Proteomics for validation of automated gene model predictions

Kemin Zhou#, Ellen A. Panisko*, Jon K. Magnuson*, Scott E. Baker*, Igor V. Grigoriev#

US DOE Joint Genome Institute,

2800 Mitchell Dr, Walnut Creek, CA 94598

*Fungal Biotechnology

Pacific Northwest National Laboratory

Richland, WA 99352

2009

The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231

DISCLAIMER

Proteomics for validation of automated gene model predictions

Kemin Zhou#, Ellen A. Panisko, Jon K. Magnuson*, Scott E. Baker*, Igor V. Grigoriev#*

 # US DOE Joint Genome Institute,

2800 Mitchell Dr, Walnut Creek, CA 94598

**Fungal Biotechnology*

Pacific Northwest National Laboratory

Richland, WA 99352

Abstract

High throughput liquid chromatography mass spectrometry (LC-MS) based proteomic analysis has emerged as a powerful tool for functional annotation of genome sequences. These analyses complement the bioinformatic and experimental tools used for deriving, verifying and functionally annotating models of genes and their transcripts. Furthermore, proteomics extends verification and functional annotation to the level of the translation product of the gene model.

Key Words: protein, peptide, proteomics, mass spectometry, LC-MS, gene model, genome, proteome, annotation, splice site, intron, exon, BLAST, FASTA, eukaryote, *Phanerochaete chrysosporium*.

1. Introduction

The explosion of genome sequencing projects in the last decade has been an enormous boon to biological researchers. However, the quantity of DNA sequence data presents a challenge for annotation and use. The development and refinement of algorithms for annotation of genome sequences has been crucial for providing *in silico* predicted gene models. This is especially true for eukaryotic genomes with the additional complexity of introns and exons. Expressed sequence tag (EST) data are critical for experimental verification of predicted transcripts from the gene models. The placement of splice sites, information about transcription start sites and untranslated regions of the gene can be derived from the EST data.

Analogous to the use of EST data, the use of high throughput LC-MS based proteomic data to experimentally verify and adjust predicted models for translation products of the gene models has begun to emerge. An organism can be grown under diverse sets of conditions and the extracted proteins pooled before analysis in order to maximize the proportion of the proteome that is observed. Alternatively, specific growth conditions at specific developmental stages or time points can be analyzed separately to extend the functional annotation of a proportion of the proteome, e.g., associate a subset of proteins with a particular metabolic state. Similar to EST data, proteomics data can be used to verify splice sites in transcripts where the identified peptides span an exon/intron/exon boundary. But proteomics data adds an additional dimension of functional annotation, as it can be used to verify predicted translation start and stop sites, signal peptide cleavage sites and post-translational protein modifications.

2. Materials

2.1 NCBI BLAST: The programs needed for BLAST based mapping of peptides are available for a variety of operating systems for free download from the NCBI (www.ncbi.nlm.nih.gov/BLAST/download.shtml; Altschul et al 1997).

2.2 Genome sequence: A complete or high coverage draft genome sequence.

2.3 Gene models: Gene models are generated for a given genome by automated gene calling software, such as Genewise (Birney & Durbin, 2000) or Fgenesh (Salamov & Solovyev, 2000).

2.4 Peptide sequence data from global proteomic experiment(s): Must be in a format, such as FASTA, that is compatible with command-line batch BLAST analysis.

3. Methods

3.1 Proteomic analysis. Global proteomic analysis is used to determine the sequences of peptides present in a protein sample.

3.2 Peptide mapping. Peptides that have been identified in the previous step are mapped to the genome sequence using the tblastn tool with the following options:

```
tblastn -e 1000 -W 2  -F F -f 6 -K 50
```

The expectation value (-e option) is given the large value of 1000 in order to collect short matches to the genome. A very short word size of 2 (-W option) is used to increase the sensitivity of the blast algorithm. Turning off the filtering (-F option) is also essential for successfully matching short peptides to the genome. A lower threshold value is used to extend the length of the matches (-f option). Because the lower thresholds generate a lot of matches, and only one or two are true matches the number of alignments reported is lowered to 50 (-K option).

