TEL AVIV UNIVERSITY אוניברסיטת תל-אביב
RAYMOND AND BEVERLY SACKLER FACULTY OF EXACT SCIENCES
SCHOOL OF PHYSICS AND ASTRONOMY
הפקולטה למדעים מדויקים ע"ש ריימונד ובברלי סאקלר
בית-הספר לפיזיקה ולאסטרונומיה

Research of the Enzome using Specific Peptides

Dissertation submitted towards the degree of

Doctor of Philosophy

by

Uri Weingart

Submitted to the Senate of Tel Aviv University

January 2011

This work was carried out under the supervision of

Professor David Horn

Acknowledgments

I would like to thank my PhD supervisor, Professor David Horn for sharing his incredible insight, uncompromisingly critical vision and commitment in setting our research in the right direction, sometimes into places that I never dreamed.
Thanks to Prof. Horn my years of Bioinformatics research in the University of Tel-Aviv have been an experience of a lifetime.
I look forward to continuing to work with Prof. Horn in the years to come.

I would like to thank the Edmond J. Safra Foundation for their support.

I would like to thank my children, Liat and Shai who in one of those peculiarly delightful twists of life became providers of parental advice and encouragement to their student father.

And finally, I would like to thank my wife Tania for her support, encouragement, and love that made my work possible.

Abstract

Recently there has been a rapid growth in the number of putative proteins derivable from new genomic and metagenomic data. The extended use of environmental shotgun sequencing to study diverse microbial systems has made metagenomics a vastly growing field leading to a flux of data, calling for development and application of new tools that allow its investigation. Conventional tools for predicting the function of a protein from its sequence are based on sequence-similarity or sequence-motifs. The focus of this work is description of a prediction method that is applicable to large numbers of sequences. Its purpose is finding whether each protein in the data is an enzyme and, if so, what its EC classification is. This Data Mining of Enzymes (DME) is based on the Specific Peptide (SP) method and is carried out by comparing the sequences of all proteins with a list of all SPs and looking for matches of the latter in the data.

SPs are strings of amino-acids, extracted from enzyme sequences using the motif extraction algorithm MEX, which will be described below. They are selected for their specificity to levels of the Enzyme Commission (EC) 4-level functional hierarchy. We describe the construction and description of the utilization of enzymatic SPs for the purpose of prediction of enzymatic function and other biological functions of single proteins, metaproteomes, genomes and metagenomes.
The first two chapters describe in detail generation of SP datasets and the subsets utilized to generate biological predictions, such as active, metal and binding sites, gene ontology, gene name and taxonomic lineage of queried sequences.

The third chapter is devoted to DME, whereby we mean a method of applying SPs to proteomic data in order to obtain enzymatic predictions. We present and establish our methodology and derive enzymatic predictions for three large metagenomes, one of them (Sargasso Sea metagenome) exceeding 1 million proteins. For the latter we present all our enzymatic assignments, including a number of proteins having two or three enzymatic functional properties.
We present the concept of enzymatic profile of a proteomic sample and demonstrate it on the data that we have studied.

Chapter four presents the analysis of genomes and demonstrates the capability of SPs to uncover enzymatic genes on a full genome without any prior knowledge of gene boundaries. We present the concept of SP scaffolding to determine approximately the beginning and end of a gene. Enzymatic genes are annotated both by SPs, specifying their EC assignments, and by FSPs specifying gene names (or protein families). We show the capability to detect shifts in nucleotide sequences caused by addition or deletion of single or few nucleotides in a coding sequence, detecting traces of genetic evolution. We present, as an example, the analysis of H. Pylori 26995, showing the enzymatic predictions for its full genome.

Chapter five presents application of the methodology described in the previous four chapters to analysis of short reads of metagenomes, without requiring gene reconstruction or contig formation. We present the SPSR (SP Short Reads) method used to generate enzymatic function predictions, and the TSPSR (Taxon Specific Peptides Short Read) method used to generate taxonomic lineage predictions.

We apply these concepts to several metagenomes and derive their enzymatic and taxonomic signatures.

While most of the emphasis of the research was focused on processing of large volume Bioinformatic data, several web tools were provided to demonstrate the concepts developed as part of this thesis in an on-line mode. Chapter seven describes these web tools.

The remarkable simplicity and versatility of the SPs is demonstrated in the different chapters of this thesis: Utilization of Specific Peptides consists of searching a pattern (SP) within an amino-acid sequence (of a protein or 6 frame translated nucleotide sequence of a genome or short-read) followed by calculation of coverage and analysis of the results. We show in this thesis that this simplicity and flexibility allows for a variety of different biological predictions utilizing very similar methodologies.

Contents

1. Introduction: The SP methodology

Annotation of protein function facilitates understanding of biological processes. Currently there is a widening gap between the number of proteins being identified by sequence genomic methods and their predicted function. In the last few years several approaches have been developed in order to bridge that increasing gap. The classical approach involves transfer of annotation from a functionally characterized protein to its functionally uncharacterized homologs [1.8]. Other methods consist of phylogenetic methods, application of sequence motifs, structural similarity and structure patterns [1.10].
Motif based approaches have been presented by others in the past and they include Prosite[1.11], MEME[1.12], eMotif algorithm and eMotif database[1.13] and Protein Sequence Motifs [1.14].

One of our main goals is to provide a comprehensive methodology to predict enzymatic functions of proteins.
Greater availability of sequence data and decreasing cost of computer resources make motifs' based approaches increasingly attractive and feasible and makes their use one of the essential tools of sequence analysis [1.11].
In that spirit we employ and investigate the Specific Peptides (SPs) approach.
SPs are strings of amino acid motifs which are unique to a branch of the Enzyme Commission functional classification (EC). SPs were reported in 2007 by Kunik et al. [1.1]. A year later, Meroz and Horn explained their biological roles in enzymes [1.2]. Data Mining of Enzymes, presented here, is a methodology used to predict the enzymatic function of proteins from their sequences using SPs.

The work presented expands the concepts introduced by Kunik et al and Meroz and Horn to a methodology capable of providing enzymatic predictions for enzymatic functions at large, including metaproteomes and metagenomes. We assign the conglomerate of all enzymes of a species the designation of "Enzome", and use it also for metagenomic studies.

The core of the methodology consists of searching for the SPs within a queried amino acid sequence and analyzing the resulting SP hits in order to predict the EC of the amino acid sequence. One of the most important conclusions of the work presented here is that the best predictor to determine EC of queried sequences using SP hits is the length of coverage of the SPs on the queried sequence, which is the number of amino acids in the queried sequence coinciding exactly with the SPs.

The search of SP hits within the queried sequence consists of simple, deterministic exact searches of a pattern within text. Searches of SPs within queried proteins have been optimized for large volume processing using well known algorithms, such as the Knuth Morris Pratt string matching text-processing method [1.3]. The method easily lends itself to large volume processing which was conducted in parallel in a farm of computer servers using tools such as Condor.

Analysis of the SP hits is a very short prediction algorithm. This algorithm consists of calculating the length of amino acid coverage of SPs on the queried sequence.

The construction of SPs uses the Motif Extraction (MEX) algorithm [1.5]. MEX is an unsupervised method that neither requires multiple sequence alignment nor relies on over-representation. MEX is run against annotated Swiss-Prot enzyme sequences. The results of the MEX run are a set of motifs. These motifs are distilled using specificity criteria and the final product is a set of SPs. The process is described in detail in the Methods section.

The process of generating new SPs has been run twice in the last few years.
The frequency of future runs will be driven by the need to catch up with new annotations in Swiss-Prot.

We have also generated "Annotated SPs" (ASPs) which are SPs specific not only to a certain branch of the EC number tree but also to specific active sites, metal sites, binding sites, taxonomies or GO annotations. These ASPs can be tagged with those attributes and can serve as predictors for these attributes. By searching for the particular subsets of ASPs within a protein, it is possible to provide predictions for active sites, metal and binding sites of the queried proteins. Annotated SPs will be discussed in detail in a separate chapter below.
Similarly, we construct Taxa Specific Peptides (TSPs), Gene Ontology based Specific Peptides (GSPs) and Family Specific Peptides (FSPs) all of which will be described and discussed in a separate chapter below.

An online system that predicts the EC, active, metal and binding sites has been built as part of this research and is available at http://adios.tau.ac.il/DME

While technology does not constitute the main focus of this research, it had a prime role supporting analysis of large volumes of data which dominated this work.
One of the most important aspects of DME is not only its simplicity but the capability to use it with great ease in parallel mode processing, which provides significant advantages when researching the large amounts of data collected.
We develop the SP approach further, to exploit the multitude of data available from short reads directly. We present tools designed to derive taxonomic signatures directly from short reads without utilization of 16S rRNA as a taxonomic indicator.
This is of significant importance in view of the fact that in many cases composition of metagenomic data is unknown and contig assembly from short reads leaves many singletons behind. Such short read singletons can be studies with SPSR to provide an enzymatic spectrum and some taxonomic signatures.

References

1.1 Kunik V, Meroz Y, Solan Z, Sandbank B, Weingart U, Ruppin E, Horn D: Functional representation of enzymes by specific peptides. PLOS Comp Biol 2007, 3(8):e167.

1.2 Meroz Y, Horn D: Biological Roles of Specific Peptides in Enzymes. Proteins: Structure, Function, and Bioinformatics 2008, 72(2):606-612.

1.3 Christian Charras - Thierry Lecroq; Exact matching algorithms: www-igm.univ-mlv.fr/~lecroq/string/

1.4 Boeckmann B., Bairoch A., Apweiler R., Blatter M.-C., Estreicher A., Gasteiger E., Martin M.J., Michoud K., O'Donovan C., Phan I., Pilbout S. and Schneider M., The Swiss-Prot Protein Knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res. 31:365-370(2003).

1.5 Solan Z, Horn D, Ruppin E, Edelman S: Unsupervised learning of natural languages. Proc Natl Acad Sci USA 2005, 102:11629-11634.

1.6 Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: Basic local alignment search tool. J Mol Biol 1990, 215:403-410.

1.7 Kunik V, Solan Z, Edelman S, Ruppin E, Horn D: Motif Extraction and Classification. http://neuron.tau.ac.il/~horn/publications/csb05.pdf

1.8 Espadaler J, Eswar N, Querol E, Avilés F, Sali A, Marti-Renom M, and Oliva B: Prediction of enzyme function by combining sequence similarity and protein interactions: BMC Bioinformatics. 2008; 9: 249

1.9 Tian W, Skolnick J. How well is enzyme function conserved as a function of pairwise sequence identity? J Mol Biol. 2003;333:863–882. doi: 10.1016/j.jmb.2003.08.057.

1.10 Friedberg I. Automated Function Prediction: the Genomic Challenge Briefings in Bioinformatics (2006) 7(3):225-242

1.11 Sigrist CJA, Cerutti L, de Castro E, Langendijk-Genevaux PS, Bulliard V, Bairoch A, Hulo N.: PROSITE, a protein domain database for functional characterization and annotation. Nucleic Acids Res. 38(Database issue)161-6 (2010). PubMed: 19858104

1.12 Timothy L. Bailey and Charles Elkan, "Fitting a mixture model by expectation maximization to discover motifs in biopolymers", Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology, pp. 28-36, AAAI Press, Menlo Park, California, 1994.

1.13 Jimmy Y. Huang and Douglas L. Brutlag, "The EMOTIF Database," Nucleic Acids Res., 2001, 29, 202-204.

2. Specific Peptide Lists

2.1 Brief overview of the Enzymes Commission Numbers

The Enzyme Commission Number set (EC) is a numerical classification scheme for enzymes, based on the chemical reactions they catalyze. Classification is done at four levels, in the format $n_1.n_2.n_3.n_4$, where each subsequent level represents a finer classification. There exist repositories for enzyme classification, amongst them Swiss-Prot/ENZYME via the following link http://expasy.org/enzyme/ and Brenda at http://www.brenda-enzymes.org. Table 2.0 below shows the classification of enzymes by the highest level of the EC number.

Classification of Enzymes at EC level 1

Group	Reaction catalyzed	Typical reaction
EC 1 - Oxidoreductases	To catalyze oxidation/reduction reactions; transfer of H and O atoms or electrons from one substance to another	Oxidation A + O → AO Reduction AH + B → A + BH
EC 2 - Transferases	Transfer of a functional group from one substance to another. The group may be methyl-, acyl-, amino- or phosphate group	AB + C → A + BC
EC 3 -Hydrolases	Formation of two products from a substrate by hydrolysis	AB + H2O → AOH + BH
EC 4 -Lyases	Non-hydrolytic addition or removal of groups from substrates. C-C, C-N, C-O or C-S bonds may be cleaved	RCOCOOH → RCOH + CO2
EC 5 - Isomerases	Intramolecule rearrangement, i.e. isomerization changes within a single molecule	AB → BA
EC 6 -Ligases	Join together two molecules by synthesis of new C-O, C-S, C-N or C-C bonds with simultaneous breakdown of ATP	X + Y+ ATP → XY + ADP + Pi

Table 2.0: Classification of Enzymes at EC level 1

Data Mining of Enzymes (DME) is our methodology in which we annotate each SP with a single, unique EC and utilize its hits on a queried sequence to provide an enzyme number prediction for the sequence of amino acids checked.
DME consists of two distinctive phases: Construction of SPs and utilization of the SPs to predict enzymatic function of proteins, metaproteomes or metagenomes. Construction of SPs is conducted using UniProtKB/Swiss-Prot (Swiss-Prot) [1.4] as the EC annotation oracle. Generation of SPs is a labor intensive process which lasts several weeks and has been conducted twice during the duration of this project. Once the process has completed, SPs can be used repeatedly.

2.2 Definition and Construction of the Specific Peptides

Specific Peptides (SP) are strings of amino acid motifs which are unique to a branch of the Enzyme Commission functional classification (EC). We show a sample of a few SPs in table 2.1 below. Each SP is assigned with a specific, unique EC number.

Specific Peptide	EC	EC Description
ELLAELFNIP	2.4.1.17	Glucuronosyltransferase.
KQFGHEY	2.7.4.22	UMP kinase.
LKDRLYT	6.1.1.5	Isoleucine--tRNA ligase.
MIDLVIGYTAIQ	4.1.1.39	Ribulose-bisphosphate carboxylase.
VVLQHQMP	3.1.1.1	Carboxylesterase

Table 2.1: Example of a few SPs from the Production SP set

Our current Production SP set contains 148,395 SPs. Figure 2.1 and 2.2 shows the number of SPs at various lengths.

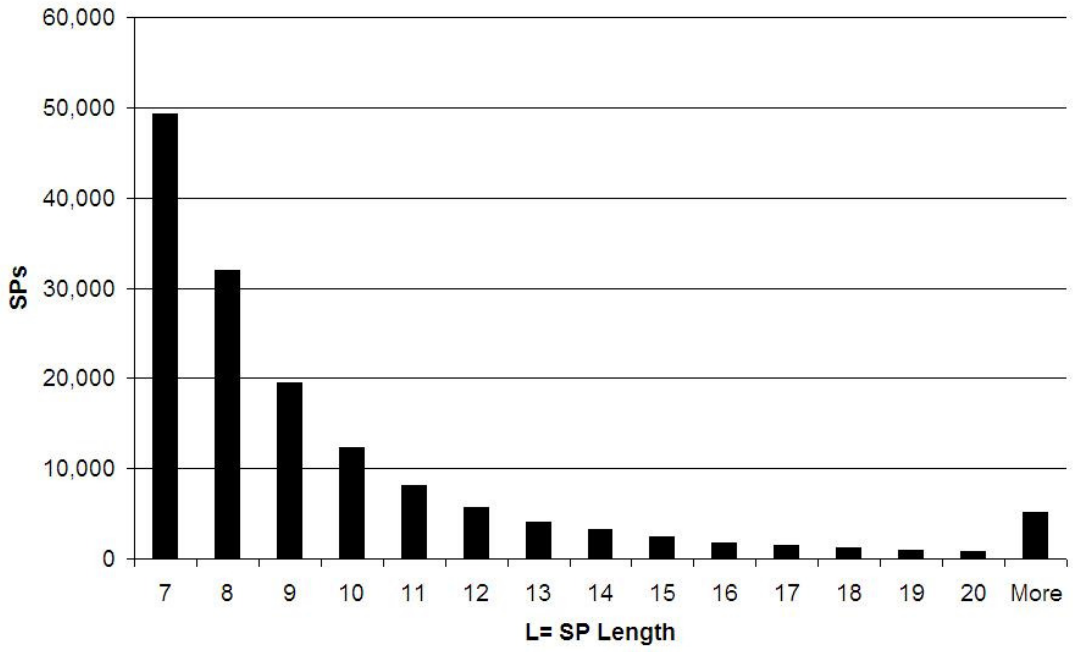


Figure 2.1: Histogram of the Production SP set by L=Length of the SP

Figure 2.2 shows the number of SPs at various lengths subdivided at the high EC number of the SP.

