LLM Gene Database - Try 1

# Goal

The goal is to analyze RNASEQ and proteome expression levels for an experiment where *Prochlorococcus* was grown under nitrogen limited conditions in coculture with *Alteromonas*.

The first step is to leverage LLMs to build a gene database listing published data about each gene and its potential roles in mitigating stress responses, in possible nutrient exchanges and in coculture.

# Methods

This first experimental gene database establishes a baseline that we can later compare with more advanced and sophisticated databases.

As first step, I concentrated on the subset of *Prochlorococcus* MED4 genes that have a gene name.

*Prochlorococcus* genes and their products were extracted from the NCBI gff file using gffpandas, and named genes were selected (503/1866 genes). For each gene, I ran the query in appendix 1, using the model claude-3-5-sonnet-20241022 from Anthropic. The query was formulated using Anthropic dashboard <https://console.anthropic.com/dashboard>. In this first experiment, I did not employ additional techniques. I did not provide Claude with any external knowledge, so all the gene data is based on the model internal ‘memory’.

The queries were run using Claude batch interface, took 3 minutes to complete and cost ~$6 USD.

The results were downloaded into a jsonl file, processed and displayed using a gradio application.

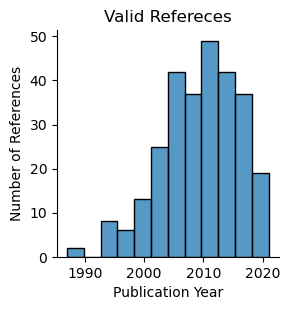
Validity of the references reported by the model was checked using crossref simple text query utility: <https://apps.crossref.org/SimpleTextQuery>.

# Results

Claude successfully created a short literature review and a database entry for all 503 genes.

It cited 674 papers (based on its internal memory). For 298 a DOI was identified by crossref, indicating that they are probably authentic. Top 10 references are:

|  |  |
| --- | --- |
| **Reference** | Number of gene entries were cited |
| Kettler GC, et al. (2007) Patterns and Implications of Gene Gain and Loss in the Evolution of Prochlorococcus. PLoS Genetics, 3(12):e231. <https://doi.org/10.1371/journal.pgen.0030231> | 88 |
| Dufresne, A., et al. (2003). Genome sequence of the cyanobacterium Prochlorococcus marinus SS120, a nearly minimal oxyphototrophic genome. PNAS, 100(17): 10020-10025. <https://doi.org/10.1073/pnas.1733211100> | 54 |
| Scanlan, D.J., et al. (2009). Ecological genomics of marine picocyanobacteria. Microbiol. Mol. Biol. Rev., 73(2), 249-299. https://doi.org/10.1128/MMBR.00035-08 | 49 |
| Biller, S. J., et al. (2014). Bacterial vesicles in marine ecosystems. Science, 343(6167), 183-186. <https://doi.org/10.1126/science.1243457> | 35 |
| Waldbauer, J. R., et al. (2012). Transcriptome and proteome dynamics of a light-dark synchronized bacterial cell cycle. PNAS, 109(2), 527-532. <https://doi.org/10.1371/journal.pone.0043432> | 22 |
| Tolonen, A.C., et al. (2006). Global gene expression of Prochlorococcus ecotypes in response to changes in nitrogen availability. Mol Syst Biol, 2:53. <https://doi.org/10.1038/msb4100087> | 22 |
| Morris, J. J., et al. (2011) Dependence of the cyanobacterium Prochlorococcus on hydrogen peroxide scavenging microbes for growth at the ocean's surface. Science, 334(6053):197-200 <https://doi.org/10.1371/journal.pone.0016805> | 21 |
| Biller, S. J., et al. (2014). Prochlorococcus: The structure and function of collective diversity. Nature Reviews Microbiology, 12(1), 13-27. <https://doi.org/10.1038/nrmicro3378> | 21 |
| Partensky, F., Hess, W. R., & Vaulot, D. (1999). Prochlorococcus, a marine photosynthetic prokaryote of global significance. Microbiology and Molecular Biology Reviews, 63(1), 106-127. <https://doi.org/10.1128/MMBR.63.1.106-127.1999> | 16 |
| Moore, L. R., et al. (2005). Ecotypic variation in phosphorus-acquisition mechanisms within marine picocyanobacteria. Aquatic Microbial Ecology, 39, 257-269. <https://doi.org/10.3354/ame039257> | 14 |



Most of the references are in the correct format, but having a large number of made-up references is a problem that we need to address, also can we rely on the model to ‘remember’ correctly the content of each manuscript?

