Detailed Data Requirements for PhageMatch Prototype

Note: The genomic data (complete sequences and annotations) is already available. The following specifications focus on the complementary data needed.

1 Phenotypic Data

Phenotypic data is essential for understanding the real-world behavior of phages and their interactions with bacterial hosts. This data will help train the model to predict the effectiveness of phages against specific bacterial strains.

1.1 Phage Lytic Activity Data

- **Plaque-Forming Units (PFU):** Experimental data on the number of plaques formed by a phage on a bacterial lawn. This indicates the phage's ability to infect and lyse the bacteria.
- **Lysis Time:** The time taken for a phage to lyse a bacterial culture. This is often measured in hours or minutes and is critical for understanding the speed of phage action.
- **Efficiency of Plating (EOP):** A measure of how efficiently a phage can infect and lyse a bacterial strain compared to a reference strain. EOP is typically expressed as a ratio (e.g., 0.1, 1.0, etc.).
- **Burst Size:** The average number of phage particles released from a single infected bacterial cell. This is a key metric for understanding the reproductive potential of a phage.

  Host Range Data:

1.2 Host Range Data

- **Cross-Infection Assays:** Data on the range of bacterial strains that a particular phage can infect. This is typically determined by testing the phage against a panel of bacterial strains and recording which strains are susceptible.
- **Narrow vs. Broad Host Range:** Information on whether a phage has a narrow host range (infects only a few strains) or a broad host range (infects many strains).

1.3 Bacterial Susceptibility Data

- **Minimum Inhibitory Concentration (MIC):** Data on the minimum concentration of a phage required to inhibit bacterial growth. This is often measured in plaque-forming units per milliliter (PFU/mL).
- **Resistance Profiles:** Data on bacterial strains that have developed resistance to specific phages, including the mechanisms of resistance (e.g., mutation in phage receptors).

2 Proteomic Data

Proteomic data provides insights into the molecular interactions between phages and bacteria, particularly the proteins involved in phage attachment and infection.

2.1 Phage Structural Proteins

- **Capsid Proteins:** Data on the proteins that form the phage capsid, including their sequences and structural information.
- **Tail Fibers and Receptor-Binding Proteins:** Data on the proteins responsible for recognizing and binding to bacterial surface receptors. These proteins are critical for determining host specificity.
- **Lysins and Holins:** Data on the proteins involved in breaking down the bacterial cell wall (lysins) and creating pores in the bacterial membrane (holins) during the lytic cycle.

2.2 Bacterial Surface Proteins

- **Receptor Proteins:** Data on bacterial surface proteins that serve as receptors for phage attachment. Examples include outer membrane proteins (OMPs), lipopolysaccharides (LPS), and pili.
- **Antibiotic Resistance Proteins:** Data on bacterial proteins that confer resistance to antibiotics, as these may also influence phage susceptibility.

3 Transcriptomic Data

Transcriptomic data provides insights into how gene expression changes in both bacteria and phages during infection. This data is critical for understanding the dynamic interactions between phages and their hosts.

3.1 Bacterial Gene Expression Data

- **RNA-seq Data:** Transcriptomic data from bacterial strains before and after phage infection. This data should include information on upregulated and downregulated genes, particularly those involved in stress responses, metabolism, and phage defense mechanisms (e.g., CRISPR-Cas systems).

- **Time-Course Data:** RNA-seq data collected at different time points during phage infection to capture the temporal dynamics of gene expression.

3.2 Phage Gene Expression Data

- **RNA-seq Data:** Transcriptomic data from phages during different stages of the infection cycle (e.g., early, middle, and late gene expression). This data should include information on genes involved in DNA replication, structural protein synthesis, and lysis.
- **Temporal Expression Patterns:** Data on how phage gene expression changes over time during the infection process.

4 Clinical and Environmental Metadata

Clinical and environmental metadata provide context for phage-bacteria interactions, helping to refine the model's predictions for real-world applications.

4.1 Clinical Outcomes Data

- **Patient Response Data:** Data on the outcomes of phage therapy in clinical settings, including reduction in bacterial load, patient recovery rates, and any adverse effects.
- **Bacterial Load Measurements:** Quantitative data on bacterial load before and after phage therapy, often measured in colony-forming units (CFU/mL).
- **Treatment Duration:** Data on the duration of phage therapy required to achieve a clinical response.

4.2 Environmental Data

- **Temperature and pH:** Data on the environmental conditions (e.g., temperature, pH) under which phage-bacteria interactions occur. This is particularly important for applications in agriculture and environmental microbiology.
- **Nutrient Availability:** Data on how nutrient availability (e.g., carbon sources, nitrogen sources) affects phage-bacteria interactions.

5 Experimental Validation Data

Experimental validation data is critical for testing the accuracy of the model's predictions.

- **Controlled Lab Experiments:** Data from controlled experiments where phage-bacteria interactions have been experimentally confirmed. This includes data on phage efficacy, host range, and lysis dynamics.
- **Cross-Validation Data:** Data from independent experiments that can be used to validate the model's predictions. This ensures that the model generalizes well to new data.

6 Pre-Trained Models and Embeddings

To reduce the need for extensive training data, we will leverage pre-trained models that have already learned biologically meaningful patterns from large datasets.

- **DNABERT:** A pre-trained model for nucleotide sequences that can be fine-tuned for phage and bacterial genomic data.
- **ProtBERT/ESM:** Pre-trained models for protein sequences that can be used to analyze phage structural proteins and bacterial surface proteins.

7 Conclusion

The datasets listed above are critical for developing the PhageMatch prototype. These datasets will enable us to train and validate machine learning models that can accurately predict phage-bacteria interactions, ultimately optimizing phage therapy for clinical, agricultural, and industrial applications.

If you have access to any of these datasets or can guide us to relevant sources, we would greatly appreciate your assistance. Please let us know if you need any further details or clarification.