Problem Description

An important field of research for biologists is the identification of new biological products with novel properties. However with the addition of each component to a chain of biological molecules, the chemical space grows exponentially. (For example considering the relatively limited case of the twenty 'standard' AAs and a peptide of length 4, there are 20^4 possible combinations... Around 160,000!)

An exhaustive search would be impossible, so instead we identify biological products produced in the wild. For an organism to spend biological resources producing some product, we can assume it confers some evolutionary advantage. For example, a strain of bacteria might produce a biological product that kills another competitor species of bacteria (which we can then use to ourselves kill a potentially deadly strain of bacteria). In this way we can use the natural algorithm of evolution as a heuristic to explore the chemical space for us and identify potentially interesting biological products.

We can then try to identify these potentially interesting biological products by, for example, analysis of chemical structure for similarity to other interesting products, or by observing an interesting behaviour across several species which produce a particular product. However we also want to isolate the particular biosynthetic gene cluster (BGC) which produces this novel product - this is a generalisation of peptidogenomics, the matching of peptides to geonomes. By observing species with the same BGC produced this product (clustering), we can generally assume this BGC has a good chance of producing the product in question.

Pep2Path is a piece of software that can probabilistically match peptidic products (a subset of our field of interest) to BGCs. Given the output from a mass spectrometer we can reconstruct composition of a peptide and order of amino acids at consecutive mass-peaks, using a mass-translation table, and given a geonome sequence, we can use tools like antiSMASH and NRPSPredictor to identify potential candidate BGCs in the geonome. Pep2Path then takes these two outputs and calculates the most likely matches of sequence to BGC.

We wish to implement a similar functionality that can be integrated into a local codebase, where we match biological products in general to candidate BGCs, using both the algorithm identified in the original Pep2Path paper and other assorted techniques. This also requires the implementation of a program to extract sequence tags from our local set of mass spectrometry data.