Specific Peptides Facilitate Metagenomic Analysis

Uri Weingart*, Erez Persi* and David Horn+
School of Physics and Astronomy, Tel Aviv University, Tel Aviv 69978, Israel
**Supported in part by fellowships granted by the Edmond J. Safra Bioinformatics program at TAU*
+ Corresponding author. Email: horn@tau.ac.il

Abstract

Specific Peptides (SPs) are sequence markers (1) for enzymatic functionality extracted from Swiss-Prot data. When found on large strings of genomic or proteomic origin SPs provide quick enzymatic annotations. The methodology of Data Mining of Enzymes (2) uses the criterion of coverage length (overall number of amino-acids in consistent SP hits) ≥ 7 to provide EC annotations at levels 3 or 4, thus specifying the biochemical function of an enzyme on the basis of its sequence. It has been applied to Sargasso Sea Data (3) uncovering 220K enzymes among 1 M protein sequences. A user-friendly tool that displays occurrences of SPs on any protein sequence that is presented as a query, together with the EC assignments due to these SPs, is available at ***http://adios.tau.ac.il/DME***

Recently SP usage has been extended to direct search on short reads (4). By collecting all short reads where SPs of a given EC can be located, an estimate is provided for the abundance of its relevant genes, thus generating an enzymatic spectrum of its genomic or metagenomic source. Moreover, some of its taxonomic decomposition can be deciphered using a subset of SPs belonging to aaRS enzymes. An SPSR tool providing SP hits on queried lists of short-reads is available at ***http://horn.tau.ac.il/SPSR***

A novel usage reported here is employing a subset of SPs for the task for species counting in metagenomic data (5). Using a list of 4000 SPs of length ≥ 9, belonging to a subset S61 of EC:6.1.1. aaRS enzymes that are single genes in bacterial genomes, we identify their occurrences on given lists of short reads or contigs. Identifying the largest number of reads associated with one SP, we propose an algorithm that constructs a minimal number of fused strings that differ from each other, thus serving as estimates for the different genes that could have led to the observed reads or contigs. Short reads lead to bounds on numbers of families, while long reads or contigs lead to lower-bound estimates of numbers of strains and species. This method can serve as complement to conventional 16S rRNA analysis.

References

1. V. Kunik et al, PLOS Comp. Bio. 2007, 3(8):e167.
2. U. Weingart, Y. Lavi and D. Horn, BMC Bionformatics 2009, 10:446.
3. J. C. Venter et al, Science 2004, 304:66.
4. U. Weingart, E. Persi, U. Gophna and D. Horn, BMC Bionformatics 2010, 11:390
5. E. Persi, U. Weingart, S. Freilich and D. Horn, 2011, in preparation