VESPA: Visual Exploration and Statistics to Promote Annotation for Prokaryotic Genomes

Bobbie-Jo M. Webb-Robertson* Elena S. Peterson† Jeffrey L. Jensen‡

Mark A. Kobold§ Hyunjoo Walker** William R. Cannon†† Lee Ann McCue‡‡

Pacific Northwest National Laboratory

ABSTRACT

Visual Exploration and Statistics to Promote Annotation (VESPA) is an interactive visual analytics tool that integrates high-throughput data into a genomic context to facilitate the discovery of structural mis-annotations in prokaryotic genomes. Data is evaluated via visual analysis across multiple levels of genomic resolution, linked searches and interaction with existing bioinformatics tools. We highlight the novel functionality of VESPA and core programming requirements for visualization of these large heterogeneous datasets for a client-side application.

KEYWORDS: visual integration, proteogenomics, genome annotation.

INDEX TERMS: J.3 [Life and Medical Sciences]: Biology and genetics; I.3.8 [Computing Methodologies]: Applications

1 INTRODUCTION

The analysis of multiple high-throughput (HTP) molecular data types offers promise to better understand biological systems. A core benefit of HTP 'omic technologies are that they offer whole-cell molecular profiles to yield a system-level snapshot of an organism. However, data-driven analyses from these technologies are reliant upon accurate gene locations (structural annotation). The procedural aspects of genome sequencing and assembly have become relatively inexpensive, yet the full, accurate structural annotation of these genomes remains a challenge.

While next-generation sequencing transcriptomic data (RNA-Seq), global microarrays, and tandem mass spectrometry (MS/MS)-based proteomics have demonstrated immense value to genome curators as individual sources of information, integrating these data types to validate and improve structural annotation remains a major challenge. Current visual and statistical analytic tools are generally focused on a single data type, or existing software tools are retrofitted to analyze new data forms. For example, several prokaryotic genome browsers tools, such as ARTEMIS [1] and Gbrowse [2], have been used for proteogenomics via their capability to compare different gene annotation models. However, to use these tools with proteomics data requires transformation of the MS/MS peptide identifications into a standard format, such as a general feature file (GFF). This requires significant data formatting on the side of the user. In addition, finding locations of interest based on peptides that are outside the defined annotation is challenging.

We present VESPA, a novel desktop Java™ application focused on assisting scientists with the annotation of prokaryotic genomes through the integration of peptide-centric proteomics and transcriptomics data with current genome location coordinates, https://www.biopilot.org/docs/Software/Vespa.php.

2 FUNCTIONALITY

VESPA has been designed to provide simple ingest of data, responsive visualizations, filtering and standard exports. To create a project, the user uploads data in standard formats available for each data type (Excel, CSV, WIG, SAM) removing data reformatting issues. Mapping data types to the genome is performed by the software. This is a unique feature of VESPA and is preferable as locations of peptides and probes do not have to be recomputed by the user with updates to the underlying GFF file. Summary statistics are gathered for probes and peptides, including how many have been mapped to the annotated Open Reading Frames (ORFs) and how many belong to a region not associated with an annotated ORF (orphan peptide or probe).

A screenshot of the VESPA user interface is shown in Figure 1 with a project loaded for a dataset collected for *Synechococcus sp.* PCC 7002, which includes both MS/MS proteomics and RNA-Seq data. Three levels of resolution work concurrently to allow the user to more easily navigate the genome. Focusing on the middle visualization, the ORFs that were listed in the GFF are in blue. Each ORF is shown in the appropriate reading frame and peptides from the proteomics data file that are located within the boundaries of an ORF are highlighted in light blue. Peptides that are not associated with a defined ORF (orphan peptides) are shown in yellow with a gray box highlighting the potential coding region. The user can select this region (circled in red), bringing up a sequence-level view, or right click on the potential ORF to launch a BLAST search at NCBI to help determine if this region contains a missed gene call. RNA-Seq data is layered as a histogram in orange. The peptide and RNA-Seq data can be filtered from the visualization using criteria on the top bar. Peptides can be filtered based on redundancy and/or the likelihood of being observed by MS [3]. RNA-Seq data can be filtered based on physical count at each base pair (top center bar).

