Project

Simple Annotation Pipeline
- Ranjit Kumaresan

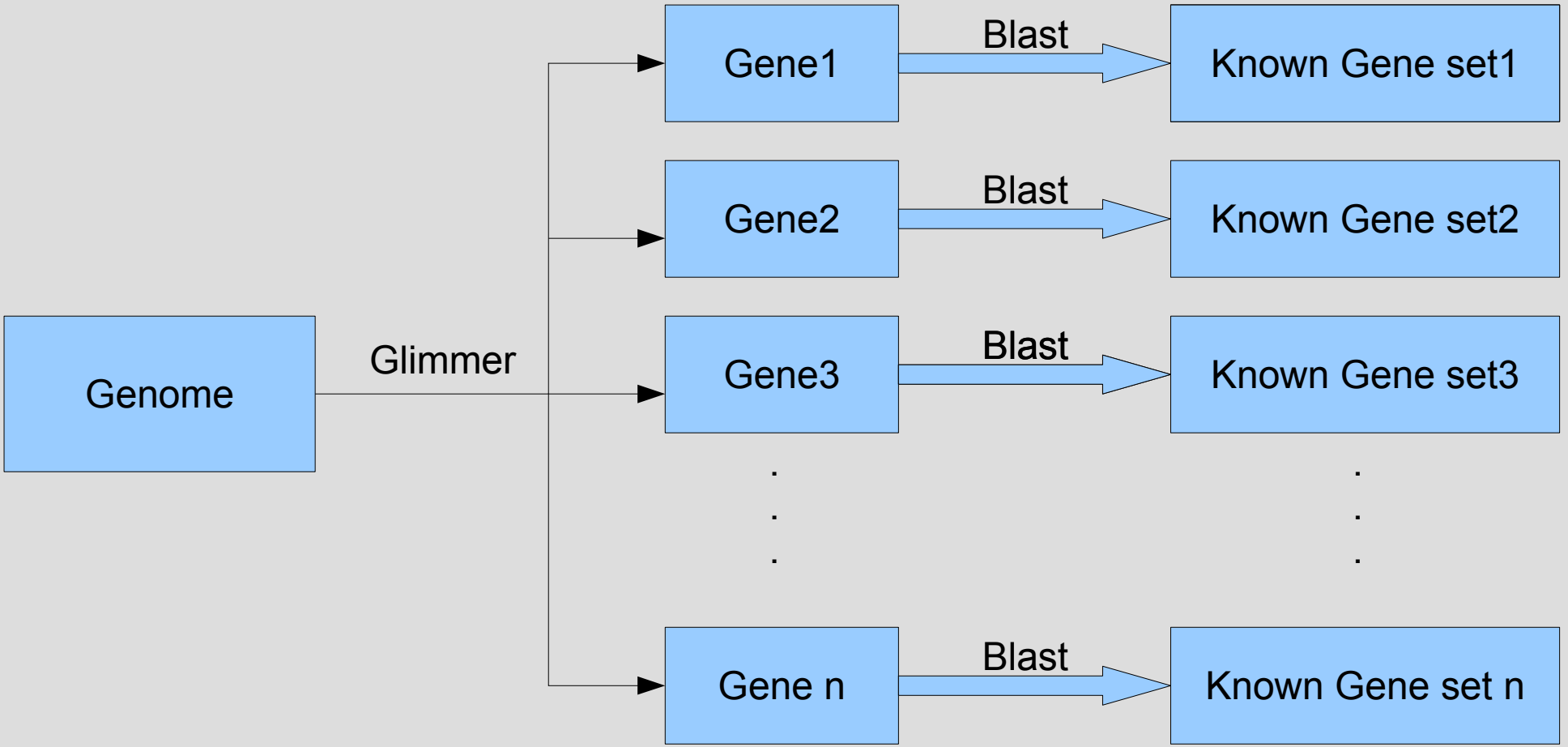
Simple Annotation Pipeline

"Run a gene finder on a genome, identify a set of genes, then compare these against GenBank using blast to determine their probable function. This project will require you to learn how to use the gene finder Glimmer, and how to run batch Blast jobs against GenBank."