3.3 Quality assessment. After the peptides are mapped to the genome, a best match is selected for each peptide. The quality of the peptide matches are assessed with respect to percent sequence identity and percent coverage. The peptide is further categorized based on the presence or absence of a gap. Based on the collective criteria the peptides are assigned to one of the following categories:

Perfect match:	100% sequence identity without gaps in the alignment covering more than 98% of the peptide sequence.
Split match:	two non-overlapping fragments of a peptide exhibit perfect matches to two genomic sequences in close proximity. Total coverage is equal to 100% of the peptide.
Imperfect match:	80-99% sequence identity, 90-98% coverage.
Imperfect split:	two fragments of a peptide with maximum overlap of two amino acids have imperfect matches to two regions of genome sequence in close proximity. Total coverage is close to the entire peptide length.
Uncertain:	the remainder of the mapped peptides

Several top matches of the same quality are retained to reflect gene duplication or non-unique peptides in the genome sequence (See Note 4.1 and figure 1 for an example).

3.4 Validation of predicted genes. Comparison of the coordinates of mapped peptides with the coordinates of predicted gene models in the genome sequence provides experimental support for the predicted genes. Given sufficient peptide coverage of a predicted gene model, the boundaries of the protein coding portion of the gene and splice sites can be verified by the experimental proteomics data (see Note 4.1 for an example).

4. Notes

4.1 Example. 4,825 peptides were used for mapping to the version 2.0 genome assembly of *Phanerochaete chrysosporium* and for validations of predicted gene models (v2.1) for this assembly. The average peptide length was fourteen amino acids. The distribution of peptide lengths is shown in Fig 1. The peptides were mapped to 5,135 locations on the genome including 4,149 (81%) perfect matches (Fig. 2). The difference between the number of locations and the number of total peptides reflects the existence of non-unique peptide matches and split matches. A total of 224 split matches were detected with 107 perfect and 117 imperfect split matches providing support for splice boundaries in genes containing two or more exons. Excluding the uncertain peptides, support was provided for 2,795 exons representing 1,440 of 10,048 predicted gene models.

4.2 Database searching. There are a number of different algorithms available for searching mass spectrometry generated peptide data against protein databases. In the example in Note 4.1, Sequest was used with the following cutoffs for fully tryptic peptides:

charge state 1: Xcorr ≥1.9

charge state 2: Xcorr ≥ 2.2

charge state 3: Xcorr ≥ 3.75

all must have a $\Delta CN^2 \geq 0.1$

Xcorr is a statistical estimate of the cross-correlation sequence of a random process. In this application, it is a measure of the quality of the match of the peptide derived from the mass spectral data to the peptide determined from translation of the genome sequence. In general, the

most important aspect of accurate peptide identification is the use of stringent parameters, regardless of the algorithm.

4.3 Peptide format. Peptide sequences should be in FASTA format.

4.4 Data display. The procedure is designed for work with the JGI Genome Portal (Figure 3; genome.jgi-psf.org) and the underlying mySQL database. However, this procedure can easily be modified for any genome browser that displays "features", such as gene models, on "tracks" mapped back to genome sequence contigs or scaffolds. The Generic Model Organism Database (GMOD) Project has produced a mostly open-source visual genome database that displays information "tracks" (www.gmod.org; Stein et al 2002). Visual display is critical for manual genome curation. At a genome-wide level, text-based lists of gene models with associated numbers and coordinates of peptides can be used effectively for analysis of automated annotation.

5. References

Altschul, SF, Madden, TL, Schäffer, AA, Zhang J, Zhang, Z, Miller, W, Lipman, DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search Programs. *Nucleic Acids Res.* 25:3389-3402.

Birney E. & Durbin R. 2000 Using GeneWise in the *Drosophila* annotation experiment. *Genome Res* **10**, 547-548.

Salamov AA & Solovyev VV. 2000 Ab initio gene finding in *Drosophila* genomic DNA. *Genome Res.* **10**, 516-522.

Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW, Arva A, Lewis S. 2002. The generic genome browser: a building block for a model organism system database. *Genome Res.* 12(10):1599-610.

Fig 1. Peptide length distribution for the set of *Phanerochaete chrysosporium* peptides in the example (note 4.1).