The most unique capability of VESPA is its search function. For the example in Figure 1, a search for all potential ORFs with at least two orphan peptides was performed, which identified 5 regions of interest. Quickly, these were identified as most likely 1) 1 mis-identification of the start codon, 2) 1 novel gene (highlighted in Figure1), 3) a functional frameshift (shown as two

902 Battelle Blvd, J4-33, Richland, WA 99352
*e-mail: bj@pnnl.gov
†e-mail: elena@pnnl.gov
‡e-mail: jeff.jensen@pnnl.gov
§e-mail: mark.kobold@pnnl.gov
**e-mail: hyunjoo.walker@pnnl.gov
††e-mail: bill@pnnl.gov
‡‡e-mail: leeann.mccue@pnl.gov

ORFs) and 3) a false discovery. The novel gene in Figure 1was evaluated via BLAST [4] (blast.ncbi.nlm.nih.gov) and found to be highly homologous to a 30S ribosomal protein S16 across multiple species. Search for ORFs with one orphan peptide finds 356 potential mis-annoations; automated triage of these for the user is for future work.

3 Software Design

There are currently two modules in VESPA, an independent data analyzer module and the user interface platform that are installed together as one application. These modules are built completely in Java™ with an embedded H2 (http://www.h2database.com) database.

The VESPA data analyzer ingests and processes all the data associated with a project and stores it in the database. It relies on the Apache POI (http://poi.apache.org) libraries to handle reading and writing Excel™ files. The data analyzer is independent from the user interface, so that it can process data independently or load projects while the user is preforming other actions. After processing and storing the data, it serves up the objects to the user interface for visualization, searching, and filtering capabilities.

The VESPA user interface is built using the Netbeans Platform (http://netbeans.org/features/platform) and relies heavily on Java 2D for its visualizations. Each "window" in the application is an independent Netbeans module. However each of the modules uses a common object model of the underlying data retrieved from the database. This object model combined with the Platform' event communication layer is how we keep our visualizations responsive and in sync.

To reduce lags while scrolling or zooming, pertinent data is kept in a bugger ahead and behind of the current visualization area. The various modules are registered with the platform at different levels depending on what types of events they should respond to. The platform then handles the communication among the modules which are coded to respond to the appropriate events when received. Because each module is independent they can be added or removed without disruption and can be reused in other settings. This also allows for quick coding of new modules. A new module simply needs to implement the platform API and register itself at one of the levels of interaction.

The visualizations themselves are straightforward Java 2D renderings based on the common data model. Each "rendered object" (ie, ORF, peptide, probe) is a Java 2D Shape object. We use transforms to handle the zoom and scale operations. Java 2D stroke objects are used for rendering the RNA-Seq data.

References

[1] K. Rutherford, J. Parkhill, J. Crook, T. Horsnell, P.Rice, M. Rajandream and B. Barrell. Artemis: sequence visualization and annotation. *Bioinformatics*, 16:944-945, 2000.

[2] R. Podicheti, R. Gollapudi, and Q. Dong. WebGBrowse – a web server for GBrowse. *Bioinformatics*, 25: 1550-1551, 2009.

[3] B. Webb-robertson, W. Cannon, C Oehmen, A. Shah, V Gurumoorthi, M. Lipton, and K. Waters. A support vector machine model for the prediction of proteotypic peptides for accurate mass and time proteomics. *Bioinformatics*, 26: 1677-1683, 2010.

S. Altschul, W. Gish, W. Miller, E. Myers and D. Lipman. Basic local alignment search tool. *J Mol Biol*, 15:403-410, 1990.

