A Visualization and Evaluation Platform to Facilitate Prokaryotic Proteogenomic data analysis – VESPA

Elena S Peterson[1], Jeffrey L. Jensen[2], Hyunjoo Walker[2], Alexandra C. Schrimpe-Rutledge[3], Lee Ann McCue[4], Samuel H. Payne[3], Joshua N. Adkins[3], Bobbie-Jo M. Webb-Robertson[4*]

[1]Scientific Data Management, [2] Software Systems and Architecture, [3]Biological Separations and Mass Spectrometry, and [4]Computational Biology and Bioinformatics, Pacific Northwest National Laboratory, Richland, WA

The acceleration of genome sequencing has led to reliance primarily on machine annotation, which has led to an accumulation of misannotated genes. One approach to enhance current annotation efforts is proteogenomics; the use of mass spectrometry to identify peptides from expressed proteins in the context of all potential open reading frames (ORFs). However, because existing genome browsers are not designed to search outside the defined annotation, proteogenomic data analysis is currently an arduous task requiring significant manual manipulation of the data and tedious browsing through existing platforms. We present VESPA, which is a specialized prokaryotic genome browser that facilitates proteogenomic data analysis. VESPA includes advanced ingest, visualization, query, and export capabilities that highlight differences between the observed peptide data and the current annotation. The software has been developed using the Java SDK and is a NetBeans platform based application. It uses a modular approach and therefore the overall set of visualizations and query panels can be customized (https://www.biopilot.org/docs/Software/index.php).

With VESPA, we have eased the burden on the user of integrating proteogenomic data with genomic data files by requiring only a simple text or excel list of peptides as the input proteomics data. VESPA compares the peptides against translations in all reading frames to identify candidate locations for each peptide, thus removing the task of mapping peptide coordinates prior to visualization. Genome sequence and annotations are uploaded as standard FASTA and GFF files, respectively. VESPA then displays the genome at four resolutions; (1) the full genome displaying density of peptides or proteins, (2) linear regions displaying sub-sections of the genome, (3) six-frame translation at a high level, and (4) six-frame translation sequence view. In appropriate views, peptides that do not map to the current annotation (orphans) are highlighted for easy discovery. The most exciting capability of VESPA in respect to proteogenomics is the ability to perform specialized queries. The user can query on the different types of entities: proteins, peptides, orphan peptides, and probes. Of particular interest to proteogenomic investigators are both un-annotated ORFs and sequences upstream from the current annotation, for which one or more sources of orphan peptide evidence are observed. Queries can identify these regions of interest based on user defined settings, such as at least 2 orphan peptides between two stop codons. These query results can then be exported with genomic coordinates, as well as sequence, to facilitate further investigation with tools such as BLAST.

This work was supported by the National Institute of General Medical Sciences at the National Institutes of Health (NIH) under grant 1R011GM-084892 (BJWR). Supporting data was, in part, generating under National Institute of Allergy and Infectious Disease at NIH through interagency agreement Y1-A1-4894-01 (JNA) ((www.SysBEP.org).

**corresponding author, bj@pnl.gov, 509/375-2292*