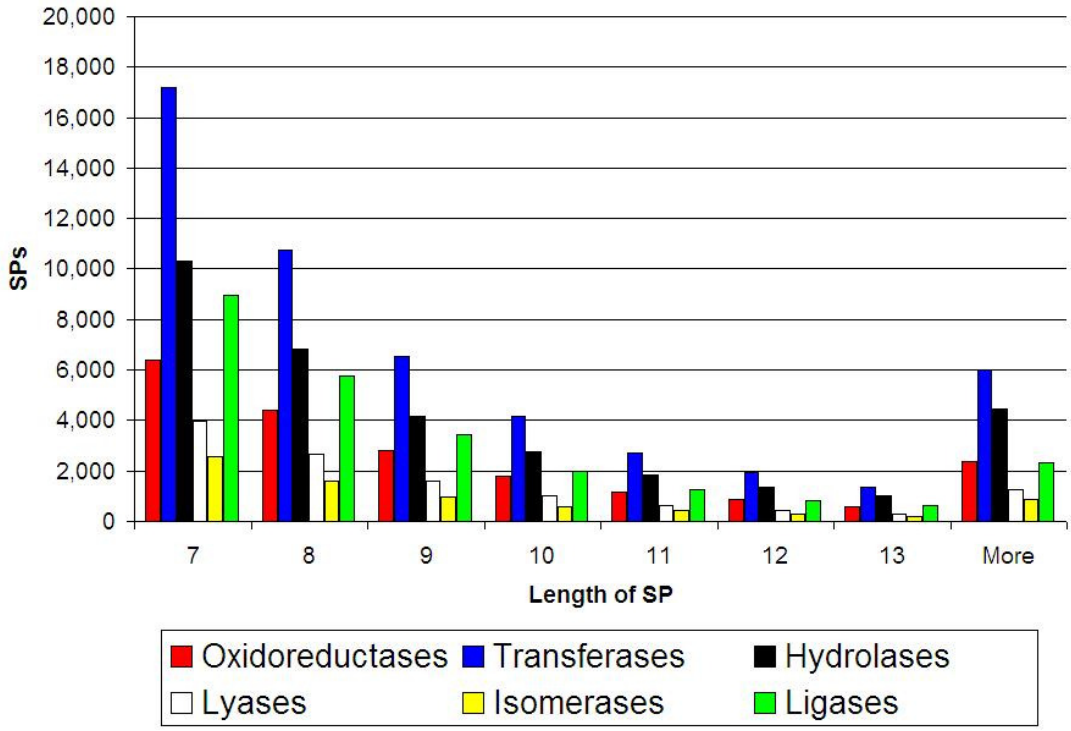


Figure 2.2: Histogram of the SPs according to the length of the SPs and high level of their EC number.

Construction of SPs is conducted using UniProtKB/Swiss-Prot (Swiss-Prot) [2.2] as the EC annotation oracle.
We use the Motif Extraction (MEX) algorithm [2.3]. Its input consists of sequences of amino-acids belonging to enzymes selected from Swiss-Prot. MEX [1.7] is a motif extraction algorithm used as a method for extraction of words from corpora of alphabetic strings. It is based on creating a super-graph whose vertices are the elements of the alphabet, and threading through it various strings of the corpus. A motif is chosen as such according to its multiple appearances in varying contexts within the corpus. MEX is applied here to find sequence-motifs within biological data. The alphabet is that of 20 amino-acids, and the strings are protein sequences.
MEX extracts motifs from proteins sequential data in an unsupervised manner.
MEX motifs are deterministic strings in contradistinction to position-specific weight matrices or regular expressions. MEX does not require any preprocessing of multiple sequence alignment nor does it rely on over-representation of k-mers. Moreover, the length of the motif is not pre-determined or constrained. The results of the MEX run are sets of motifs. In the context of the work presented below, the functionality of MEX consists solely to extract motifs from sequences of amino acids. In independent steps following MEX processing the resulting motifs are distilled using specificity criteria and the final product is a set of SPs. The entire process is described in detail in the "Construction of SPs" section below.
Swiss-Prot is used as the EC annotation source because it is manually curated and therefore more accurate than other sources. Another component of Uniprot, UniProtKB/TrEMBL (TrEMBL) consists of many putative, machine-generated annotations. As of August 2010, 519,348 (5%) proteins are annotated currently in

Swiss-Prot. 11,636,205 (95%) entries are annotated in TrEMBL. The EC annotations in TrEMBL are not manually-curated, and their accuracy and reliability is lower compared to the ones in Swiss-Prot. It is possible to use the whole set of UniProt annotations (Swiss-Prot and TrEMBL together). We refrained from including TrEMBL in the interest of increasing the accuracy of the predictive power of our predictions. Another consideration is that inclusion of TrEMBL would have required significantly greater resources to generate the SPs. This decision comes at the price of impacting negatively recall.
We found that utilization NCBI, Kegg [2.3], Brenda [2.4] or a combination of them with Swiss-Prot does not significantly increase the number of reliable annotations, and their addition would have multiplied the volume of the sequences to be processed. This in turn would have increased the computer resources needed to generate the SPs and rendered the process highly ineffective.

2.3 Construction of the SP sets

Construction of enzymatic SPs consists of a mixture of unsupervised and supervised procedures and involves the following steps:

Step	Description
1.	Selection of the Training set
2	Selection of subsets of enzymes from Swiss-Prot
3	Motif Extraction Utility
4.	Assignment of an EC to the MEX Motifs
5.	Elimination of nonspecific EC MEX motifs and promotion of remaining MEX motifs to “SPs”
6.	Optimization of the SP dataset

Table 2.2: List of steps required for the construction of SPs

Selecting the Training Sets

Training sets consists of subsets of singly-annotated enzymes in Swiss-Prot. Selection of the Training set has been done using as a discriminating parameter the field “Date Integrated into Swiss-Prot”. The process of generating new SPs has been run twice in the last few years. It was run first on 2006 with a training set of 89,854 singly annotated enzymes with a “Date-Integrated” annotation in Swiss-Prot before July 1st, 2006 (“Training Set #1”). The process was run a second time in 2009, with a training set consisting of 201,169 enzymes from singly annotated enzymes in Swiss-Prot with a “Date Integrated” indicator before July 29th 2009 (“Training Set # 2”).

We denote first the SP Production dataset generated in 2006 from Training Set #1 as “Production SP dataset V1.0”. Similarly, we denote the SP dataset generated from Training Set #2 in 2009 as “Production SP Dataset V2.0”.

Figure 2.3 shows the rapid growth of protein annotations in Swiss-Prot throughout the years.

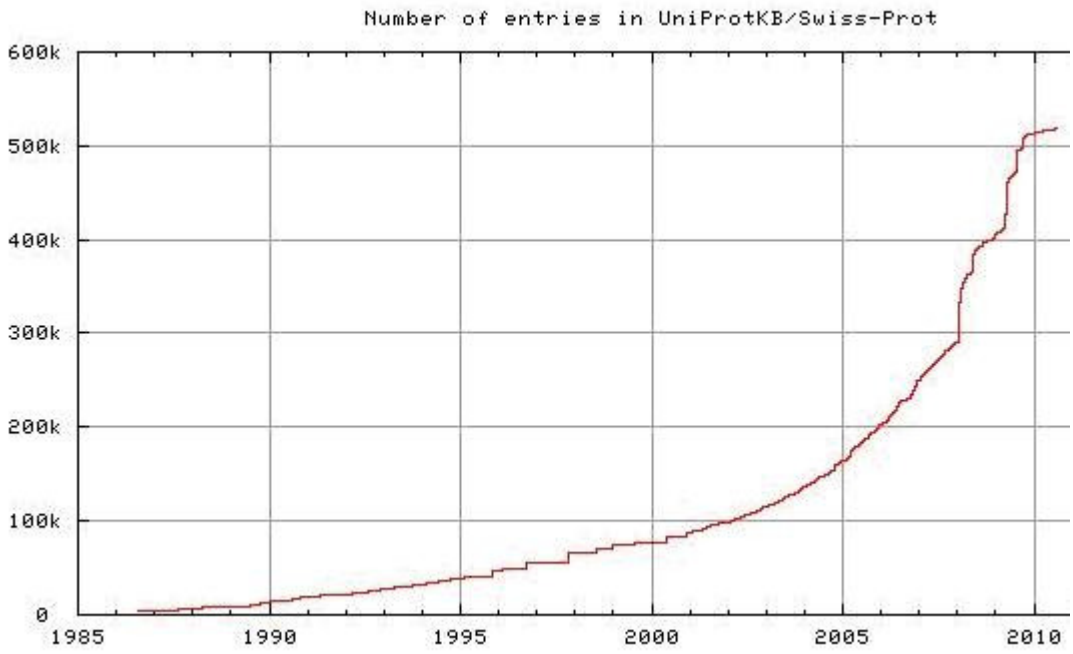


Figure 2.3: Growth of number of entries in Swiss-Prot throughout the years [2.14]

The rapid growth of Swiss-Prot annotations, shown in Figure 2.3, is the principal driving force to generate new sets of SPs.
Our goal is that new SPs will reflect the new Swiss-Prot annotations and thus increase precision and recall capabilities of our predictions.
We define precision and recall in the context of this work as follows:

$$PRECISION = \frac{N[P \mid P]}{N[P \mid P] + N[P \mid DP]}$$

$$RECALL = \frac{N[P \mid P]}{N[P \mid P] + N[P \mid DP] + N[NP \mid P]}$$

Where P|P represents the number of cases where the model prediction coincides with that of the expert, P|DP where the expert provides a different EC assignment, and NP|P where the model provides no prediction for enzymes whose EC assignments are given by the expert.

A year after generating the second SP dataset, we have seen sporadic cases of incorrect predictions when analyzing enzymes in Swissprot that belonged to Training Set #2. Those incorrect predictions result from EC numbers changed by Swiss-Prot after Training Set #2 was generated. We predict ECs according to the old annotation which has changed in the meantime. The remedial action to overcome this challenge is to generate SP sets from fresh Training Sets as often as possible.

Selection of subsets of enzymes from Swiss-Prot

This step can be run in supervised or unsupervised mode. It consists of selecting a number of subsets of a few thousand singly-annotated enzymes from Swiss-Prot to be processed by MEX. Multi-functional enzymes in Swiss-Prot (enzymes that have multiple ECs annotation per protein) are discarded.

In supervised mode, each one of the Swiss-Prot subsets consists of enzymes containing the same EC level 4 numbers. In unsupervised mode, each one of the subsets contains a few groups of enzymes with the same EC level 3 number, such that each selected subset contains a few thousand enzymes. We found that the set of SPs generated using the supervised selection method was larger by 30% than the unsupervised set. However, recall did not improve significantly using supervised selection. Instead, unsupervised selection required less computer resources and human manual labor minimizing total run time. Therefore, the current production set of SPs was generated using the unsupervised selection method. Using SPs generated in supervised selection mode is not a justified process for large-scale predictions because of performance considerations.

Motif Extraction Utility

The Motif Extraction Utility is an unsupervised step which consists of running the Motif Extraction algorithm against each previously selected enzyme subgroups. We used 6 subgroups of enzymes when constructing SPs V2.x. The input to each one of the MEX runs are the sequences of enzymes belonging to this subgroup and the output are motifs. It is important to note that MEX does not address EC number at all: Its function is to generate motifs from sequences provided. Assignment of EC numbers to the MEX motifs and review for their EC specificity (across all subgroups of enzymes) is done in separate subsequent steps and procedures, described below. The size of each selected group was limited only by the available machine resources. The constraining factors are memory and the CPU's power to run several MEX runs simultaneously, but typically the orders of magnitude are a few thousand enzymes per run. MEX motifs with lengths of lower than 5 amino-acids are discarded because they never meet specificity requirements.

Assignment of an EC number to the MEX Motifs

MEX Motifs are then assigned EC labels by searching for them in the Training Set and assigning them the EC annotation of the Swiss-Prot enzyme they hit. The search of the motifs harvested from MEX within the Training set consists of a simple search of patterns within text, which we optimized for performance using the Knuth Morris Pratt algorithm, described by Charras and Lecroq [2.13].

Elimination of nonspecific EC MEX motifs

Specificity for motifs harvested from MEX is determined as follows:
MEX Motifs that have a single EC level 4 label are promoted to the SP set and assigned to that particular EC level 4.
MEX Motifs that have different EC Level 4 labels but have the same EC level 3 number are promoted to the SP set and assigned that particular EC level 3.
MEX Motifs that have different EC Level 3 and 4 labels but have the same EC level 2 number are promoted to the SP set and assigned that particular EC level 2.
MEX Motifs that do not have the same EC Level 2, 3 and 4 labels but have the same EC level 1 label are promoted to the SP set and assigned that particular EC level 1.
All the other MEX motifs are deemed non-EC specific and are discarded.
At this point the resulting motifs are considered SPs and we proceed to optimizing the resulting SPs dataset.

2.4 Comparison of SP dataset V1.0 to SP dataset V2.0

We are interested to compare set V1.0 and V2.0 not only because of the mere growth from V1.0 to V2.0 but to observe their internal structure: Their composition by SP length and by the EC level assigned to each SP.

The size of the SP set increased from 87,017 in SP set V1.0 to 312,465 SPs in SP set V2.0. Figure 2.4 below shows comparative cross sections of SP sets V1.0 and V2.0. While the number of SPs more than tripled, their breakdown by EC level remains constant.

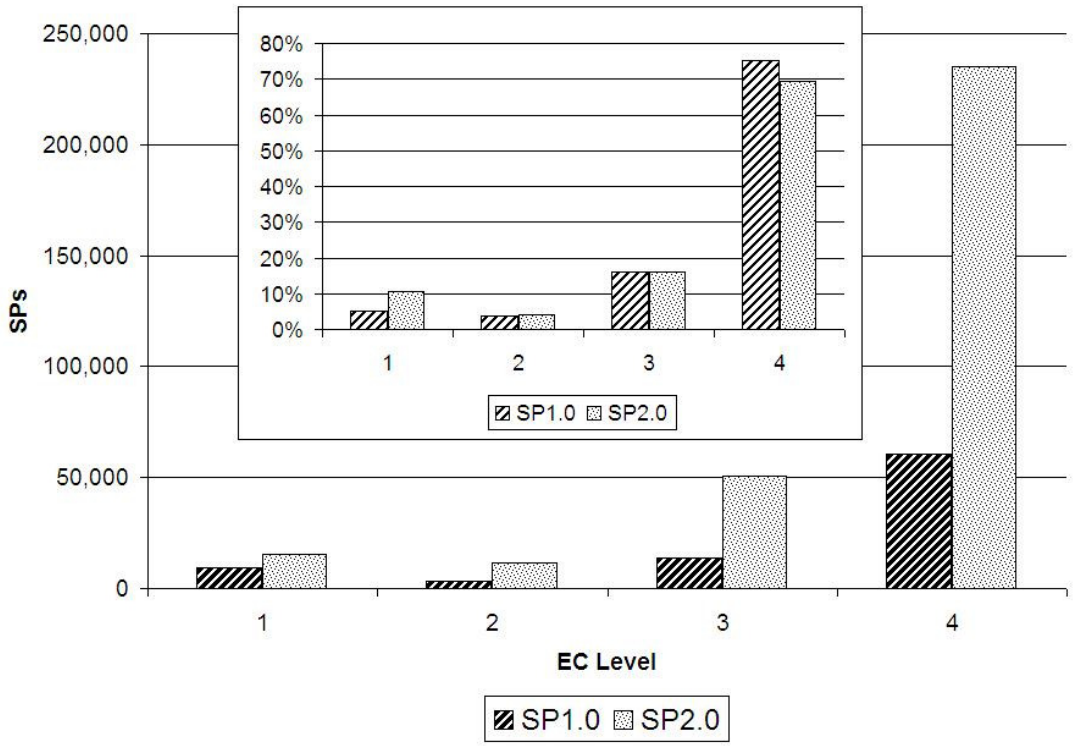


Figure 2.4. Comparative histograms of SPs V1.0 (2006) and V2.0 (2009) by EC level.