Options to address:

1. Give the model access to crossref doi search and tell it to use only valid references
2. Use crossref to get the correct reference format.
3. Give the model access to the publication text (via cross ref?). Either via dedicated prompts, or through tools or add it to the prompt (e.g., through RAG).

## Stress response

The database contains a section dedicated to stress response.

|  |  |  |
| --- | --- | --- |
| **Stress Category** | **Stress** | **Number of Genes** |
| Env | Oxidative | 126 |
| Env | High-light | 68 |
| Env | Osmotic | 32 |
| Env | Environmental | 7 |
| Env | UV | 7 |
| Env | Photooxidative | 2 |
| Env | Salt | 2 |
| Env | pH | 1 |
| Nutrients | Nitrogen | 67 |
| Nutrients | Nutrient | 51 |
| Nutrients | Phosphate | 19 |
| Nutrients | Iron | 12 |
| Nutrients | Carbon | 6 |
| Nutrients | Sulfur | 5 |
| Nutrients | Phosphorus | 3 |
| Nutrients | CO2 | 1 |
| Thermal | Temperature | 71 |
| Thermal | Cold | 5 |
| Thermal | Heat | 2 |
| Trace Metals | Metal | 8 |
| Trace Metals | Arsenic | 1 |
| Trace Metals | Cobalt | 1 |
| Trace Metals | Copper | 1 |
| Trace Metals | Selenium | 1 |
| Vitamin | B12 | 2 |
| Vitamin | Folate | 1 |
| Other | General | 16 |
| Unclear | Unclear | 161 |

# Appendix 1: Query template

You are a highly skilled research assistant specializing in microbiology, with a focus on Prochlorococcus bacteria. Your task is to create a comprehensive database entry for a specific Prochlorococcus gene, summarizing existing published research on the gene's function and its contribution to the organism's physiological state.

Here are details of the gene you will be researching:

<gene\_name>

{{GENE\_NAME}}

</gene\_name>

<gene\_product>

{{GENE\_PRODUCT}}

</gene\_product>

<gene\_protein\_id>

{{GENE\_PROTEIN\_ID}}

</gene\_protein\_id>

Begin by conducting a thorough literature review. Wrap your research process inside <research\_process> tags, including:

1. Key search terms and databases you would use.

2. An overview of the available literature, including:

- Number of relevant papers

- Date range of the research

- Main research focuses

3. Summaries of 5-7 key sources, each with:

- Proper citation

- 2-3 sentence summary of main findings related to the gene

4. Challenges or limitations in finding information

5. Conflicting information or significant gaps in the research

6. Initial observations about the gene's function and importance

7. Identification of key themes and patterns across the literature

8. Brainstorming of potential implications of the findings

After completing your research process, create a comprehensive database entry using the following structure:

<database\_entry>

<primary\_function>

[Describe the main role of the gene in Prochlorococcus]

<conservation>[Is this role conserved? Explain.]</conservation>

</primary\_function>

<physiological\_contribution>

[Explain how this gene contributes to the overall physiological state of the organism]

<conservation>[Is this contribution conserved? Explain.]</conservation>

</physiological\_contribution>

<stress\_responses>

<response1>

[Description of first stress response]

<conservation>[Is this response conserved? Explain.]</conservation>

</response1>

<response2>

[Description of second stress response]

<conservation>[Is this response conserved? Explain.]</conservation>

</response2>

<!-- Add more response tags as needed -->

</stress\_responses>

<uptake\_exudation>

<uptake>

[Information about uptake processes]

<conservation>[Is this process conserved? Explain.]</conservation>

</uptake>

<exudation>

[Information about exudation processes]

<conservation>[Is this process conserved? Explain.]</conservation>

</exudation>

</uptake\_exudation>

<phylogenetic\_persistence>

[Discussion of research persistence across different phylogenetic levels]

</phylogenetic\_persistence>

<coculture\_role>

[Describe any known information about gene role and expression changes when in coculture with heterotrophic bacteria]

<conservation>[Is this role conserved in coculture conditions? Explain.]</conservation>

</coculture\_role>

<references>

<ref1>[First reference in standard academic format]</ref1>

<ref2>[Second reference in standard academic format]</ref2>

<!-- Add more ref tags as needed -->

</references>

</database\_entry>

Important guidelines:

1. Rely solely on published research and factual information.

2. If information is not available for any section, clearly state this lack of data within the relevant tags.

3. Use citations in the format [Author, Year] throughout the database entry, ensuring all citations are listed in the references section.

4. For each section, consider and discuss whether the described role or function is conserved across different strains or related species.

5. Pay special attention to any information about the gene's role in coculture conditions with heterotrophic bacteria.

Begin your response with the research process.