Problem Description

An important field of research for biologists is the identification of new peptidic products with novel properties. However with the addition of each amino acid to a peptide sequence, 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 peptidic products produced in the wild. For an organism to spend biological resources producing some peptide, we can assume it confers some evolutionary advantage. For example, a bacterium might produce a peptidic product that kills another competitor species of bacteria (which we can then use to ourselves kill a potentially deadly species 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 peptidic products.

We can then try to identify these potentially interesting peptidic 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 peptide. However we also want to isolate the particular biosynthetic gene cluster (BGC) which produces this novel peptide - this is 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 peptide in question.

Pep2Path is a software that can probabilistically match peptidic products 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.

However, Pep2Path is not very well-maintained, and its codebase has proven to be largely unmaintainable and inextensible. We wish to implement a new software which can perform some of the useful functionality of the original Pep2Path, while offering more useful features and the potential for future development.