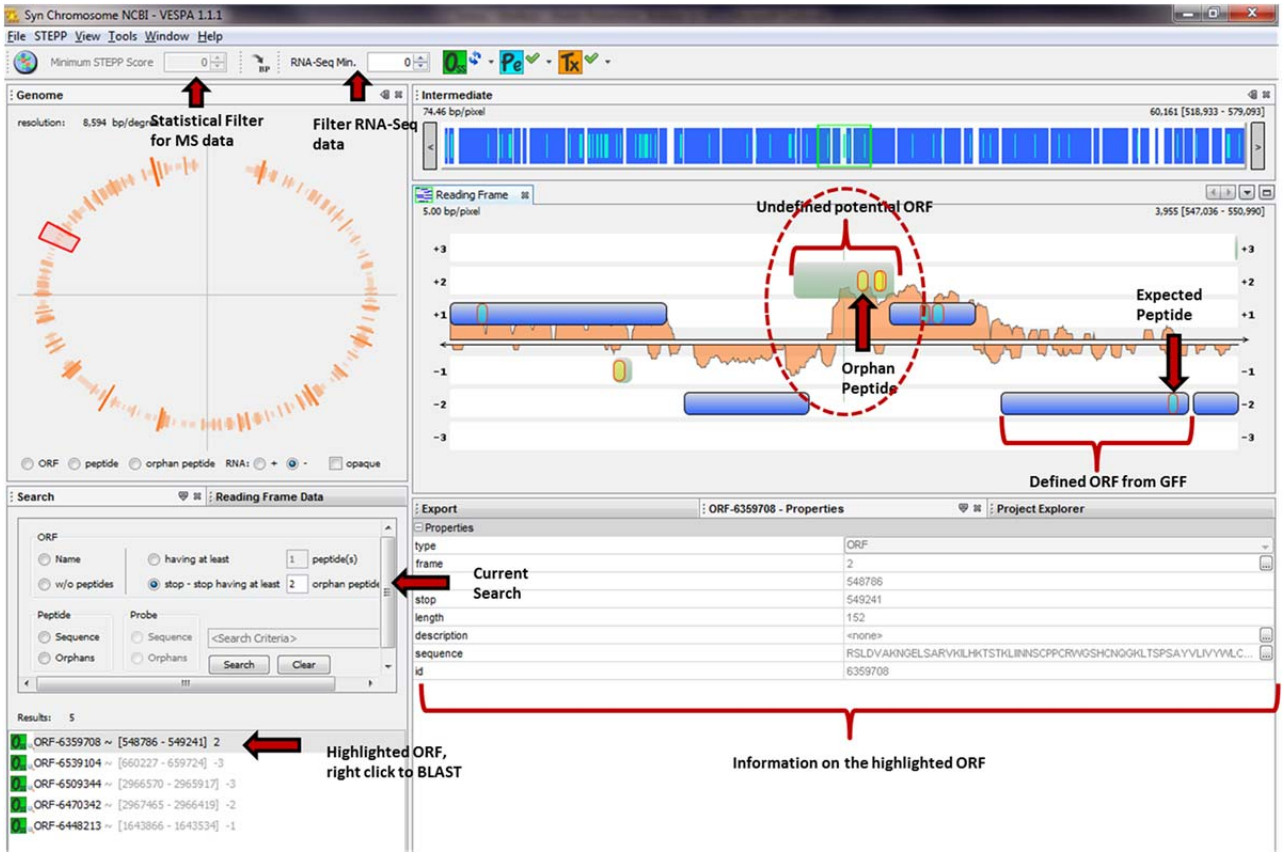

Figure 1. A static screenshot of the current VESPA visualization for proteomics and RNA-Seq datasets collect for an experiment on *Synechococcus*. The middle view shows the results of a query for any ORF region with at least two observed peptides that do not match a gene listed in the genome feature file (gray box). Highlighted in the box are orphan peptides (yellow). Confirmatory evidence that this is a previously un-called gene is provided by RNA-Seq data (orange histogram).