Figure 2.5 below shows that the general distribution pattern of the SP lengths remained the same. Analysis of the histogram showing normalized values shows that the proportional number of SPs with lengths L=6 through L=9 is greater in the set SP V1.0 as compared to SP set V2.0. The trend reverses for L > 9. This difference can be explained by the fact that we utilized different MEX parameters in each case.

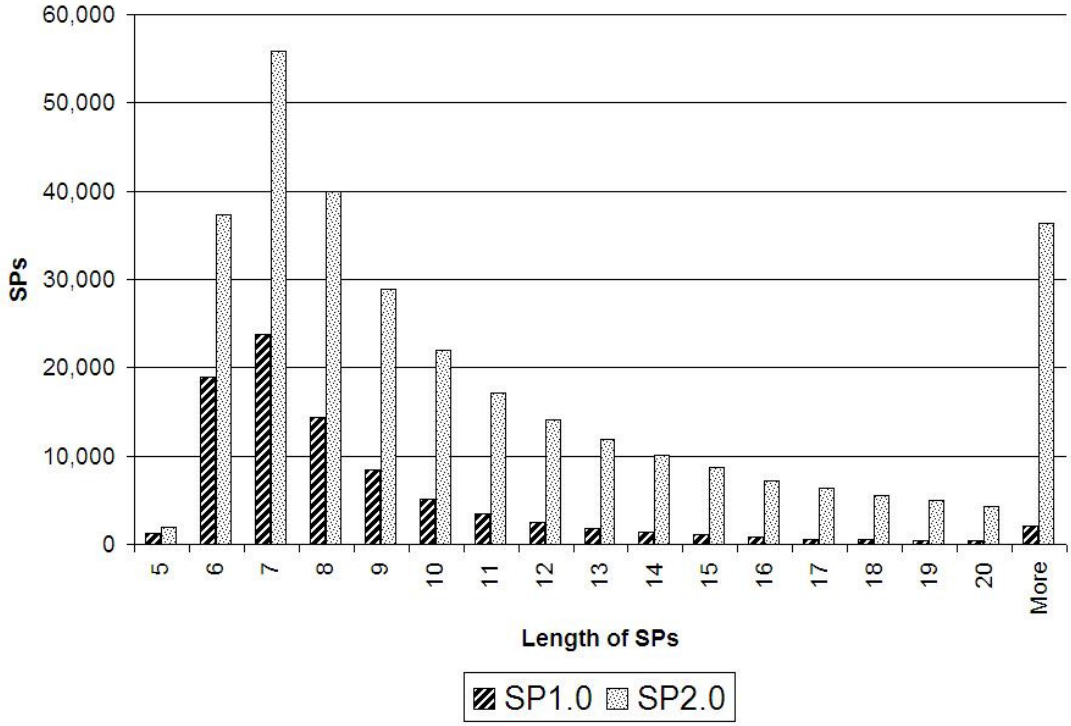


Figure 2.5: Comparative histogram of SPs V1.0 and V2.0 by SP length

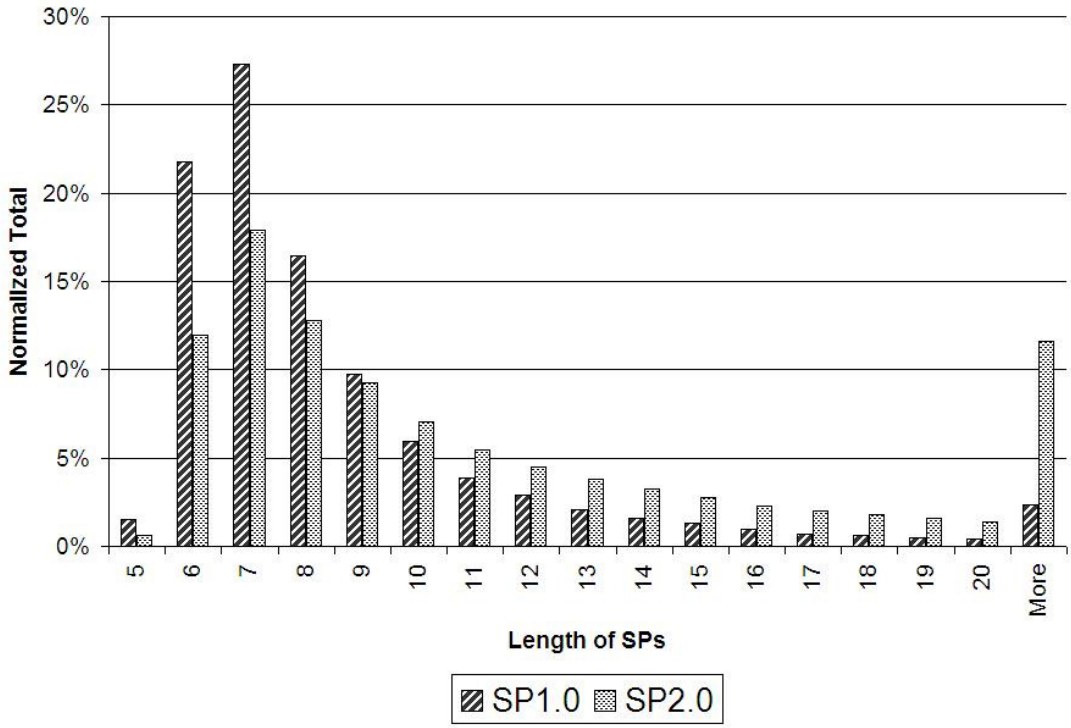


Figure 2.6: Comparative histogram of SPs V1.0 and V2.0 by SP length in normalized values

2.5 Optimization of SP dataset V2.0

The dramatic increase from 88,017 SPs in SP set V1.0 to 312,465 SPs in SP set V2.0 motivated us to research optimization methods designed to minimize run time of large scale metaproteomic and metagenomic predictions. That is because increase in number of SPs has a significant negative impact on resources required to generate predictions. This optimization is crucial because as the number of Swiss-Prot annotations will increase in the future, so will the corresponding SP dataset. Optimization of the SP dataset entails removing a subset of SPs. Our goal was to maximize the performance benefits reaped from the trimming while making sure, at the same time, to limit the negative impact to precision and recall values.

We conducted the following optimization actions on SP set V2.0:

1. Removed all SPs assigned an EC level 1 and EC level 2 from the SP set. The rationale is that predictions at EC level 1 and EC level 2 are not as significant as predictions at EC level 3 or EC level 4.
2. Discarded all SPs with length $L < 7$ amino-acids. The rationale behind this trimming is that SPs with length 5 or 6 amino-acids contribute significantly to accidental hits and thus to false predictions. The threshold of $L < 7$ was derived by observing the statistics of random hits on tests performed and will be discussed below in a separate chapter.

Table 2.3 below shows a summary of the intermediate SP sets built as part of the optimization process of SP V2.0.

Version	Description	Number of SPs in the Set
V1.0	SP set built in 2006 – Based on Training set #1	87,017
V2.0	SP set built in 2009 – Based on Training set #2	312,465
V2.1	SP dataset based on V2.0 discarding all SPs with length less than 7	273,186
V2.15	SP dataset based on V2.1 discarding all "containing" SPs	159,775
V2.3	SP set based on SP set V2.0 – discarding SPs with $L \leq 6$ and SPs with EC levels 1 and 2 and "containing" SPs.	148,395
V2.4	Based on SPs V2.3 and V1.0 that meet specificity criteria. Added SPs from V1.0 to V2.3. The process will be described in detail below.	170,491

Table 2.3: Summary of the intermediate SP sets built as part of the optimization process of SP V2.0

Qualitative analysis of the optimization

SP set V2.3 is considered at the present our Production SP set. We must ascertain the impact of trimming 164,070 SPs from set V2.0 to generate set V2.3.
Before presenting this analysis, we must present three definitions which are central to our methodology and assist comparing the performance of these two datasets:

Consistent SPs are SPs whose EC belongs to the same branch in the EC number tree. For example: Three different SPs with EC=2.4, 2.4.1 and 2.4.1.227 are consistent.

Consistent coverage of SPs on a queried sequence is the total number of amino-acids that are in agreement with all consistent SP hits.
Prediction Threshold Coverage is the minimum consistent coverage to generate a prediction on a queried amino-acid sequence
These metrics will be discussed in detail in the following chapters.
A brief illustration is shown using Table 2.4 below, which consists of all SP hits (SP V2.3) on Swiss-Prot enzyme B9DIX2 - AMPA_STACT, LAP, Leucine aminopeptidase for the bacteria Staphylococcus carnosus (strain TM300), which is annotated in Swissprot with two ECs: EC=3.4.11.1 (Leucine aminopeptidase) and EC=3.4.11.10 (Bacterial leucyl aminopeptidase)

SP hit Location within the enzyme	Hitting SP	EC of Hitting SP
338	EVLNTDAEGR	3.4.11
341	NTDAEGRL	3.4.11
336	TVEVLNTDAEGRL	3.4.11.1
370	TLTGAAVA	6.1.1.14

Table: 2.4: SP hits (SP V2.0) on Swissprot enzyme B9DIX2.
The first three SPs have consistent ECs with coverage L=13 amino-acids at EC level four for EC=3.4.11.1 (Leucyl aminopeptidase)
The fourth SP has a coverage of 8 amino-acids for EC=6.1.1.14 (Glycyl-tRNA synthetase).
As an example of the impact of choosing different prediction thresholds we select two different prediction thresholds and inspect the corresponding results. We select first a threshold coverage of L=13 amino-acids and provide a prediction of EC=3.4.11.1 (consistent with the Swissprot annotation). In this case, we regard the fourth SP hit an accidental hit. Setting up the prediction threshold at L=7, we predict a double function enzyme: EC=3.4.11.1 and EC=6.1.1.14.
An important conclusion of this work (see below) is that the optimal prediction threshold for proteomic searches is L=7. Increasing the prediction threshold increases precision but could affect recall negatively. We will discuss in detail in the following chapter the optimization of the prediction threshold.

We compare the quality of SP set V2.0 and SP set V2.3 by calculating precision and recall values using one of our test sets, "Test Set #3" which consists of one thousand random annotated Swiss-Prot enzymes with integrated date after July 27th 2009, a single EC annotation at EC level 3 or EC level4.
Figure 2.5a below shows the resulting precision curve using SP sets V1.0, V2.0, V2.3 and V2.4 varying L=prediction threshold coverage.

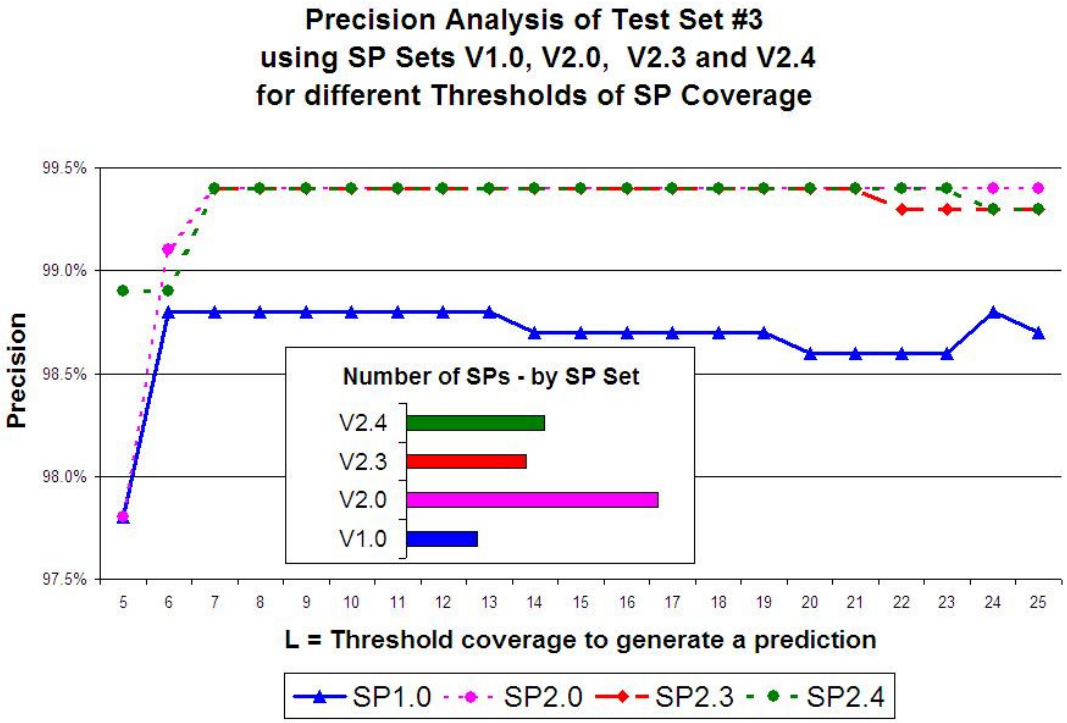


Figure 2.5a: Precision as function of L=threshold coverage to generate a prediction for SP sets V1.0, 2.0, 2.3 and 2.4 for Test Set #3.

The inset in the figure represents the number of SPs by SP set. Given the construction of SPs (No complete mutual inclusion between SPs within a specific set and EC specificity) we expect and see that performance of an SP set improves with increasing numbers of SPs in the set.
For threshold prediction coverages $L \geq 8$ precision power of sets V2.0, V2.3 and V2.4 are identical even though the size of the SP set V2.3 is less than half of set V2.0 and 13% smaller than SP set V2.4.

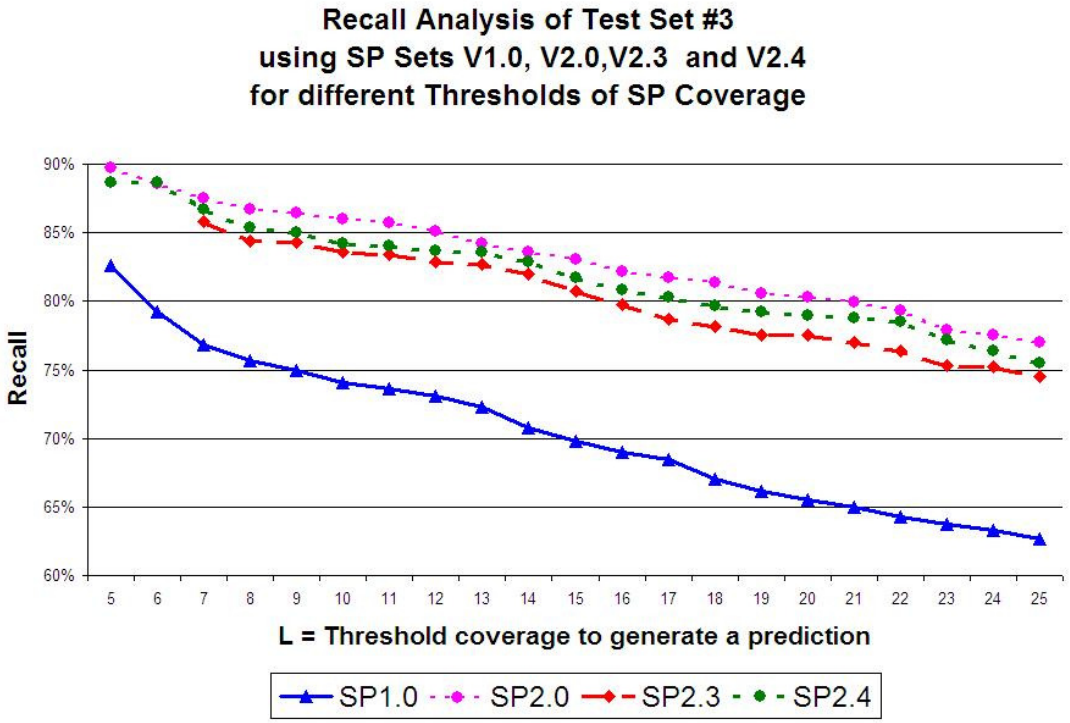


Figure 2.5b: Recall as function of L=threshold coverage to generate a prediction for SP sets V1.0, 2.0, 2.3 and 2.4 for Test Set #3.

