

# Specific Peptides Facilitate Metagenomic Analysis

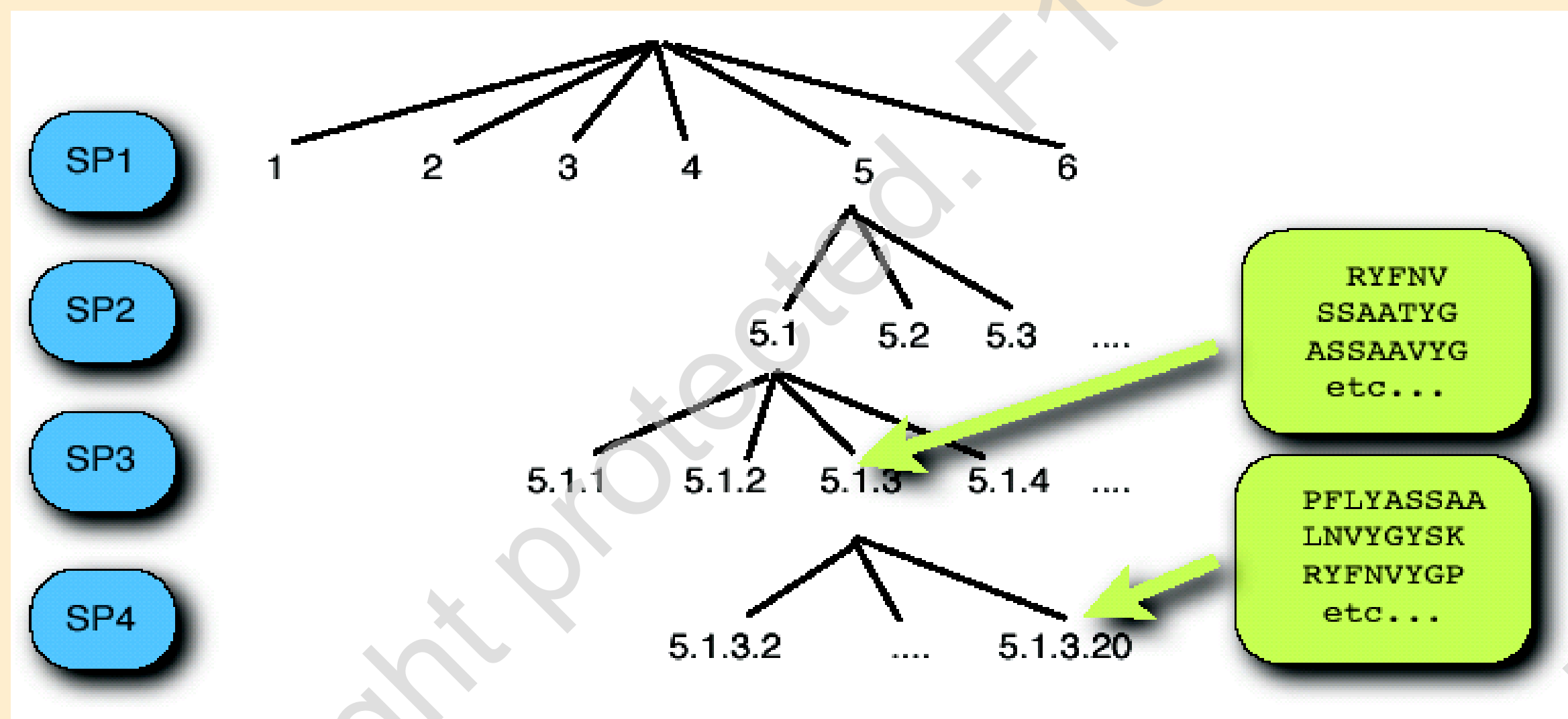
Uri Weingart\*, Erez Persi\* and David Horn

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.

## 1) EC hierarchy and specific peptides



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

## 2) Examples of SP occurrences on sequences of enzymes

```
5.1.3.20
545048|HDD_HABIN YCLDREIPFFVAS S AATYG-DTKVFEERE---FEGPLNV Y GYS K FLEDOYVNNILPE-AKSPVCGRYFNVYGP 174
Q9CL97|HDD_PASMU YCLDREIPFFVAS S AATYG-DTKVFEERE---FEGPLNV Y GYS K FLEDOYVNNILPE-AKSPVCGRYFNVYGP 174
Q7V988|HDD_HABEDU FVDVROIPFLIAS S AATYG-DAANFIEERK---FEGPLNV Y GYS K FLEDOYVNNILPE-ANSAICGGFYFNVYGP 175
Q82JN4|HDD_YERPE FCLDRSIPFLIAS S AATYG-GRDTNFIEDRQ---YEQPLNV Y GYS K FLEDOYVNNILPE-ADSOICGGFYFNVYGP 175
P67910|HDD_ECOLI YCLEREIPFLIAS S AATYG-GRDTNFIEDRQ---YEQPLNV Y GYS K FLEDOYVNNILPE-ADSOICGGFYFNVYGP 175
Q7N1E6|HDD_CHEVO YCLEREIPFLIAS S AATYG-GRDTNFIEDRQ---YEQPLNV Y GYS K FLEDOYVNNILPE-ADSOICGGFYFNVYGP 176
Q51061|HDD_NEIGO WQDERIPFLIAS S AATYG-KGEIFREERE---LEKPLNV Y GYS K FLEDOYVNNILPE-ADSOICGGFYFNVYGP 177
Q9W966|HDD_BURPS ACLAQOTFLIAS S AATYG-SRRFYERE---FEAPLNV Y GYS K FLEDOYVNNILPE-ADSOICGGFYFNVYGP 174
Q7W609|HDD_BORBR YCAQERYFLIAS S AATYG-GSSVVEDPA---NEHPLNV Y GYS K FLEDOYVNNILPE-ADSOICGGFYFNVYGP 172
Q72ET7|HDD_DESVH LCMETGAFINIAS S AATYG-DGSLGSDDDTTLMLKPLNM Y GYS K QLFDLWAYREGRL---DGIALSKFYFNVYGP 176
***
```

```
SP | peptides
SP4 | PFLYASSAA LNVYGYSK YGYSKFLFEDVVR RYFNVYGP YNVYGPPE FSSSATVYG
    | IPYVESFPTG MVEQIL LLRYFNP YENVAGA
SP3 | SSAATYG ASSAAVYG RYFNV
```

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)  $\geq 7$
  - 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)

