University of Glasgow | School of Computing Science

Honours Individual Project Dissertation

Pep2Path v2

Ross McBride

Abstract

Education Use Consent

Contents

1 Introduction

One particular problem of interest for biologists is the identification of new biological products with novel properties. Until relatively recently, this process was done manually by trained experts. But with the recent advent of the field of bioinformatics, we can now bring immense computational resources to bear on problems like this, saving precious expert time.

Pep2Path (Medema et al. (2014)) is a tool that aims to accelerate drug identification by matching mass-spectrometry outputs to identified biosynthetic gene clusters (henceforth 'BGCs') and in the original paper the authors demonstrate robust benchmarking results. However, despite the utility of Pep2Path, it is implemented as a few monolithic scripts, with little room to revise elements of it in accordance with new developments and research without a complete rewrite. In this paper we present a small suite of software tools implementing a platform for functionality like Pep2Path, and the software process by which this software was designed and created.

1.1 Aims

During this project, we aim to:

- Implement a set of software tools that can achieve similar functionality to the original Pep2Path, subject to correctness testing.
- Achieve a degree of flexibility in the implementation such that this software can be responsively updated to changes in its environment.
- Follow good software practises in the implementation such as loose coupling such that we can achieve software reuse, a cornerstone of software development

1.2 Motivation

Various biological compounds, such as peptides, form assembly-chains of smaller compounds, such as amino acids. If we look at this problem as a computer scientist might, the first thing to come to mind in identifying useful drugs within this space might be an exhaustive search of all combinations of molecule. (Of course, for other molecules questions of topology and structure come into question...) This quite quickly runs into problems, as with the addition of each component to a chain of biological molecules, the chemical space grows exponentially. For example, if we consider the relatively limited case of the twenty proteinogenic amino acids and a peptide of length 4, there are 20^4 possible combinations... Around 160,000! We would then have to identify which, if any, of these have useful properties.

Without some heuristic (for example being able to infer properties from chemical composition), an exhaustive search of the chemical space is impossible. Instead we identify biological products produced in the wild, by living organisms - so-called 'natural products'. From an evolutionary perspective, we expect that if organisms are expending resources on producing a particular biological product, then it confers a survival advantage. For example, a strain of bacteria might produce an antibiotic that kills another competitor species of bacteria, which we can then use ourselves to kill a potentially deadly strain of bacteria.

To identify potentially interesting natural products we could, for example, analyse chemical structure for similarity to other interesting compounds, or by observing similar behaviour across several species which produce a particular product. However, we also want to match this natural product to the BGC that produced it. One way of achieving this is to cluster together species that share a BGC and produce the natural product in question - we can generally assume this pattern is meaningful and the BGC produces the natural product.

Mass spectrometry allows us to identify the chemical composition of compounds - by breaking up molecules and then measuring the mass lost, we can infer the *potential* composition of a compound by the masses of its components using a mass-translation table. Mass spectrometry readings, or amino acid sequences, are one of the key inputs to Pep2Path, which provides two algorithms. NRPS2Path matches potential chemical compositions generated by mass spectrometry analysis to BGC predictions using antiSMASH, (Blin et al. (2017)) whereas RiPP2Path matches these potential chemical compositions to the six-frame translation we can extract directly from the genetic sequence data of the RiPP-producing BGCs. Pep2Path can then automatically score these two sets of data against one another, filling a key part in an automated drug-identification pipeline, where previously investigation of natural products would be based on hand-identification.

However, the original Pep2Path is written almost entirely using singular scripts, one for each algorithm, with some custom data formats and some miscellaneous extra files. As a result, it mixes concerns such as parsing input and generating sequence tags, and at times relies on data structures passed around the entire script. This design makes Pep2Path difficult to update, but there are a number of reasons why it might be advantageous to do so. For one, the scoring method the authors use is based on antiSMASH 2.0 (Blin et al. (2013)) the current version at the time. Since the release of Pep2Path, there have been antiSMASH versions up to 4. (Blin et al. (2017)) While the original Pep2Path scoring method is backwards-compatible, antiSMASH version 4.0 offers the SANDPUMA ensemble algorithm, which collates the Stachelhaus Code and SVM predictors that antiSMASH 2.0 uses along with several others to offer a more accurate BGC prediction.

Furthermore, we may want to investigate product classes other than NRPS and RiPP, the classes of biological product on which Pep2Path operates - there are many others, including polyketides and saccharides. It might also be useful to integrate different algorithms for identifying biological products from sequence data such as the RiPPQuest (Mohimani et al. (2014)) method for lanthipeptides (a subtype of RiPP), decouple the mass spectrometry process from the Pep2Path process or otherwise update the scoring mechanism with new developments. We also expect more generally that having a more transparent software design might lead to easier discovery of bugs or integration with other codebases.

2 Background

3 Analysis/Requirements

4 Design

5 Implementation

6 Evaluation

7 Conclusion

A Appendices

7 Bibliography

K. Blin, M. H. Medema, M. Kazempour, M. A. Fischbach, R. Breitling, E. Takano, and T. Weber. antiSMASH 2.0–a versatile platform for genome mining of secondary metabolite producers. *Oxford Academic*, 2013.

K. Blin, T. Wolf, M. G. Chevrette, X. Lu, C. J. Schwalen, S. A. Kautsar, H. G. S. Duran, E. L. C. de Los Santos, H. U. Kim, M. Nave, J. S. Dickschat, D. A. Mitchell, E. Shelest, R. Breitling, E. Takano, S. Y. Lee, T. Weber, and M. H. Medema. antiSMASH 4.0-improvements in chemistry prediction and gene cluster boundary identification. *Oxford Academic*, 2017.

M. H. Medema, Y. Paalvast, D. D. Nguyen, A. Melnik, P. C. Dorrestein, E. Takano, and R. Breitling. Pep2Path: Automated Mass Spectrometry-Guided Genome Mining of Peptidic Natural Products. *PLOS Computational Biology: Software*, 2014.

H. Mohimani, R. D. Kersten, W. T. Liu, M. Wang, S. O. Purvine, S. Wu, H. M. Brewer, L. Pasa-Tolic, N. Bandeira, B. S. Moore, P. A. Pevzner, and P. C. Dorrestein. Automated Genome Mining of Ribosomal Peptide Natural Products. *American Chemical Society*, 2014.