From Figure 2.5b we see the impressive increase by more than 10% in recall values going from set V1.0 to set V2.0 V2.3 and V2.4. In addition, we see the impact of trimming set V2.0 by half is minimal – around 2%.

The performance implications of reducing the SP set by half are highly significant, bearing in mind that many of our analyses consist of large volume processing. An example of the scale of magnitude of volumes processed is the predictions on the Sargasso Sea metaproteome [2.6], which consisted of 1,001,986 sequences which had to be checked for occurrence of every one of the SPs.
Because of performance considerations, we consider today SP set V2.3 our Production SP set.
Potential uses of SP sets V2.0 and V2.4 consist of analysis of small samples of data, where performance is a minor consideration. Precision and recall results in these cases would be better than using SP set V2.3.

Unless otherwise noted in this work, this is the SP dataset used for analyses. It consists of 148,395 SPs with EC level 3 and 4 whose SP length $L \geq 7$ amino-acids.

Table 2.4a below shows an analysis of the number of SPs added and lost between SP sets V1.0 and V2.3.

Description	Number of SPs
Total SPs in set V1.0	88,017
Total SPs in set V2.3	148,395
SPs common to SP set V1.0 and SP set V2.3	38,067
SPs in SP set V1.0 and not in SP set V2.3 (Lost)	48,950
SPs in set V2.3 and not in SP set V1.0 (Gain)	110,328

Table 2.4a: Analysis of number of SPs added and removed from SP set V1.0 when building SP set V2.3

We see from Table 2.4a that a very large number of SPs (56%) were lost from SP set V1.0 going on to SP set V2.3. There are three factors that can result in that loss:
1. Different MEX runs building the motifs could have generated a motif when building SP set V1.0 and that particular motif was not generated running MEX for SP set V2.0
2. Change in Swissprot annotations cause SPs in set V1.0 to change their EC annotations
3. Larger data sets may have new occurrences of old motifs on different sequences, causing previous SPs to lose their specificity.

Going from set V1.0 to set V2.3, we see an increase of the number of SPs by a factor of 3.6 upon a 2.2-fold increase in input protein number. This can be explained by the fact that Swissprot is hand curated and thus adds to its repository strongly homologous enzymes and therefore an enrichment of 2.2 in the number of enzymes generates more redundancy than adding non-homologous enzymes, which in turn, enables Mex to mine more motifs that ultimately become Specific Peptides.

2.6 Addition of existing production SPs to a newly created SP dataset

Starting from the second production release of SPs, we can take advantage of prior production releases of SPs and add them to the new production set, as long as they meet the following EC specificity and inclusion criteria requirements:

1. The EC Level of the additional SPs is three or four (EC Level one and two provide much less insightful information, hence we focus on EC levels three and four only)
2. Additional SPs from old Production release do not fully include any SP from the new production set of SPs and meet EC specificity requirements within the new training set of enzymes

The rationale for increasing the size of the SP Production set is to increase the predictive capabilities of the set, mainly recall figures.
Using this strategy, we generated SP set V2.4 from the second production SP set, V2.3. Production set V2.3 was expanded from 148,395 SPs to 165,095 SPs using the 87,017 SPs belonging to SP V1.0.

Table 2.5 below shows that SP set V1.0 contributes 13% of the total SPs to the newly created SP set. Of the 87,017 SPs in set V1.0, 21,336 SPs (25%) are carried to the new dataset; 4,636 SPs from the set V2.3 (3.1%) are disqualified.

EC Level	SPs from V1.0	SPs from V2.3	SPs from V1.0	SPs from V2.3
3	3,532	26,863	2%	16%
4	17,804	116,896	11%	71%
Total	21,336	143,759	13%	87%
	165,095		100%	

Table 2.5: Contributions of SP set V1.0 and V2.3 to SP Set 2.4

2.7 Utilization of SP exact and fuzzy matches for DME predictions

MEX extracts exact motifs from the training set of enzymes. Motifs are subsequently distilled to include only ones that meet EC specificity criteria.
We researched whether there is a potential to expand the scope of SP utilization by taking advantage of fuzzy matching using BLOSUM matrices (18). The benefit of this expansion could be increasing recall of predictions.
One part of the expansion can be done by introducing modifications to MEX so that it generates sets of motifs that are not exact but support fuzzy similarity between them up to a certain preset BLOSUM threshold. Such expansion was out of the scope of the present work.
We researched expanding the comparison of SP hits from exact hits to fuzzy hits using BLOSUM matrices [2.8]. This was performed using the "BL2SEQ" utility [2.9] which compares two input sequences directly in order to assess the similarities between them.
In order to quantify the quality of predictions using fuzzy matching we tested 99 random enzymes using SPs V1.0. We searched for each one of the 87,018 SPs within each one of the 99 test enzymes using BL2SEQ.
Table 2.6 below shows analysis of the SP hits on enzyme A2ZAB5 - Serine/threonine-protein kinase SAPK3 which is annotated in Swissprot with an EC=2.7.11.1. Each row shows the SP that BL2SEQ found as the highest scoring segment, the EC of the SP, the length of the identical sequence and conserved sequence. The table is sorted by the difference between the length of the SP and the length of the identical sequence. We consider the SP first entry to be the "BL2SEQ SP prediction" – in this case EC=2.7.11.1, in agreement with the Swissprot annotation.

SP	EC of Nearest SP (Prediction)	L SP	L Identical	L Conserved	eValue	High Scoring Pair Alignment
VGTPAYIAPEVLSRREYDGK	2.7.11.1	20	19	20	0	VGTPAYIAPEVLSR+EYDGK
PKSTVGTPAYIAPEVL	2.7.11.1	16	16	16	0	PKSTVGTPAYIAPEVL
CHRDLKLENTLLDGS	2.7.11.1	15	15	15	0	CHRDLKLENTLLDGS
PRLKICDFGYSKSS	2.7.11.1	14	12	14	0	PR+KICDFGYSKS+
ADVWSCGVTLYVM	2.7.11.1	13	13	13	0	ADVWSCGVTLYVM
MAYSTVGTPDYIAPEIF	2.7.11.1	13	11	12	3E-05	STVGTP YIAPE+
RICNAGRFSEDE	2.7.11.1	12	10	11	7E-05	+IC AGRFSEDE
RICSAGRFSEDE	2.7.11.1	12	10	12	5E-05	+IC+AGRFSEDE
TSTFCGTPNYIAPEILRG	2.7.11.13	11	9	10	0.0004	GTP YIAPE+L
VHRDLKPENLLLASK	2.7.11.17	11	9	9	0.002	HRDLK EN LL
VMELCAGGELF	2.7.11.1	11	9	9	0.001	VME AGGELF
HLAIVMEYA	2.7.11.1	9	9	9	0.0007	HLAIVMEYA
REIINHRSL	2.7.11.1	9	9	9	0.0007	REIINHRSL
RFSEDEAR	2.7.11.1	8	8	8	0.003	RFSEDEAR
EYAAGGE	2.7.11.1	7	7	7	0.005	EYAAGGE
YIAPEVL	2.7.11	7	7	7	0.008	YIAPEVL

Table 2.6: Analysis of SP hits on enzyme A2ZAB5 - Serine/threonine-protein kinase SAPK3 using BL2SEQ.

Aggregating the results of this analysis at EC level 3 for the 99 random test enzymes selected and comparing the results to predictions using exact matches we find the following results:

	a	b	d	e		
	TP	**FP**	**DME NP**	**DME NP and Swissprot NP**	**Precision**	**Recall**
Exact Matching	73	11	9	6	87%	78%
Fuzzy matching using BL2SEQ	62	11	21	5	85%	66%

Table 2.7: Precision recall analysis for 99 test enzymes using exact and fuzzy SP matching.

Fuzzy SP matching is computationally costly and does not contribute to precision or recall and therefore most of our studies were conducted using exact SP matching. However, we should not discard totally this method and should be used as a second layer prediction tool, as there are some cases where it can provide additional insight to the exact SP predictions. As an example of such case we show analysis of an enzyme

from Test Set #2: Enzyme A0KIC9 - BIOD_AERHH, Dethiobiotin synthetase is annotated in Swissprot with an EC=6.3.3.3.
With exact matching, we find two SP hits using the Production SP2.3 set with a consistent coverage of 12 amino acids, as shown in table 2.8 below:

SP	SP Location	EC of the SP	EC Description
FVTGTDT	6	6.3.3.3	Dethiobiotin synthase
GTDTDVGKT	9	6.3.3.3	Dethiobiotin synthase

Table 2.8: Exact SP hits on Swissprot enzyme A0KIC9 belonging to Test Set #2.

The amino-acid of this sequence looks as shown in figure 2.6 below - we highlight the amino acids matching the SPs.

```
MVKSFFVTGTDTDVGKTLVARTLLLEFAAHGLRCAGYKPISAGCARTPDGLRNLDAVLLQ
EAASLPLPYDLVNPYAYEPPIAPHIAASEARDAITLKGLSDGLRQIEQAGAELVVVEGAG
GWFLPLDRKHLLSDWVKQENMPVIMVVGAKLGCLNHALLTFAAIRNNNLPVAGWVINRLH
GSMSHYQENLDTLRGLLPAPFLGEIPFVNNPLEADLRGRLDISPLL
```

Figure 2.6: Exact SP hits on enzyme A0KIC9

Using the fuzzy matching method, we search each one of the 148,395 SPs from the SP2.3 Production set within this enzyme using BL2SEQ. This search produces 17 fuzzy hits as shown in Table 2.8 below. EC of fifteen of these SPs coincide with the EC annotated for this enzyme – 6.3.3.3 and two do not. The matched coverage for EC=6.3.3.3 is 85 amino acids and the matched coverage for other hits is 15 amino acids.

SP	EC	Matched inexact sequence
FITATGTDIGKTYVTALIIK	6.3.3.3	F+T T TD+GKT V ++
GTDTDVGKT	6.3.3.3	GTDTDVGKT
GTDTEIGKT	6.3.3.3	GTDT++GKT
NPYTFAEPTSPHI	6.3.3.3	NPY + P +PHI
EICPYSIEEPLAPRLAMKRAGR	6.3.3.3	PY+ E P+AP +A A
PAIAPHLAAREAGVELSAARLH	6.3.3.3	P IAPH+AA EA
QLLQAGAEMVKIEGAG	**2.1.2.11**	Q+ QAGAE+V +EGAG
LVRERGADLVVIEGMGRA	**2.7.1.33**	+ + GA+LVV+EG G
NDIKKLFIEGAGGLMVPLNEQDTWLDFLKLTRIPVILVVG	6.3.3.3	+EGAGG +PL+ + D++K +PVI+VVG
EGAGGWFTPLS	6.3.3.3	EGAGGWF PL
EGAGGWRVP	6.3.3.3	EGAGGW +P
PVVLVVGVRLGCI	6.3.3.3	PV++VVG +LGC+
LVSAIKVGCINHTLLTINEL	6.3.3.3	+V K+GC+NH LLT
RLGCISHALLT	6.3.3.3	+LGC++HALLT
GCINHALLT	6.3.3.3	GC+NHALLT
PLAGWVANRIDP	6.3.3.3	P+AGWV NR+
VDPATSRLEENLATLAERLPAPCLGRVPRL	6.3.3.3	S +ENL TL LPAP LG +P

Table 2.8: Fuzzy SP hits on enzyme A0KIC9

Figure 2.8 below shows the mapping of the fuzzy hits on the enzyme. Highlighted in green the identities and positive matches for EC=6.3.3.3 that coincide with the exact SP matches – both with the correct prediction EC.
Highlighted in blue are the fuzzy matches that coincide with the annotated EC that were not discovered using exact matching.
Highlighted in yellow the identities and positive matches with other ECs - accidental hits.

MVKSF**FVTGTDTDVGKT**L**V**AR**TL**LLEFAAHGLRCAGYKPISAGCARTPDGLRNLDAVLLQ
EAASLPLPYDLV**NPY**A**YEPPIAPHIAA**S**EA**RDAITLKGLSDGLR**QI**E**QAGAELVVVEGAG**
GWFLPLDRKHL**L**S**DWVK**QENM**PVIMVVG**A**KLGCLNHALLT**FAAIRNNNL**PVAGWV**I**NRL**H
GSM**S**H**YQENL**D**TL**RGL**LPAP**F**LG**E**IP**FVNNPLEADLRGRLDISPLL

Figure 2.8: BLSEQ2 fuzzy SP hits on enzyme A0KIC9.

This figure demonstrates a benefit of using fuzzy matches: large areas of similarity which were not discovered using exact SP hits were discovered fuzzy matches. The price is introducing accidental SP hits, which would decrease precision.

An aspect that has weighty implications on the feasibility of utilization of fuzzy matching is performance.
Elapsed time to conduct exact matching for enzyme A0KIC9 against all 148,395 SPs totaled 3 seconds. The same process with fuzzy matching using BL2SEQ on the same production server took six hours and eighteen minutes which is 7,560 times more than the exact matching process.

2.8 Annotated Specific Peptides – ASPs

46% of the 201,169 enzymes in the Training Set #2 carry annotations in Swiss-Prot of "Active site," "Binding site" and "Metal binding site" at specific locations of single amino acids. As stated by Meroz and Horn [2.10], SPs cover these functionally important sites significantly more than other loci on proteins, indicating biological significance of these SPs.
SP matches that overlap such sites are compiled and the corresponding SPs are denoted as Annotated Specific Peptides (ASPs). We compiled a list of 27,457 ASPs. All ASPs appear at least four times in the training set, and the location of the annotation is consistent in the different appearances. Most ASPs carry single annotations (7,442 active sites, 9,247 binding sites and 8,567 metal binding sites); 2,171 ASPs carry two annotations; and 30 ASPs carry all three annotations.
We annotate each one of the ASPs with its EC, Activity function (Active Site, Binding or Metal Binding) and a bit map of the location of the active amino-acid marking the active location.
Table 2.9 below shows an example of two ASPs.

ActSP	Bit Map pointing to Active Site	EC	Active Site Description	Bit Map pointing to Metal Site	Metal Site Description
HEIDHLNG	0**1**000000	3.5.1.88, Succinyl-diaminopimelate desuccinylase		**1**0000000	Iron
LITSD**EE**G	00000**1**00	3.5.1.18, Succinyl-diaminopimelate desuccinylase	Proton acceptor	0000000**1**	Cobalt or zinc 2

Table 2.9: Example of two ASPs

Table 2.10 below shows the distribution of ASPs according to their function

Active sites	Metal Binding sites	Binding sites	Number of Annotated SPs
		x	9,871
	x		9,003
	x	x	865
x			7,777
x		x	731
x	x		753
x	x	x	34

Table 2.10: Number of ASPs by Annotation Type in SP set V2.3

Using only the annotated subset of the SPs dataset produces high precision predictions but severely impacts recall. The reason is that active, binding and metal annotations for enzymes in Swiss-Prot exist for only half of the training set and therefore the resulting Annotated set cannot represent the full information that exists in the whole Production SP set. We use the ASPs as a supplementary source of information within our online web tool, http://adios.tau.ac.il/DME which displays SP and ASPs hits on a queried sequence of amino-acids. Figure 7.2 shows sample of analysis of an enzyme using our online system demonstrating display of Active, Metal and Binding sites.

2.9 GSPs – Gene Ontology based SPs

The gene ontology GO [2.11] covers three domains: cellular component, the parts of a cell or its extracellular environment; molecular function, the elemental activities of a gene product at the molecular level, such as binding or catalysis; and biological process, operations or sets of molecular events with a defined beginning and end, pertinent to the functioning of integrated living units: cells, tissues, organs, and organisms.
We built SPs, that are also specific to the Biological Process annotations of GO - GSPs, which contain biological process information. The objective of building such SPs is to generate biological process predictions based on their hits on queried sequences.

Similarly to the case of Family SPs and Annotated SPs, GSPs is a subset of the Production SP set to which we assigned expanded attributes, in this case, biological process information. The procedure required to build the GSPs consists of searching for SPs from the Production SP set within the Training Set. A large portion of the enzymes hit contain biological process annotations and those are inferred to the hitting SPs. Filtering is conducted so that the SPs selected belong uniquely to a chain of biological processes. Each GSP is assigned the whole vertical branch of the gene-ontology tree which contains the leaf hit by the GSP. The purpose of this assignment is to provide the capability to generate cross sections at different branch levels of the gene-ontology tree. The Production GSP set consists of 30,452 GSPs. Table 2.11 below demonstrates a sample of three random GSPs showing the Biological GO processes assigned to them. For each of the GSPs, we highlighted the biological process documented in Swissprot for the Training Set enzyme hit by the GSP. We derived all other leaves in the vertical branch and assigned them to the GSP using the tree structure provided in the Gene Ontology Website [2.11].