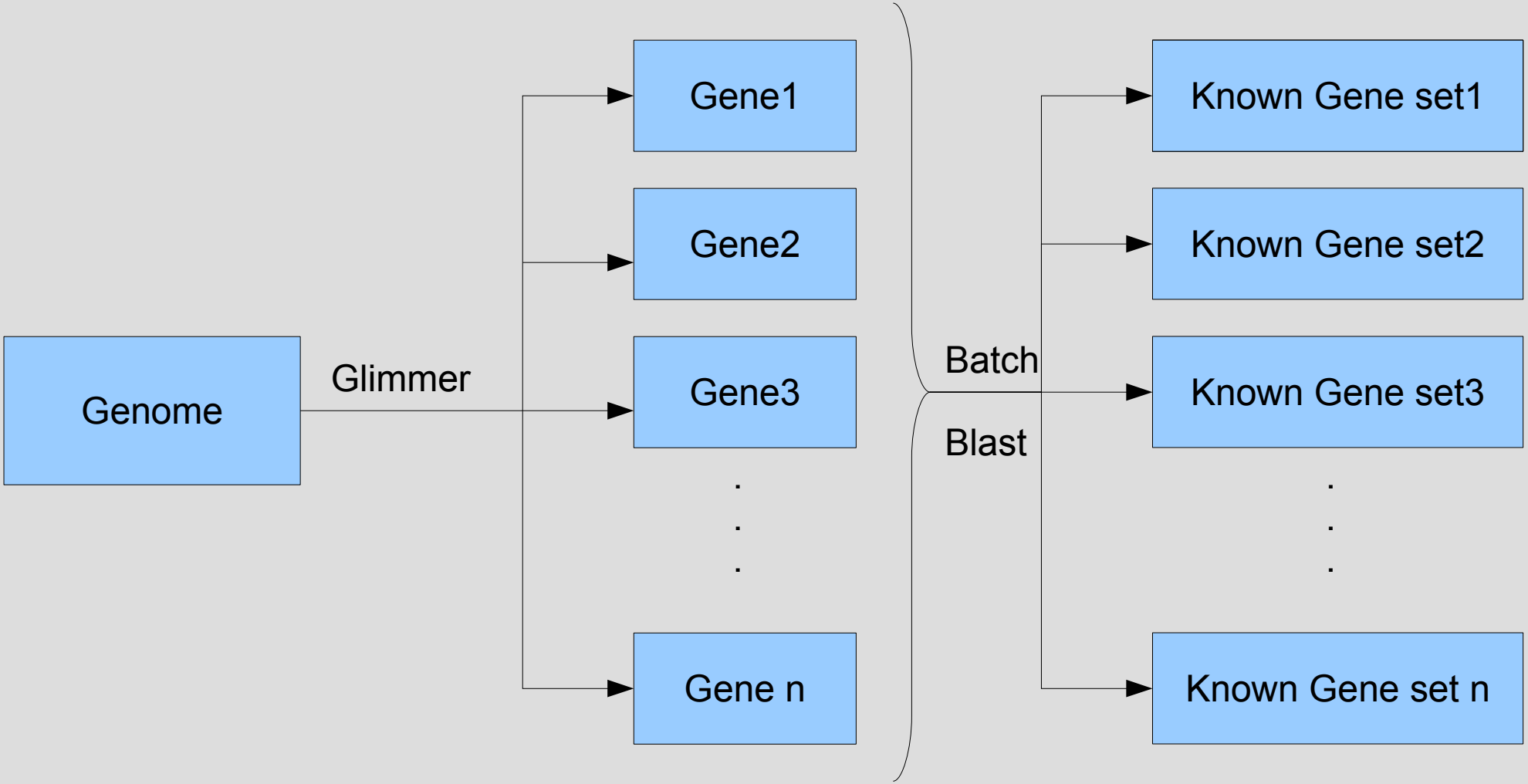
Software to be used

- Glimmer (stands for Gene Locator and Interpolated Markov ModelER) is a bioinformatics ystem for finding genes that uses the interpolated Markov model formalism.
- GenBank (sequence database) is an open access, annotated collection of all publicly available nucleotide sequences and their protein ranslations.
- Blast (Basic Local Alignment Search Tool) is an algorithm for comparing primary biological sequence information, such as the amino-acid sequences of different proteins or the nucleotides of DNA sequences. A *BLAST search* enables a researcher to compare a query sequence with a library or database of sequences, and identify library sequences that resemble the query sequence above a certain threshold.

Normal manual process



What this project does



Relevance

- Blast is a very popular software.
- Most people (biologists) using Blast don't know there is a batch processing option to Blast.
- Also an Annotation pipeline is definitely useful to a biologist

Extensions

- Capabilities of a genomic sequence can be determined using Blast output.
- Example: it is possible to identify what set of proteins are required for digesting food and extracting energy out of it
- Given an unknown genome and a function, whether the genome contains genes that can perform the function.

Challenges

- Making the Annotation pipeline user friendly.
- Blast output is huge.
- Functions of genes is not easily extractable from the Blast output (not defined in all cases).