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

- Specific Peptides (SPs) are sequence markers (frames 1,2) for enzymatic functionality extracted from Swiss-Prot data.
- When found on large strings of genomic or proteomic origin SPs provide quick enzymatic annotations (frame 3).
- They can also be directly applied to short read metagenomic data (frame 4) in order to extract an enzymatic spectrum (frame 5), as well as taxonomic classification of its bacterial species.
- Moreover, a particular subset of SPs underlies a novel algorithm for species-counting (frame 6). The latter can serve as complement to conventional 16S rRNA analysis in microbial metagenomics.

4) The SPSR (Specific Peptide hits on Short Reads) Methodology





Vered Kunik, Yasmine Meroz, Zach Solan, Ben Sandbank, Uri Weingart, Eytan Ruppin and David Horn. Functional representation of enzymes by specific peptides. PLOS Computational Biology 2007, 3(8):e167.

 Search SPs of length ≥ 7 amino-acids on all 6 frame translations of every Short Read • Each SP hit assigns its EC label to the corresponding Short Read • Accumulate Specific Peptide Short Read (SPSR) events according to EC labels • Multiply by factors to obtain prediction of enzyme content (protein numbers) • Factor concept is tested on simulations of short reads of the *E coli* genome for different choices of short read length • Factors are deduced from analysis of artificial metagenomes constructed of various combinations of 7 out of 11 well-annotated bacteria (training set).

• Method is tested on metagenomes of 11 bacteria (test set) for which short reads of length 50 nucleotides are produced randomly from NCBI genomes, with 5 fold coverage of each full genome. • Taxonomic association is deduced from Taxon Specific SPs (TSPs) that belong to the set S61 of single-gene aaRS enzymes.

Uri Weingart, Erez Persi, Uri Gophna and David Horn Deriving enzymatic and taxonomic signatures of metagenomes from short read data.BMC Bioinformatics 2010, 11:390



of 7 out of 11 bacteria from the test set



5) Enzymatic and Taxonomic Signatures of Metagenomes generated by SPSR



Bacillus cereus (strain A

Salmonella typhimuriu Shigella flexneri Salmonella typh Escherichia coli (K12) Caulobacter crescentu

Thermotoga petrophila

_eptospira biflexa serovar Pato

Alignment is ordered according to SPs (in red). Spaces are inserted to highlight annotations of active and binding sites.

3) Data Mining of Enzymes

- Utilization of Specific Peptides for large volume enzymatic prediction
- Find whether protein is an enzyme and, if so, what is its EC classification
- Use coverage length (overall number of amino-acid in consistent SP hits) ≥ 7 Testing method on 20K novel enzyme entries find precision 99% recall 92%
- Application to Sargasso Sea data (Nature 2004) uncovers 220K enzymes among 1M putative protein sequences
- Comparison of enzymatic spectrum with human gut microbiome of Gill et al (Science 2006)

Leading EC categories

• EC=6.1.1 aminoacyl tRNA synthetases • EC= 1.1.1 (alcohol dehydrogenases with NAD+ or NADP+ as acceptor)



Analysis of 3 metagenomes from Dinsdale (Nature 2008) shows similarities between two sets and anomalous behavior in Soudan Black data.

Short reads have average length of 103 nucleotides (SD 18).

| | Class | Edwards | CARMA | S61TSP |
|---|-----------------------|---------|-------|--------|
| Comparison of class predictions within proteobacteria for the Soudan Red data with the methods of Edwards (16S rRNA) and of Carma (protein based). | Alphaproteobacteria | 40% | 37% | 45% |
| | Gammaproteobacteria | 54% | 40% | 45% |
| | Betaproteobacteria | 2% | 8% | 8% |
| | Epsilonproteobacteria | 0% | 2% | 2% |
| | Deltaproteobacteria | 3% | 13% | 0% |
| | | | | |

6) Species Counting in Metagenomic Data

- Use 4000 SPs of length \geq 9 belonging to a subset S61 of 6.1.1. enzymes (aaRS) that are single-genes in bacterial genomes
- Identify lists of reads (or contigs) carrying the same SP and choose largest lists
- Algorithm constructs minimal number of fused strings that differ from each other, serving as estimates for the independent genes that could have lead to the observed reads
- Short reads lead to bounds on numbers of families, while long reads or contigs lead to lower-bound estimates of numbers of strains, species and genera.
- Method can serve as complement to conventional 16S rRNA

Example of Short Reads

ort read (translated to amino-acid strin ILTSSSPEGARDFLVPSRLNPGKFYALPQAPQQFKQL VFFSFLLGFTKGKFYALPQAPQTILSNLFMVSGFDKYFTNC GARDFLVPSRLNPGKFYALPQAPQQFKQLIMVSGF

Contigs of human gut microbiota. Data of Qin et al (Nature 2010) of 124 individuals.



Uri Weingart, Yair Lavi and David Horn Data Mining of Enzymes using Specific Peptides.. **BMC Bioinformatics 2009, 10:446**

- EC= 3.6.3 (hydrolases catalysing
- transmembrane movement of substances).
- EC=2.7.7 Nucleotidyltransferases

Leading EC# in Sargasso Sea

| # proteins | Enzymatic activity |
|------------|---|
| 5,993 | DNA-directed RNA polymerase |
| 2,999 | NADH dehydrogenase (quinone) |
| 2,610 | DNA topoisomerase (ATP-hydrolysing). DNA gyrase. |
| 2,198 | carbamoyl-phosphate synthase (glutamine-hydrolysing) |
| 2,169 | H ⁺ -transporting two-sector ATPase. ATP synthase. |
| 2,083 | DNA-directed DNA polymerase |
| | # proteins 5,993 2,999 2,610 2,198 2,169 2,083 |

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. An SPSR tool providing SP hits on queried lists of short-reads is available at http://horn.tau.ac.il/SPSR .

Species-counting example of Rios Mesquites metagenomic short reads carrying a common SP. The first 14 rows display the short reads. X indicates inconsistency of short read with all others. After elimination of 8 reads we are left with 6 reads that can be fused into the two last rows. Thus the 10 strings indicated by X form a possible solution of the minimal chromatic number problem (for a graph whose vertices are reads and edges are inconsistency relations), resulting in a species count of 10. Since these are short reads, we estimate that this count indicates the existence of 10 different families or orders.

Erez Persi, Uri Weingart, Shiri Freilich and David Horn. Submitted

Published papers, web-tools and algorithms are available at http://horn.tau.ac.il.

- Leading SPs:
- ISRQLWWGH (EC=6.1.1.9, gene=SYV) 1488 hits. Species Count 1009. TRFPPEPNGYLH (EC=6.1.1.18, gene=SYQ) 1961 hits. Species Count 888. GEAAFYGPK (EC=6.1.1.3, gene=SYT) 1488 hits. Species Count 718.
- Adding non-leading SPs we obtain counts of 1136, 937 and 1076, respectively.

Differentiating between strains

and species. This follows trends of different S61 genes in Uniprot data as shown to

the right.

By analyzing all strings of the leading SPs and the distances between them, we conclude from SYQ (cut at distance 2) that more than 400 may account for different strains rather than different species



Statistics of Uniprot show a cutoff at d=2 aa in SYQ sequences distinguishes between strains (D) and species (F)