Sample GSP #1		Sample GSP #2		Sample GSP #3	
FTAGVGE		**PGADPEVRAA**		**DTRELDR**	
Process Number	Process Description	Process Number	Process Description	Process Number	Process Description
8152	Metabolic process	8152	Metabolic process	8152	Metabolic process
44237	cellular metabolic process	44237	cellular metabolic process	44237	cellular metabolic process
6793	phosphorus metabolic process	6139	nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	51186	cofactor metabolic process
6796	phosphate metabolic process	16070	RNA metabolic process	6732	coenzyme metabolic process
16310	**phosphorylation**	34660	ncRNA metabolic process	9108	coenzyme biosynthetic process
		6399	tRNA metabolic process	19363	pyridine nucleotide biosynthetic process
		8033	**tRNA processing**	**9435**	**NAD biosynthetic process**

Table 2.11: Three random GSPs – highlighted entries show the GO biological process for the Training Set enzyme hit by the GSP.

Because the SP hits are confined strictly to enzymes in the training set, the set of biological processes inferred to GSPs are limited to metabolic processes only, which is only a partial set of the branches of the entire biological processes tree.

We present below a sample of analysis showing the distribution of GSP hits at level 5 of the GO ontology tree for two metagenomes.
The Soudan Red Mine sample data [2.12] consists of the total microbial community taken from the oxidized sediments of the Soudan Mine (Minnesota).
The Rios Mesquites Stromatolites bacteria sample [2.12] consists of a microbial community isolated from the Rios Mesquites microbiolite in Cuatro Cienagas, Mexico.
We assume that highly represented GSP hits reflect important metagenomic biological functions and thus we can use such spectrums to conduct comparative studies of the biological functions between different metagenomes.

We illustrate this point in figure 2.9 below which shows the comparison of the number of SP hits for two different metagenomic samples.

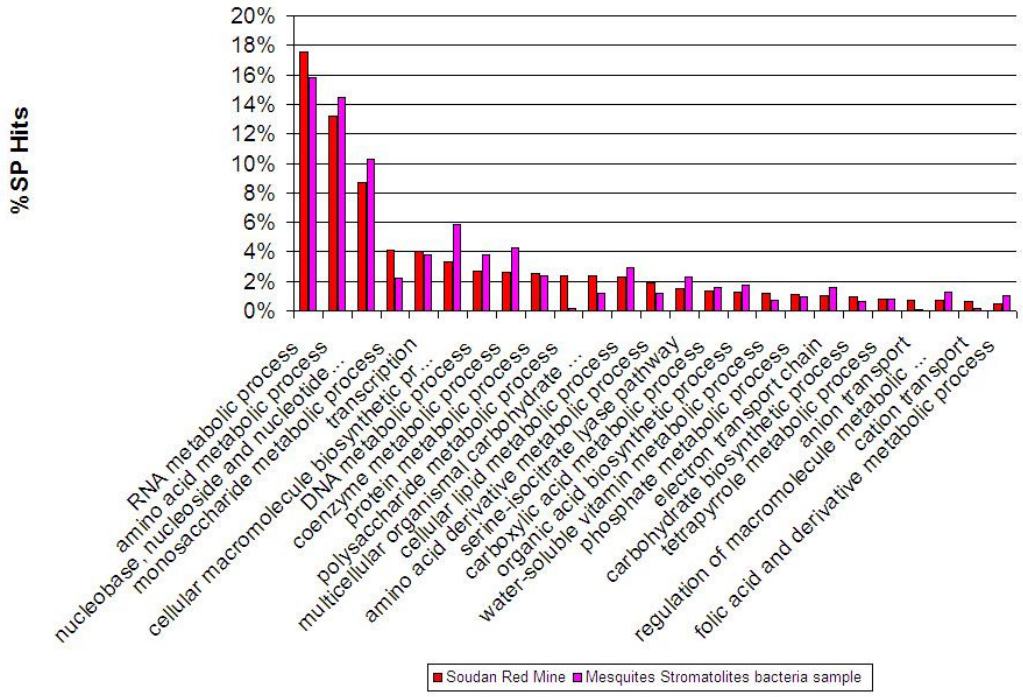


Figure 2.9: Histogram of GSP hits at level 5 of the GO ontology tree for Red Soudan Mine and Mesquites Stromatolites bacteria sample metagenomes.

2.10 Family Specific Peptides - FSPs

Swissprot assigns to each enzyme a "UniProtKB/Swiss-Prot entry name".
Every entry name is composed of two elements:

1. A mnemonic protein identification code which consists of the recommended protein name or gene name
2. Identification of the organism which is the biological source of the protein

We tag each of the SPs in the Production SP Set, SP V2.3, with the gene name of the enzyme it hits in Training Set 2 and select only SPs that hit only genes that have the same gene name, i.e. belong to the same protein family. The resulting set is labelled "Family SPs - FSPs".
Figure 2.10 below shows the leading 40 groups of enzymes by gene names in Training Set 2. The inset shows the behaviour of the distribution for the first 2,000 genes.

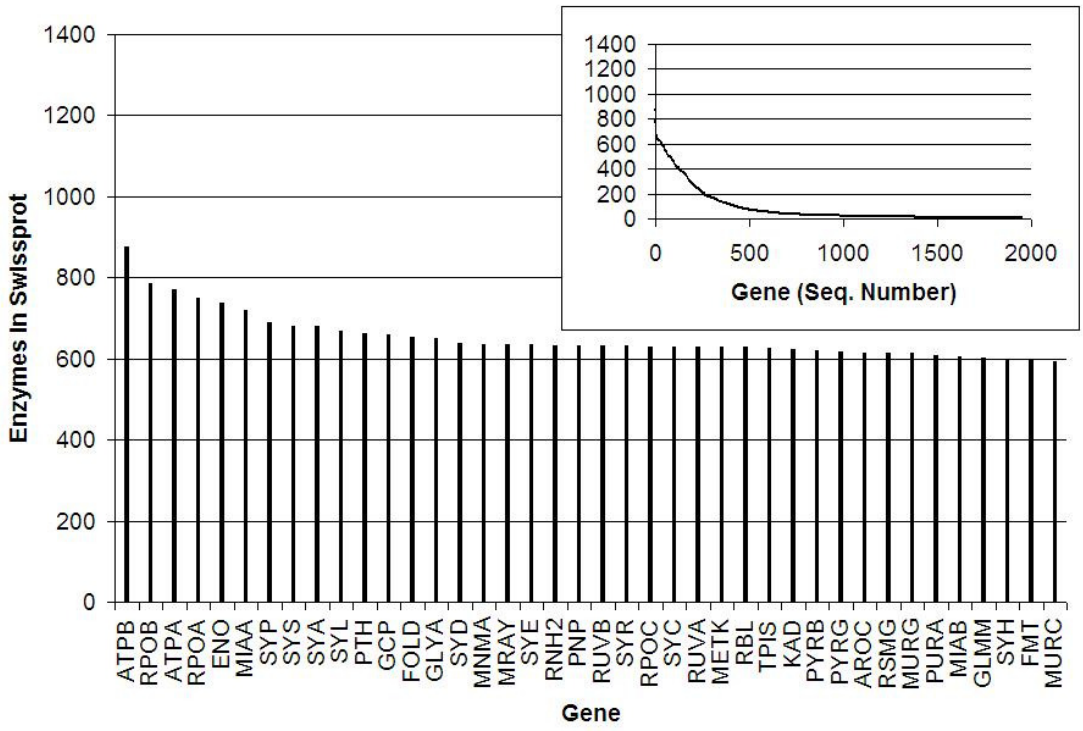


Figure 2.10: Histogram of the number of enzymes in Training Set 2 of Swissprot by gene name.

104,981 SPs (71% of all SPs) are assigned a specific gene tag and are considered FSPs.
Figure 2.11 below shows the histogram FSPs by gene name. The inset shows the behaviour of the distribution of FSPs by gene for the first 1,000 genes.
The behaviour of the distribution of enzymes by gene name as shown in figure 2.11 is similar to the behaviour of the distribution of FSPs by gene. The reason for that is that if a gene is highly represented in the Training Set, it is expected that the FSPs derived for this genes will be highly represented in the FSP set.

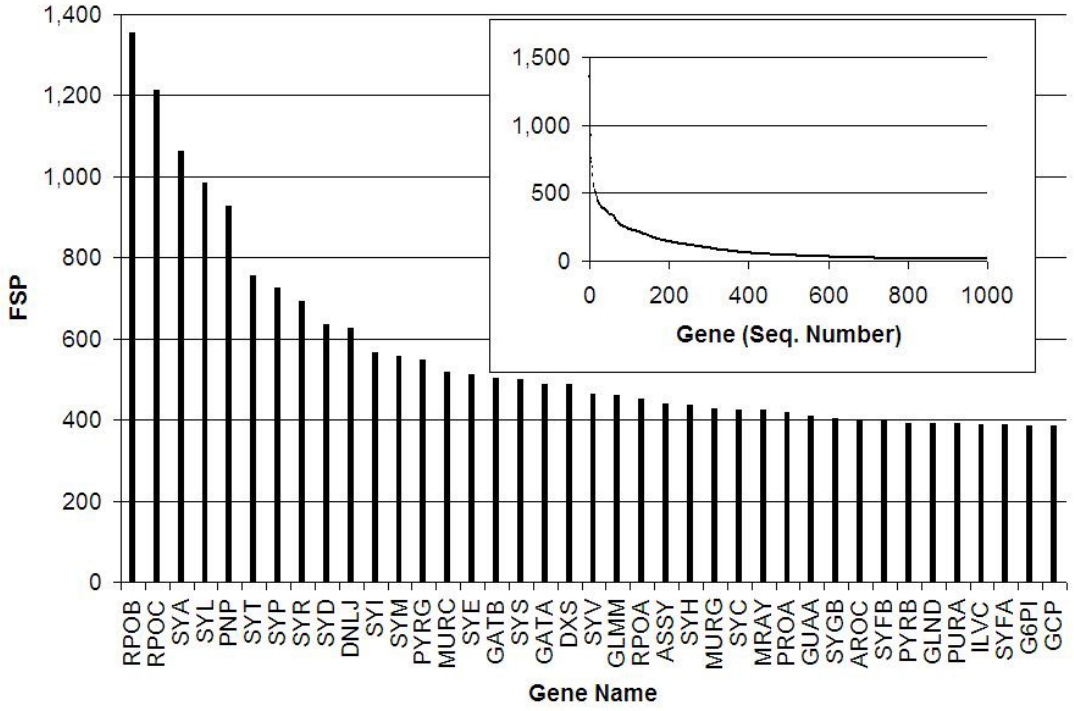


Figure 2.11: Histogram of Family SPs by gene name

FSPs are generated indirectly using ECs as classifiers in the Training Set and therefore the proportional representation of a gene family in Training Set 2 does not have to be exactly the same as in the set of FSPs.

Every FSP has a specific gene name in addition to its EC and thus the gene name assigned to the FSP can be viewed conceptually as a branch of the EC hierarchy, “EC level 5”.
As an example of this hierarchy we review the set of aminoacyl-tRNA synthetase family FSPs, which consist of 11,068 FSPs with EC=6.1.1.{x}. Figure 2.12 below shows the distribution of the FSPs by gene within each aminoacyl-tRNA synthetase group.

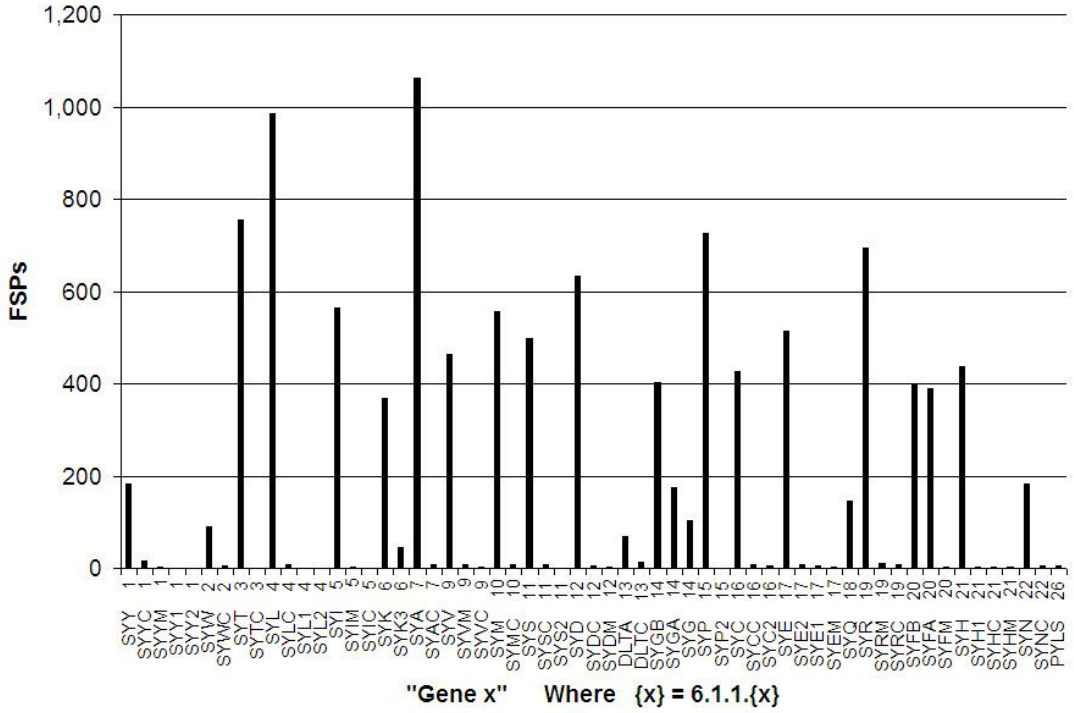


Figure 2.12: Histogram of aminoacyl-tRNA synthetase FSPs by tRNA Group and Gene.
For almost all of the groups (EC Level 4) there is a single dominant gene.

Using FSPs we can annotate the gene for a proteome, genome, metaproteome and metagenome, much the same way we generate EC predictions.
In the case of genomic and metagenomic predictions, FSPs have to be in the same Open Reading Frame to contribute to consistent coverage.
Because the FSP set is a much smaller subset of all SPs, recall is negatively impacted.
Table 2.13 below shows prediction and recall figures using SPs with length greater or equal to 7 for the complete genome of H. Pylori 26695.

SP Set and coverage threshold (L)	**N (P/P)**	**N (P/DP)**	**N (P/NP)**	**N (NP/P)**	**N (NP/NP)**	**Precision**	**Recall**
L 7 SPs 2.3 comparing ECs at level 3	244	11	207	60	1038	95.7%	77.5%
L 9 SPs 2.3 comparing ECs at level 3	225	7	40	83	1205	97.0%	71.4%
L 9 FSPs comparing FSP gene to Swissprot gene Annotation	209	25	90	107	1129	89.3%	61.3%

Table 2.13: Precision recall analysis for H Pylori 26695 using SPs v2.3 and the set of FSPs
In the third case, the precision and recall calculations are conducted not at the EC level but at the gene name level.

Gene predictions using FSPs provide more detailed biological insight information than EC predictions as demonstrated in the following two examples, analyzing genes HP0501 and HP0701 of H. Pylori 26695.
Analyzing gene HP0501 (Swissprot Access ID P55992) of H. Pylori 26695 we observe it is hit by five FSPs in the same Open Reading Frame providing a coverage length of 47 amino acids, which makes it a very solid 5.99.1.3 EC prediction.
But beyond that, we predict the gene function of this gene as DNA Gyrase subunit B.