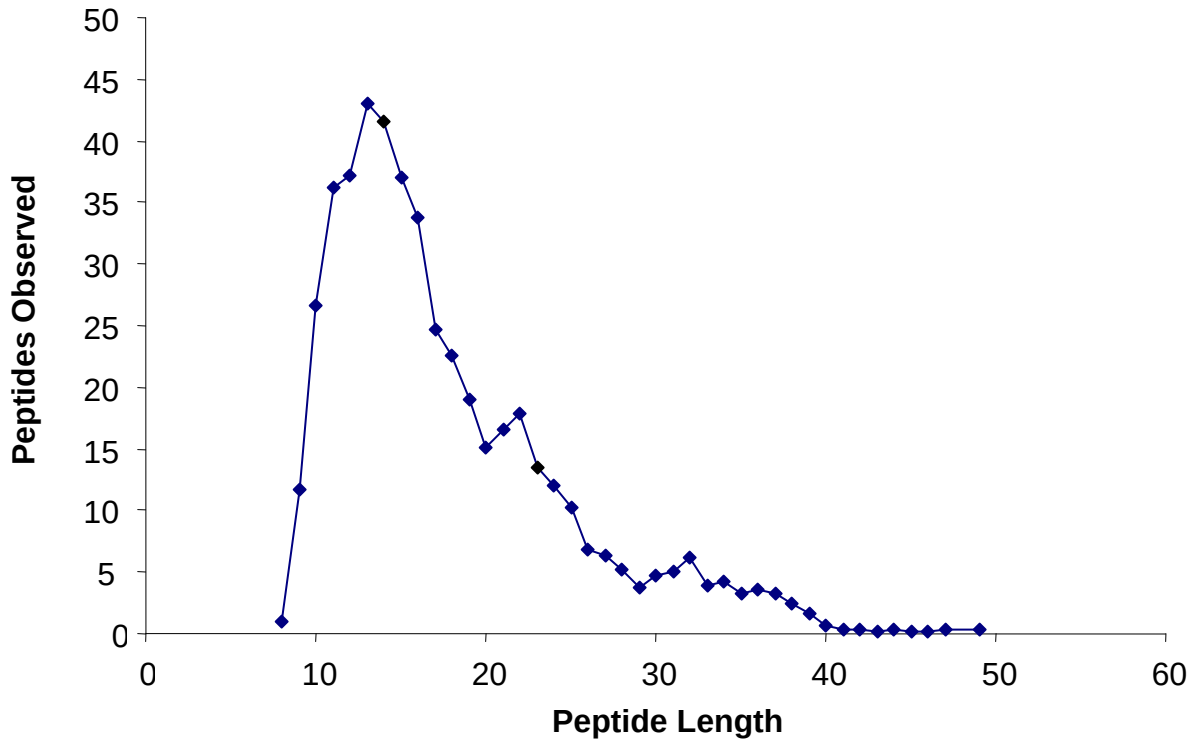

Fig 2. Distribution of mapped *P. chrysosporium* peptides according to the quality categories described in section 3.3.

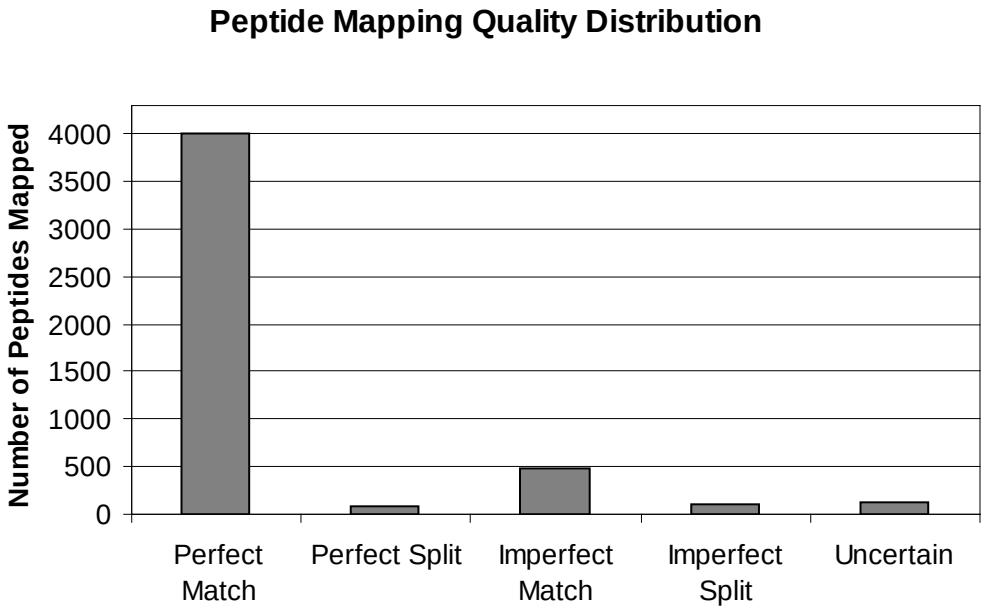

Fig 3. Display of peptide data on the JGI *P. chrysosporium* Genome Browser. Genome scaffold, black; gene model, red; mapped peptides, green (Note 4.1).