective: Search for GO annotations related to the MLH1 gene in the context of colorectal cancer in the GAF database.
Tool: gaf_query

Task 4:
Objective: Perform network analysis to determine how MLH1 interacts with other genes in the colon cancer pathway.
Tool: graph_analysis

ha:

Hypothesis

The MLH1 gene plays a pivotal role in the regulation of critical pathways involved in colorectal cancer. Its interactions, or the lack thereof, with various genes and its specific functional annotations suggest it may have a significant impact on the risk and progression of colorectal cancer. However, current data reveals limited direct interactions with other genes within the KEGG colorectal cancer pathway.

Supporting Evidence

1. **KEGG Pathway Analysis**:
    - **Empty Results for General Colon Cancer Pathway**: An initial query using 'colon cancer' yielded no results, indicating that MLH1's interactions are not broadly documented under the general term in KEGG.
    - **Specific MLH1 Interactions**: Targeted queries showed MLH1 involvement in the KEGG colorectal cancer pathway, but no direct interactions were documented with other genes.
    - **Key Pathway Association**: MLH1 is associated with the KEGG pathway ID **hsa05210** for colorectal cancer.
2. **Gene Ontology (GO) Annotations from GAF Database**:
    - **Functional Roles**: MLH1 annotations include:
        - **DNA mismatch repair** ("acts_upstream_of_or_within" GO_ID: GO:0006298, various PMIDs).
        - **Binding functions**:
            - **Protein binding** (GO_ID: GO:0005515, various PMIDs).
            - **Single-stranded DNA binding** (GO_ID: GO:0003697, PMID: 11809883).
            - **Molecular functions in mismatch repair**
            - MLH1 interacts with multiple proteins directly involved in its repair mechanisms, such as P43246, P52701, P54277, P54278, Q9UHC1, Q9UQ84.
3. **Network Analysis Findings**:

- **Forest Subarea Ratio**: The calculated forest subarea ratio for MLH1 within the colorectal cancer pathway is **0.0147**. While this reflects a specific impact footprint, the actual interactions identified within the pathway are sparse.
- **Absence of Directly Impactful Nodes**: Despite a thorough search, no other nodes (genes) were directly impacted by MLH1 in terms of significant interactions.

Interpretation

The data suggests that while MLH1 is a crucial gene in colorectal cancer, particularly through its role in DNA mismatch repair, there are significant gaps in direct interaction data within the specific colorectal cancer pathway in KEGG. This absence of interaction data could be due to several factors, including incomplete annotations in the database or the highly localized and specialized role of MLH1 that involves indirect interactions through other molecular pathways not fully captured in the KEGG database.

Next Steps

1. **Expand Database Search**: Investigate other databases or more comprehensive data sources that might contain interaction data not recorded in KEGG.
2. **Experimental Validation**: Conduct laboratory experiments to validate the functional impact and interaction of MLH1 with other genes in colorectal cancer.
3. **Integrative Analysis**: Use multi-omic approaches to explore interactions, combining genomic, transcriptomic, proteomic, and epigenetic data to obtain a holistic view of MLH1's role in colorectal cancer.

In conclusion, while the MLH1 gene appears central to colorectal cancer mechanisms, the precise manner of its interactions remains under-documented in KEGG. This elucidates a potential crucial role of MLH1 in disease risk that warrants further investigation.

Things that didn't go well

As stated before I tried a few things that probably were bad ideas. Here's some of it.

Additional Terminal Nodes

The PNG files with the pathways have very useful text nodes that show you the high level idea of what's going on at the end or beginning of a particular pathway. I found it pretty unfortunate that the KEGG data didn't have those as they would have served as nice anchors for mapping to real queries which would be more likely to connect with a phrase like `insulin resistance` for example. I thought I could enrich the data by passing the PNG to the LLM but sadly both GPT-4o and Claude 3.5 have severe limitations in spatial reasoning, to the point that they can't tell an outlined text from a plain text. There's some code for this but overall this didn't work and there were quite a few problems with it.

The centrality metric

I chose Neo4j because I wanted to have a learning experience where I actually analyse the graph with the database. I find the queries OK but they seem very slow. If I have to do this again either I need to dig deeper into how the engine works or I should just do this (all metrics) with NetworkX.

I wanted to use Katz centrality in general but that didn't exist in the Neo4j libraries and I couldn't get the HITS one to work either, so I went with a crude metric I implemented which compares the area of the subtree of a chosen node to the area of the forest of a pathway. Not sure it's very useful but it is an indication.

Perhaps I should have let the AI be more creative but judging by how it did things — it probably will need quite a bit of help for things like that.

Brittle system

I did get the system to be less brittle but overall I'd say it's still quite brittle and there are possibly quite a few places where an exception would bring it down.

Limited testing

I've implemented sanity check tests for most functionality but didn't have the time to do a proper suite for the actual core of the agent — too late :'(

Missed opportunities for abstraction

You will see some missed opportunities for abstracting away some methods in a super class or merging a few methods from subclasses into a superclass method with a good set of parameters — think all the generative methods 🙂

Given a bit more time, I'd put more work into making those more general.

Agent limitations

- For now I see pretty simple patterns of 3-4 item plans
- The graph analysis fairs (or used to fail) quite often; maybe not so much with the current fix
- There is essentially no freeform graph analysis and I had to push quite hard to get the agent to do simple things like get the right pathway title from the database before making queries.
- Reflection works quite well actually
- Neo4j queries are quite strange sometimes — very verbose and overcomplex
- I have not seen the agent trying to break down a task further — but that would require an extra too that I have not made. Would be great to have a plan re-worker.