FSP Hit	**EC**	**Family SP**
SIKVLKGL	5.99.1.3	GYRB
GRGIPVDIH	5.99.1.3	GYRB
SFVNNIKT	5.99.1.3	GYRB
QAILPLKGKILNV	5.99.1.3	GYRB
MGDEVEPRR	5.99.1.3	GYRB

Table 2.12: List of FSP hits on gene HP0501 of H. Pylori 26995

In general, Swissprot annotates DNA Gyrase subunit A enzymes, DNA Gyrase subunit B enzymes, Reverse Gyrase enzymes and others with the same EC, 5.99.1.3. In this case Swissprot annotates the name of the gene as DNA Gyrase subunit B, in agreement with our assignment
Similar results are obtained for HP0701 (Swissprot Access ID P48370 EC=5.99.1.3) for which we generate a very strong prediction of GYRA, with coverage of 60 amino acids. This prediction is in agreement with Swissprot annotation of this gene: DNA Gyrase subunit A. We aggregate Family SP names and build generic groups using the prefix of the Family name as identifier. The purpose of this construction is to provide a prediction at a family generic name in cases where two very similar FSPs hit a queried entity. This approach is similar to providing predictions at EC level 3 when SP hits coincide. An example of such generic groups is shown in tb_fam03 below. The first two Family SP names cannot be aggregated to any Generic SP Groups. Follow examples of Family names that we aggregate.

Family SP Name	Generic SP Group
1A1D	1A1D
2NPD	2NPD
5NTC	5NT*
5NTD	5NT*
AAPK1	AAPK*
AAPK2	AAPK*
AAT	AAT*
AATC	AAT*
AATM	AAT*
ABD12	ABD*
ABDH	ABD*

Table 2.13: Samples of SP Family names with the appropriate Generic SP Family Groups.

2.11 Taxon Specific Peptides – TSPs

Swissprot documents each of the enzymes in the Training set with its Taxonomic Lineage (TL). As in previous cases of Annotated SPs, Family SPs and GSPs, we expand the contents of SPs, assigning them attributes of enzymes they hit in the Training Set, this time with taxonomic lineage information.
We will adopt the following notation:

Taxonomic Lineage Level	**Notation**
Kingdom	Level 1
Phylum	Level 2
Class	Level 3

Table 2.14: Notation convention for taxonomic lineage levels

We apply the following taxonomic lineage specificity filtering rules to build the Production TSP set:

1. SPs in the Production SP set that are specific at Level three are promoted to the Production TSP set.
2. SPs in the Production SP set that are not specific at Level three but are specific at Level two are promoted to the Production TSP set.
3. SPs in the Production SP set that are not specific at Level three but are specific at Level one are promoted to the Production TSP set.
4. All other SPs are discarded

The Production TSP set consists of 134,306 TSPs distributed as follows:

Taxonomic Specificity Level	**Number of TSPs in the ` Production Set**
1	28,388
2	19,415
3	86,503

Table 2.15: Distribution of TSPs by Taxonomic Specificity Level

Sample of a few TSPs is shown in table 2.14 below.

TSP	Taxonomic Specificity Level Assigned to the TSP	Level 1	Level 2	Level 3	EC of TSP
DIDAIAVT	1	Bacteria	-	-	3.4.24.57
HLGLTPP	1	Bacteria	-	-	3.6.1
IRFEDCT	1	Eukaryota	-	-	3.6.1
LYKGKNG	1	Bacteria	-	-	1.4.4.2
VRADGVI	1	Bacteria	-	-	3.6.1
DWENPYVTL	2	Bacteria	Firmicutes	-	6.1.1.5
EFFQGFVNH	2	Bacteria	Proteobacteria	-	4.2.1.19
LAGMIKLI	2	Bacteria	Firmicutes	-	6.3.5
LYPEQRAEG	2	Bacteria	Proteobacteria	-	3.6.3.27
QRILEDD	2	Bacteria	Firmicutes	-	2.1.1.45
HLQDPLEVL	3	Bacteria	Proteobacteria	Alphaproteobacteria	2.4.2.21
LNGFYIP	3	Eukaryota	Metazoa	Chordata	1.14.14.1
LVTLLEQT	3	Bacteria	Proteobacteria	Gammaproteobacteria	2.3.1.181
RISLRPGPL	3	Eukaryota	Fungi	Dikarya	3.6.1
RKDFPTTGYTEVRYDE	3	Bacteria	Proteobacteria	Alphaproteobacteria	1.6.99.5
HLQDPLEVL	3	Bacteria	Proteobacteria	Alphaproteobacteria	2.4.2.21

Table 2.14: Sample of a few TSPs

Taxonomic lineage specificity at level 3 implies taxonomic lineage specificity at level two and one. Similarly, taxonomic lineage specificity at level two implies taxonomic lineage specificity at level one. Predictions using TSPs, the prediction are inclusive of lower levels of the Taxonomic Specificity Level Assigned to the TSP. Predictions at Taxonomic Specificity Level one are computed using contribution by TSP hits of TSPs with Taxonomic Specificity Level one, two and three.

Figure 2.13 shows the distribution of SPs and TSPs with Taxonomic Specificity Level one by the EC level 1 of the SP or TSP. The distributions are similar – rich EC classes that contribute to SPs contribute similarly to TSPs.

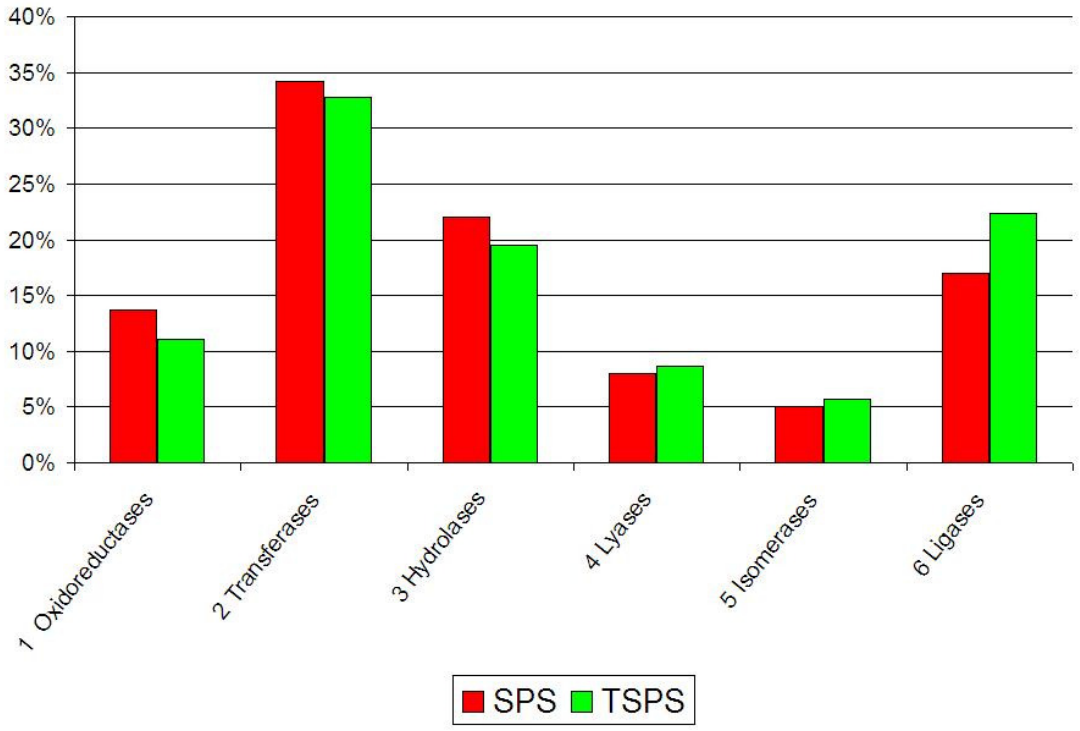


Figure 2.13: Histogram of SPs and TSP with Taxonomic Specificity Level one by the EC level 1.

Figure 2.14 shows the distribution of TSP with Taxonomic Specificity Level three by EC level 4. The inset shows the same distribution for all ECs. This distribution is important as it shows that the tRNA Aminoacyl synthetases have sufficient rich representation among TSPs.

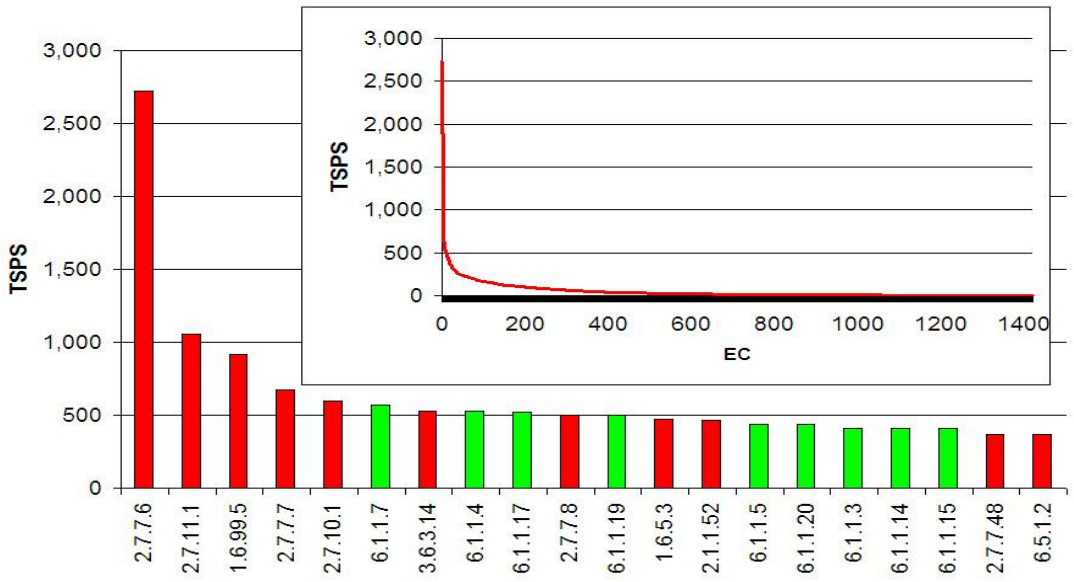


Figure 2.14: Histogram of TSP with Taxonomic Specificity Level three by EC level 4. The inset shows the same distribution for all ECs.

References

2.1 http://en.wikipedia.org/wiki/EC_number

2.2 Boeckmann B., Bairoch A., Apweiler R., Blatter M.-C., Estreicher A., Gasteiger E., Martin M.J., Michoud K., O'Donovan C., Phan I., Pilbout S. and Schneider M., The Swiss-Prot Protein Knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res. 31:365-370(2003).

2.3 Solan Z, Horn D, Ruppin E, Edelman S: Unsupervised learning of natural languages. Proc Natl Acad Sci USA 2005, 102:11629-11634.

2.4 Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M., and Hirakawa, M.; KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res. 38, D355-D360 (2010).

2.5 Chang A., Scheer M., Grote A., Schomburg I., Schomburg D. : BRENDA, AMENDA and FRENDA the enzyme information system: new content and tools in 2009. Nucleic Acids Res. 2009, Vol. 37, Database issue, D588-D592

2.6 Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu DY, Paulsen I, Nelson KE, Nelson W, Fouts DE, Levy S, Knap AH, Lomas MW, Nealson K, White O, Peterson J, Hoffman J, Parsons R, Baden-Tillson H, Pfannkoch C, Rogers YH, Smith HO: Environmental Genome Shotgun Sequencing of the Sargasso Sea.

2.7 Henikoff, S.; Henikoff, J.G. (1992). "Amino Acid Substitution Matrices from Protein Blocks". PNAS 89 (22): 10915–10919. doi:10.1073/pnas.89.22.10915. PMID 1438297. PMC 50453.

2.8 Henikoff, S.; Henikoff, J.G. (1992). "Amino Acid Substitution Matrices from Protein Blocks". PNAS 89 (22): 10915–10919. doi:10.1073/pnas.89.22.10915. PMID 1438297. PMC 50453.

2.9 http://www.ncbi.nlm.nih.gov/staff/tao/URLAPI/bl2seq.html

2.10 Meroz Y, Horn D: Biological Roles of Specific Peptides in Enzymes. Proteins: Structure, Function, and Bioinformatics 2008, 72(2):606-612.

2.11 Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium (2000) Nature Genet. 25: 25-29 PDF

2.12 Dinsdale EA, Edwards RA, Hall D, Angly F, Breitbart M, Brulc JM, Furlan M, Desnues C, Haynes M, Li L, McDaniel L, Moran MA, Nelson KE, Nilsson C, Olson R, Paul J, Brito BR, Ruan Y, Swan BK, Stevens R, Valentine DL, Thurber RV, Wegley L, White BA, Rohwer F: Functional metagenomic profiling of nine biomes. Nature 2008, 452:629-632

2.13 Christian Charras - Thierry Lecroq; Exact matching algorithms: www-igm.univ-mlv.fr/~lecroq/string/

2.14 (14) http://expasy.org/sprot/relnotes/relstat.html

3. Utilization of Specific Peptides for large volume enzymatic predictions

Data mining of Enzymes

Our article "Data Mining of Enzymes" published in BMC Bioinformatics 2009, 10:446doi:10.1186/1471-2105-10-446 is included almost in its entirety in this chapter.

Background

As mentioned in the introduction to this work, recently there has been a rapid growth in the number of putative proteins derivable from new genomic and metagenomic data [3.1]. The extended use of environmental shotgun sequencing to study diverse microbial systems has made metagenomics a vastly growing field leading to a flux of data, calling for development and application of new tools that allow its investigation [3.2]. Conventional tools for predicting the function of a protein from its sequence are based on sequence-similarity [3] or sequence-motifs [3.4, 3.5]. Here we outline a relatively simple and straight-forward method that is applicable to large numbers of sequences. Its purpose is finding whether each protein in the data is an enzyme and, if so, what its EC classification is. This Data Mining of Enzymes (DME) is based on the Specific Peptide (SP) method of [3.6], and is carried out by comparing the sequences of all proteins with a list of all SPs and looking for matches of the latter in the data.

SPs are strings of amino-acids, extracted from enzyme sequences using the motif extraction algorithm MEX [3.7]. They are selected for their specificity to levels of the Enzyme Commission (EC) 4-level functional hierarchy. We have updated the SP set of [3.6] by extracting it from all Swiss-Prot enzymes of July 1st, 2006. More details are provided in Methods.

Using SPs for prediction of enzymatic function needs some further decisions as to what to do if various SP hits on the same protein have EC assignments that are not consistent with one another. Moreover, one should decide when a single SP hit is sufficient to make a prediction. The methodology developed here relies on coverage length (overall number of amino-acids) of consistent SP hits. This is further described below, when testing performance on an enzyme test set, and when discussing a ten-organism test-set that contains non-enzymatic as well as enzymatic proteins. We develop a random model for the latter to assess the effect of accidental SP matches. The resulting methodology, which we call Data Mining of Enzymes (DME), is being applied to analyze several metagenomes.

Methods

3.1 The new SP sets

A novel method based on sequence motifs has been proposed by [3.6], who have studied enzymes in the Swiss-Prot database. They have demonstrated that enzyme functions, as represented by the four-level EC hierarchy, can be deduced from the appearance of deterministic short strings of amino-acids, denoted as Specific Peptides (SPs), on these enzymes. The SPs were derived from enzyme sequence data using an unsupervised motif extraction algorithm MEX [3.7], and filtered by the EC so that

each SP is specific to a particular EC branch, specifying the EC function that the enzyme performs. Thus, if an extracted motif is found to occur on enzymes belonging to only one EC number (i.e., 4th level of the EC hierarchy), this peptide will be declared to be an SP labeled with this EC number. If, however, the motif occurs on several EC numbers, all of which share the same 3rd-level hierarchy (i.e. the first three digits of their EC numbers are the same), the motif is declared as an SP with labeling at the third level of EC hierarchy, etc. The SPs of [3.6] comprise on average 8.4 amino-acids (SD 4.5), and were shown to compete favorably with a Smith-Waterman based SVM classifier. Usage of the SP methodology is demonstrated by our web-tool http://adios.tau.ac.il/DME. Given the sequence of an enzyme, this tool searches through the set of all SPs and finds which of them coincide with substrings of the sequence, indicating where they lie, what is the EC assignment associated with each SP, and provides the EC predicted by the DME method for the protein that is being queried.

Kunik et al [3.6] have investigated 50,698 enzyme sequences of the 48.3 Swiss-Prot release of October 2005. We have used the same methodology and applied it to all enzymes in the Swiss-Prot/Enzyme records of July 1st, 2006. The number of enzymes that have a single EC assignment is 89,854. Applying MEX and filtering it by EC levels in the same way as [3.6], we have obtained 87,017 SPs. This new 1st list of SPs serves as the basis for developing and analyzing our methodology.

In making the prediction of an EC number (i.e. 4th level of the EC hierarchy) based on one SP match, or several SP matches that have the same EC number assignment, we require that the total number of amino-acids of the protein matched with these SPs be at least seven. We refer to this number as the coverage-length L. If L at level 4, L4, is less than 7, we check for SP hits that are consistent at level 3 of the EC hierarchy, i.e. have identical first three digits in their assignments. Once again, a prediction is made if L at level 3, L3, is at least 7. In principle, the threshold of L at every EC level can be viewed as a parameter of our method. Reducing L increases recall at the expense of lowering precision, as will be discussed below.

Test data were downloaded from Swiss-Prot Release 56 on July 1st, 2008. We consider two types of test sets. The "Enzyme Test Set" consists of all enzymes integrated into Swiss-Prot between July 1st of 2006 and 2008. The "10 Organisms Test Set" consists of proteins of E. coli and 9 other bacteria (see Table 3.0 below) containing enzymes from the same period of 2006 to 2008, and all other proteins incorporated into Swiss-Prot by July 1st, 2008.

Organism	**# Proteins in test Set**
Anabaena variabilis (strain ATCC 29413 / PCC 7937).	240
Bacillus cereus (strain ATCC 14579 / DSM 31).	269
Bradyrhizobium sp. (strain BTAi1 / ATCC BAA-1182).	142
Burkholderia sp. (strain 383).	247
Cytophaga hutchinsonii (strain ATCC 33406 / NCIMB 9469).	128
Escherichia coli (strain K12).	2,932
Rhodococcus sp. (strain RHA1).	183
Solibacter usitatus (strain Ellin6076).	94
Sorangium cellulosum (strain So ce56).	51
Streptococcus pneumoniae.	223
Total	4,509

Table 3.0: List of the ten organisms used as a test-set.

A compilation of the training and test datasets together with precision and recall values is displayed in Table 3.1 below.

Dataset	**Selection Criteria from Swiss-Prot**	**Number of Proteins (and SPs)**	**Precision**	**Recall**
Training set #1	Single EC annotation and Date-Integrated before 7/1/2006	89,854 (#SPs=87,017)	100%	85%
"Enzyme Test Set"	EC annotation and Date-Integrated between 7/1/2006 and 7/1/2008	24,443	98%	70%
"Ten Organism Test-Set"	EC annotation and Date-Integrated between 7/1/2006 and 7/1/2008 and all non-enzymes before 7/1/2008	4,509	98%	76%
Training set #2	Single EC annotation and Date-Integrated before 7/27/2009	201,169 (#SPs=312,465)	100%	94%
Test Set #3	1,000 random annotated Swiss-Prot enzymes with integrated date after July 27th 2009 and a single EC annotation at EC level 3 or EC level4.	1,000	98.8%	76.8%

Table 3.1: Compilation of training and test datasets.

It includes also information about precision and recall (for definitions see below) that will be further discussed in the first Results section. These values are obtained by determining 3rd level EC assignments, using coverage-length of L3≥7. Precision values of 100% on the training sets are of course trivial results of specificity.

54% of the proteins in the 1st training set carry Swiss-Prot annotations of 'active site', 'binding site' or 'metal binding site' at specific locations of single amino-acids. SPs cover these functionally important sites significantly more than other loci on proteins, thus indicating biological significance of SPs (for an extensive discussion see [3.8], in particular Table 1 there). SP matches that overlap such sites are compiled, and the corresponding SPs are denoted as Annotated SPs (ASPs).

We have thus compiled a list of 6,078 ASPs. All appear at least four times in the training set, and the location of the annotation is consistent in the different appearances. Most ASPs carry single annotations (1,900 active sites, 1932 binding sites and 1,819 metal binding sites), 418 ASPs carry two annotations and 3 ASPs carry all three annotations.

A second set of SPs is extracted from Swiss-Prot data dated July 27th, 2009. This training set, consisting of all singly annotated enzymes, contains 201,169 proteins. It has led to 312,465 SPs. Their length distribution is presented in Additional File Fig. 3.0 below.

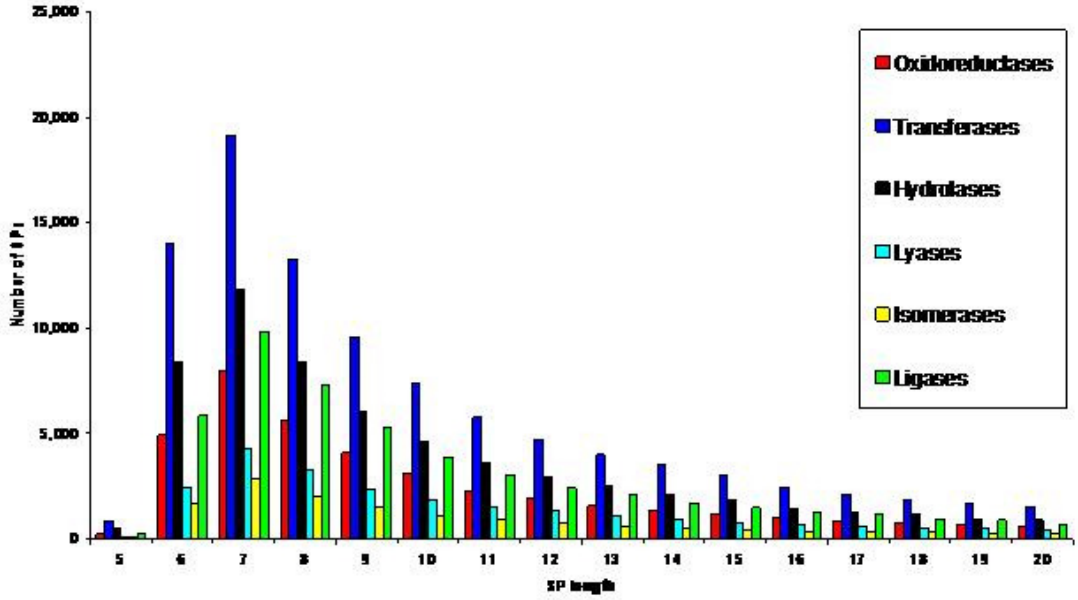


Figure 3.0: Length histogram of the 2nd SP set.

This set includes 285,485 SPs with labels corresponding to EC levels 3 and 4 (containing 257,598 SPs of length ≥ 7). Only SPs with EC labels at levels 3 and 4 are relevant for the assignment of EC level-3 annotations to proteins, and hence for the calculation of recall included in Table 3.1. It should be emphasized that only 191,275 of the Swiss-Prot annotated enzymes in the training set carry EC annotations at levels 3 and 4. They are the ones on which the EC predictions at level 3 are tested, leading to the recall result of 94%. The 2nd SP set is being used for the analysis of metagenomic data and is incorporated in our web-tool at http://adios.tau.ac.il/DME.

.

3.2 Estimate of accidental SP matches

Proteins that do not possess enzymatic functions may still have a substring that matches an SP. Such SP matches will be called 'accidentals'. Their occurrences can be modeled by SP hits on random protein sequences. Such random sequences are generated from real data by scrambling the order of the amino-acids in every protein, conserving only first-order statistics. 3 such sets were produced in order to measure

the expected random hits. Estimates of the probabilities of accidental occurrences of SPs are derived below for the 10 organism test-set and for Sargasso Sea data.

3.3 Recall-precision analysis of EC annotations in enzymes.

Comparing the results of our method with an expert-method (such as Swiss-Prot) we face three possible situations when dealing with a collection of enzymes: P|P where the model prediction coincides with that of the expert, P|DP where the expert provides a different EC assignment, and NP|P where the model provides no prediction for enzymes whose EC assignments are given by the expert. As mentioned in chapter 2, we define the following measures in terms of number of occurrences:

$$PRECISION = \frac{N[P \mid P]}{N[P \mid P] + N[P \mid DP]}$$

$$RECALL = \frac{N[P \mid P]}{N[P \mid P] + N[P \mid DP] + N[NP \mid P]}$$

This is a generalization of the common terms used in binary classification problems where P|P, P|DP and NP|P are replaced by true-positive, false-positive and false-negative correspondingly.

3.4 Recall-precision analysis of EC annotations in proteins.

Extending the previous analysis to a collection of proteins we have to add two more possibilities: P|NP, where the new method has an EC prediction whereas the expert does not have one, and NP|NP where both do not have any EC assignment. Whereas the latter corresponds to true-negative in a binary classifier, the former, P|NP could be added to P|DP as 'false-positive'. Since, however, there are many cases where the absence of an EC assignment does not imply that the protein in question is not an enzyme, we opt to define a new measure, putative novelty ratio, as the fraction of such P|NP out of all the predictions of the model:

$$PUTATIVE\_NOVELTY = \frac{N[P \mid NP]}{N[P \mid P] + N[P \mid DP] + N[P \mid NP]}$$

Other measures one can define are

$$SPECIFICITY = \frac{N[NP \mid NP]}{N[NP \mid NP] + N[P \mid DP] + N[P \mid NP]}$$

$$ACCURACY = \frac{N[P \mid P] + N[NP \mid NP]}{N[P \mid P] + N[P \mid DP] + N[P \mid NP] + N[NP \mid P] + N[NP \mid NP]}$$

They are the analogs of

$$SPECIFICITY = \frac{TRUE\_NEGATIVE}{TRUE\_NEGATIVE + FALSE\_POSITIVE}$$

and

$$ACCURACY = \frac{TRUE\_POSITIVE + TRUE\_NEGATIVE}{TRUE\_POSITIVE + TRUE\_NEGATIVE + FALSE\_POSITIVE + FALSE\_NEGATIVE}$$

in conventional binary classifications.

Results: Analysis of the Methodology

3.5 Analysis of the Enzyme Test Set using the 1st SP set.

In making the prediction of an EC number (i.e. 4th level of the EC hierarchy) based on one SP match, or several SP matches that have the same EC number assignments, we require that the total number of amino-acids of the protein matched with these SPs be at least seven. We refer to this number as the coverage-length L. In principle, the threshold of L at every EC level can be viewed as a parameter of our method. Reducing L increases recall at the expense of lowering precision. This is exemplified in Table 3.2 below, where we analyze our enzyme test set and show precision and recall at 3rd EC level as function of the L3 threshold.

L3 threshold	Precision	Recall
5	95.1%	72.4%
6	95.8%	72.3%
7	98.4%	70.0%
8	99.4%	67.1%
9	99.5%	66.2%
10	99.5%	65.4%
11	99.5%	65.0%
12	99.6%	64.8%
13	99.6%	63.9%

Table 3.2: Variation of precision and recall of DME (based on the 1st SP set) on the enzyme test-set as function of the L3 threshold.

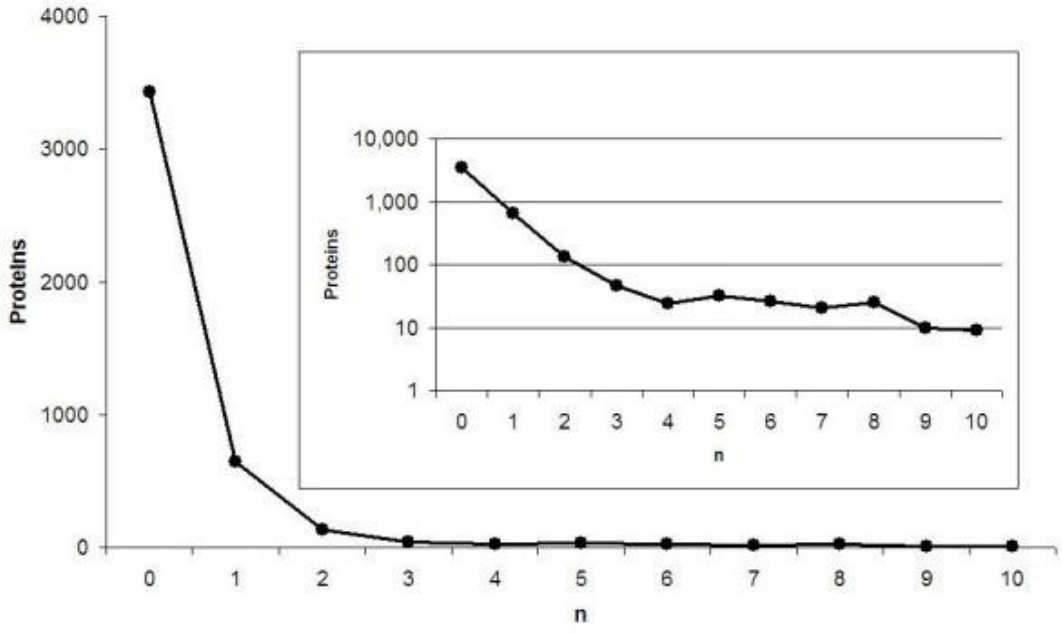
Although precision turns out to be quite high, even for low L3 values, recall is low when compared to what BLAST [3.9] can achieve on this test-set. Using the most significant outcome of a BLAST search against the 1st training set as its prediction, and limiting the most significant e-value to stay below e-05, we find BLAST precision of 98% and recall of 95%, to be compared with DME values of 98.4% and 70% when setting L3≥7. Thus while precision is similar, DME loses on recall. There is no direct relation between DME and BLAST, although high coverage-length L values of DME go usually hand in hand with very low e-values of BLAST. Differences may occur for low L values of DME, and relatively high e-values in

BLAST. We refer to Kunik et al. [3.6] for a discussion of such examples (see Table 4 there). The advantages of SPs in resolving classification problems in situations of remote homology have been discussed and exemplified by [3.8].

It is worthwhile pointing out that the fact that one can abide by such a small threshold value of $L \geq 7$ is strongly connected to our requirement that the SP matches on the protein's sequence be exact. If one were to allow for insertions or deletions or replacements, such as the BLOSUM62 matrix [3.10], this would not work. Based on various trials we may state that, whereas reliance on BLOSUM works well for BLAST searches over large sequences, it ruins predictivity and specificity of SP searches even if only single amino-acid changes are allowed.

3.6 Analysis of the ten organism test-set

The ten organism test-set contains 4,509 proteins of E. coli and 9 other bacteria listed in Additional File Table A2.1. Proteins for this dataset were downloaded from Swiss-Prot on July 1st 2008 and include all proteins that had no EC annotation in Swiss-Prot prior to July 1st 2006. The intersection between the 10 organism test-set and 1st training set used to build the SPs is void and allows us to develop and test the SP methodology on general proteomic data rather than on enzymes only. SP search on this dataset, using our 1st set of 87,017 SPs (see Methods), leads to the results shown in Fig. 3.1 below, sorted according to the number n of SP matches.



SP hits on the ten-organism test-set. The numbers of proteins in the ten organism test-set carrying n SP matches, for n = 0 to 10. The inset shows the same data on a semi-log scale, emphasizing the sharp exponential decrease for low n, partially reflecting the existence of erroneous SP hits.

Figure 3.1 - SP hits on the ten-organism test-set.

1,079 proteins have at least one SP match (or 'hit'). Some of them may be due to random hits and our task is to resolve which of the hit proteins should be recognized as enzymes and what their EC assignments should be. As before, we propose to rely on coverage length. We judge the prediction not by how many SP hits (with consistent annotations) are observed, but by L3, the number of amino-acids matched by all SP hits whose EC assignment is identical within the first three digits of the EC number. In order to have some intuition about the expected noise level, we compare in Table 3.3 below SP hits on real data with random model results for different values of L3. Entries of L3=0 refer to either no SP hits, or hits by SPs that have labels with EC levels 2 and 1 but none at EC levels 3 and 4. The columns random and stdev refer to the average and standard deviation of seven random sets. Noise is the ratio of random/real. All 4509 proteins of the ten-organism test-set were included in this search.

L3	**Real**	**Random**	**stdev**	**Noise**
0	3768	4079.86	23.87	
4	0	0.57	0.53	
5	41	53.57	9.81	1.31
6	305	307.86	14.12	1.01
7	106	59.43	8.30	0.56
8	13	4.57	1.40	0.35
9	5	0.29	0.49	0.06
10	2	0.14	0.38	0.07
11	1	0.00	0	0
12-15	25	0.71	0.76	0.03
>15	243	0	0	0

Table 3.3 Comparison of results for the ten organism test-set with those of a random model as function of coverage-length at level 3 of the EC hierarchy.

We will use L3≥7 as our threshold criterion, as in the enzyme data-set discussed in the previous section. We note that predictions based on L3=7 may still have a large uncertainty, however from L3=8 onwards random hits become very small. Our threshold criterion leads to the results displayed in Table 3.4 below, with precision=98.4%, recall=75.9%, accuracy=95.1% and putative novelty=35.2%.

	DME	**Swiss-Prot**	**# proteins**
A	P	P	252
B	P	DP	4
C	P	NP	139
D	NP	P	76
E	NP	NP	4,038

Table 3.4 DME predictions vs. Swiss-Prot EC (level 3) annotations for the 10 organism Test Set.

The interest in this exercise is twofold: to see how well our method performs on unassigned proteins, i.e. true-negatives, and how good our predictions are for putative novelties. Indeed, our accuracy turns out to be high, 95.1%, which proves that we have correct negative assignments.
Seven out of the 139 putative novelties (category C in Table 3.4) have been annotated by Swiss-Prot since July 2008, six out of which are at levels 3 or 4. All observations are consistent with the predictions, as shown in Table 3.5 below. Quoted here are also all coverage lengths on which the predictions were based. Note that also the one based on coverage length 7 has been validated. All this may be viewed as an indication (although not a proof) of the validity of DME predictions. The first six entries in Table 3.5 belong to E. coli, and the last protein belongs to Bacillus cereus.

ID	**DME Prediction (1st SP set)**	**L1**	**L2**	**L3**	**L4**	**Current Swiss-Prot EC annotation**
P06610	1.11.1	25	22	22	0	1
P07821	3.6.3	25	25	25	0	3.6.3.34
P0A9V1	3.6.3	7	7	7	0	3.6.3
P33360	3.6.3	13	13	13	0	3.6.3
P76469	4.1.2	9	9	9	0	4.1.2.n3
P77257	3.6.3.17	14	8	8	8	3.6.3
Q81IT9	3.6.1	58	58	58	0	3.6.1

Table 3.5: DME predictions for the ten-organism test-set are compared with recent Swiss-Prot EC assignments. L1 to L4 are the coverage-lengths at EC levels 1 to 4 respectively.

3.7 Classification based on Annotated SPs.

It has been noted by [3.6] and [3.8] that some of the SPs can be demonstrated to play important biological roles since they carry crucial amino-acids known to serve as active sites, binding sites or metal binding sites. Such annotations are available for 54% of the enzymes in the 1st Swiss-Prot training set. Selecting only SPs that carry these annotations we obtain a set of 6,078 Annotated SPs (ASPs), a mere 7% of all SPs. We have tested it on the enzyme test set. Using annotation predictions at the third level of EC we find precision 99.6% and recall 25.4%. The limited recall is due to the fact that ASPs have been derived from only 54% of the training set. Nonetheless they possess the advantage of being selected due to their demonstrated operational importance to the catalytic function. Because of their limited recall we have not used the ASPs as the primary tool for large scale analysis; however we list their properties in our web tool http://adios.tau.ac.il/DME. Any queried protein can be analyzed by this tool for SP hits and the expected DME prediction. The appearance of ASPs may serve as providing additional credence to the prediction, as well as specifying the positions of expected active or binding sites.

3.8 Analysis of Sargasso-Sea data

After verifying DME on the two test-sets we turn to an analysis of the 1,001,986 records in the Sargasso Sea protein data [3.10]. The average length of these proteins is 194 amino-acids, with SD=109. For this analysis we employ our 2nd set of SPs, updated on July 2009. In order to reduce random hits, we have further limited our SP set to include only peptides of length 7 amino-acids or more. Using a random set of 5,000 proteins selected from these data, we generated three randomized protein sets from which we calculated the probabilities of accidental matches. The results are displayed as function L3 in Table 3.6 below. The columns random and stdev refer to the average and standard deviation of three random sets. Noise is the ratio of random/real.

L3	Real	Random	stdev	Noise
0	3,910	4,868	5.1	
7	235	127	5.5	0.54
8	71	6	2.1	0.08
9	40	0		0
10	27	0		0
>10	717	0		0

Table 3.6 Numbers of sequences with consistent SP hits (same category at level 3 of the EC hierarchy) are compared between 5,000 proteins randomly chosen from Sargasso-Sea data, and a corresponding random model, as function of coverage-length.

Similar results are obtained for L4. The results of Table 3.6 are slightly better than Table 3.3. The reason is that we have limited ourselves here to SPs of individual length 7 or more. Once again we choose L=7 as our threshold for DME predictions. Applying DME with this threshold we obtain EC assignments at levels 3 and 4 for 220,278 proteins. All assignments are provided in Appendix 2 – Tables A2-1-A2.3.

In Fig. 3.2 we display a histogram of the 30 largest EC sub-subclasses (level 3) that emerge from our DME analysis. The category with the largest number of different proteins is 6.1.1, corresponding to aminoacyl-tRNA synthetases (aaRS). Since there are about 20 aaRS enzymes expected for each organism, this allows us to estimate the content of the metagenome to be of order of 800 species or so. Looking at level 4 annotations, i.e. at specific aaRS enzymes, we find that their numbers vary from 116 to 1326. These differences may be due both to different occurrences of aaRS sequences in the sample, and to different efficiencies of the SP methodology for different aaRSs. The order of magnitude of 1000 different species remains a reasonable estimate. The same order of magnitude can be derived from another source. Venter et al. [3.11] have provided some information about single copy proteins (Table 2 there) in trying to arrive at estimates of the number of species involved. One such protein is the gyrase subunit B enzyme, GyrB. The same enzyme has also been proposed by Watanabe et al. [3.12] for the purpose of spanning a database for identification and classification of bacteria. GyrB is one of several protein families belonging to EC 5.99.1.3 (DNA gyrase). Checking through the SPs

belonging to this EC we have found a subset that is specific to GyrB only. Using this subset we estimate the number of GyrB copies in the Sargasso-Sea data to be 1344, which is close to the number of maximal fragment depth of 924 quoted in Table 2 of [11], and is in the same ball-park as the aaRS estimate.
In addition to 6.1.1 (aaRS) enzymes we observe the following leading categories: 3.6.3 (Hydrolases catalyzing transmembrane movement of substances involving ATPases), 2.7.7 (Nucleotidyl transferases), 1.1.1 (Oxidoreductases acting on the CH-OH group of donors), and 4.2.1 (Carbon-oxygen lyases).

There are several EC numbers (i.e. level 4 of the hierarchy) that are particularly abundant. They are presented in Table 3.7 below, where we list all cases that appear more than 2000 times in the data. Some of them have already been mentioned above: the DNA gyrase, and its role in estimating the number of species, and the two ECs belonging to the subsubclass of 2.7.7 (Nucleotidyl transferases), playing important roles in RNA and DNA polymerases.

EC	# proteins	Enzymatic activity
2.7.7.6	5,993	DNA-directed RNA polymerase
1.6.99.5	2,999	NADH dehydrogenase (quinone)
5.99.1.3	2,610	DNA topoisomerase (ATP-hydrolysing). DNA gyrase.
6.3.5.5	2,198	carbamoyl-phosphate synthase (glutamine-hydrolysing)
3.6.3.14	2,169	H^+-transporting two-sector ATPase. ATP synthase.
2.7.7.7	2,083	DNA-directed DNA polymerase

Table 3.7 Leading occurrences of EC-numbers in Sargasso-Sea data

All our predictions for the enzymatic annotations of the Sargasso-Sea data are presented in Additional File Tables A2.1-A2.3 in Appendix 2. We wish to point out that some of the enzymes contain two or more EC assignments. Table 3.8 below reports some of these occurrences. Included here are the most abundant observations of dual EC assignments, sorted by the numbers of proteins exhibiting the two annotations.

Prediction a	Prediction b	# Proteins
3.5.4.25	4.1.99.12	27
3.6.3.44	2.7.1.130	6
1.1.1.205	1.7.1.7	6
2.7.1.25	2.7.7.4	6

Table 3.8: Some examples of doubly annotated enzymes uncovered by DME in the Sargasso-Sea data.

The first and the last entries in Table 3.8 have many analogs in currently known doubly-annotated enzymes in Swiss-Prot. Checking all proteins we find that the SP hits that belong to the two different EC numbers do not overlap on the protein sequences, thus falling comfortably into the categorization of two different catalytic domains. It is interesting to note that finding multiple domains is easier with SPs than it is with BLAST: we will not miss out on a small domain of a protein that may be

overshadowed by sequence similarities with a larger protein domain, and we can immediately check whether the different catalytic regions lie on disjoint sections of the protein. A full list of the doubly annotated Sargasso-Sea enzymes is presented in Appendix 2, Table A2.2. A further list of triple-enzymatic annotations is presented in Appendix 2 Table A2.3.

3.9 Human Gut Metagenome

Gill et al. [3.13] have analyzed the DNA sequences obtained from fecal DNA of two healthy adults – 'subject 7' a female aged 28 and 'subject 8' a male aged 37. We have analyzed the resulting proteins (downloaded from http://img.jgi.doe.gov/m/) with our DME method. The two proteomes of subjects 7 and 8 consist of 20,523 and 25,980 proteins correspondingly. We predict enzymatic annotations for 3,428 proteins of subject 7 and 4,102 proteins of subject 8. These numbers are relatively lower than the enzymatic content of Sargasso-Sea. Numbers of 6.1.1 enzymes are predicted to be 260 and 264 for subjects 7 and 8 respectively. Thus the number of different species contained in these samples is scaled down by two-orders of magnitude compared to the Sargasso-Sea data, which is quite reasonable given the size of the databases. Further comparisons between the three metagenomes are offered in the next section.

3.10 Enzymatic Profile

Trying to compare different metagenomes with each other one has obviously to resort to some normalization method. Normalizing the results of a bar chart like Fig. 3.2 below by the total number of enzymes that we find, we obtain a spectrum characteristic of the genome or metagenome we study, which we will refer to as its enzymatic profile.

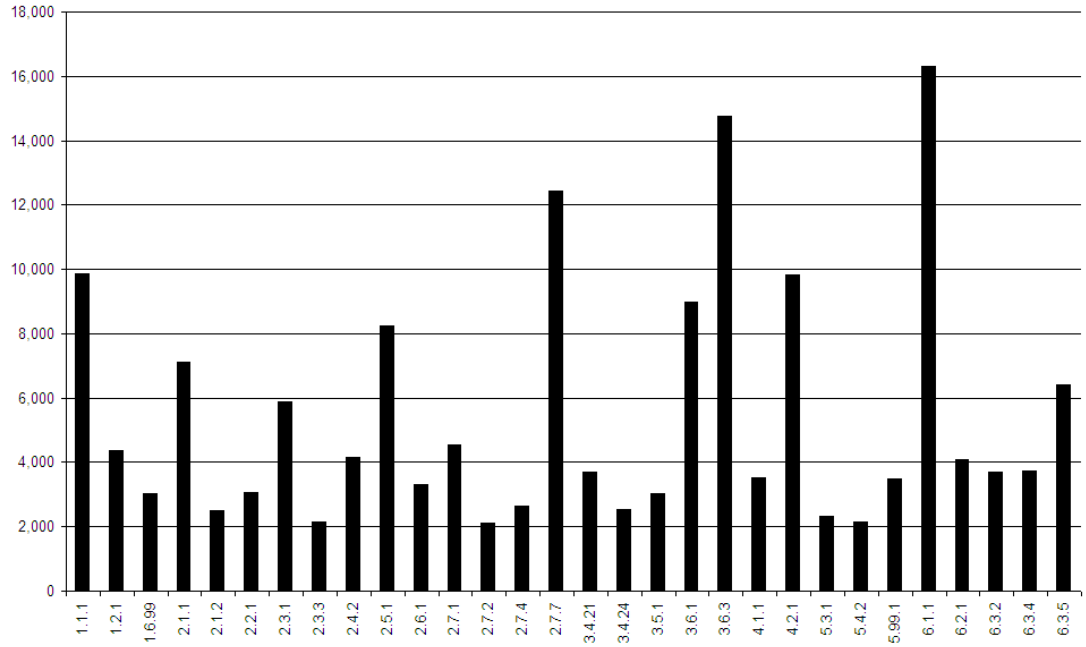


Figure 3.2: Numbers of enzymes predicted by DME in the Sargasso-Sea data. Shown are the thirty leading level 3 EC categories.

Figure 3.3 below depicts such profiles for the examples studied in this paper, the Sargasso Sea one, and the two gut metagenomes, all based on DME predictions. Since all three are bacterial metagenomes the leading EC categories are quite similar. The identities of the leading categories have already been described in the previous section.

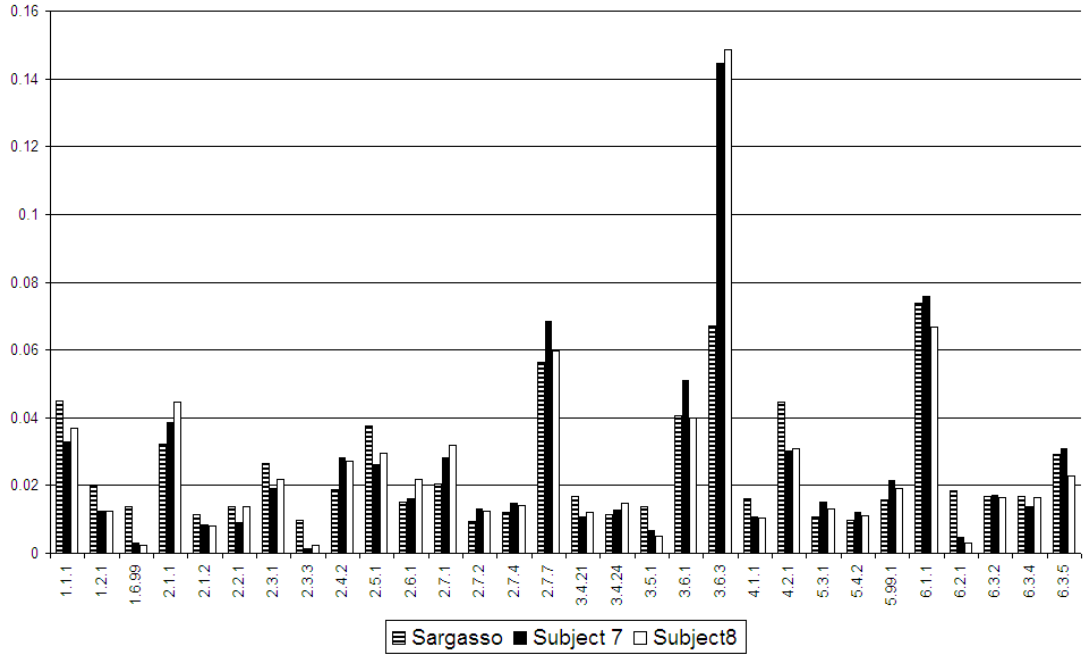


Figure 3.3 - Enzymatic profiles of three metagenomes. Compared are the relative numbers of identified enzymes in the 30 leading sub-subclasses (EC level 3) of the Sargasso-Sea meatagenome with those of the gut microbiomes.

In spite of the obvious similarities, there exist differences among the three histograms. We use the absolute value of the difference of any two distributions as the difference measure (theoretically limited to vary between 0 and 2). Taking into account all level-3 EC predictions we obtain the distances between the different distributions presented in Table 3.9 below.

Metagenome	Sargasso	Subject7	Subject8
Sargasso	0	0.42	0.41
Subject7	0.42	0	0.18
Subject8	0.41	0.18	0

Table 3.9 Absolute values of differences between enzymatic profiles based on the DME predicted distributions at level 3 of EC.

As expected, the two gut metagenomes are the closest pair.
It has been emphasized by [3.14] and by [3.15] that the functional characteristics of a metagenome vary with the environment in which it is being found. Hence we expect the genetic enzymatic profiles to vary accordingly.

Our exercise shows that the gross features of microbial communities may be similar, thus more attention will have to be paid to smaller details, in particular emphasizing the cases where the relative differences between EC categories are the largest. This may become a useful tool in the future.

We wish to close this section by emphasizing that the three metagenomic profiles are different from those derived from the genome of E. coli, and very different from human. The comparisons are presented in Fig. 3.4 below, drawn according to the top 20 categories of E. coli, and in Figure 3.5 below, displaying the top 20 categories of human.

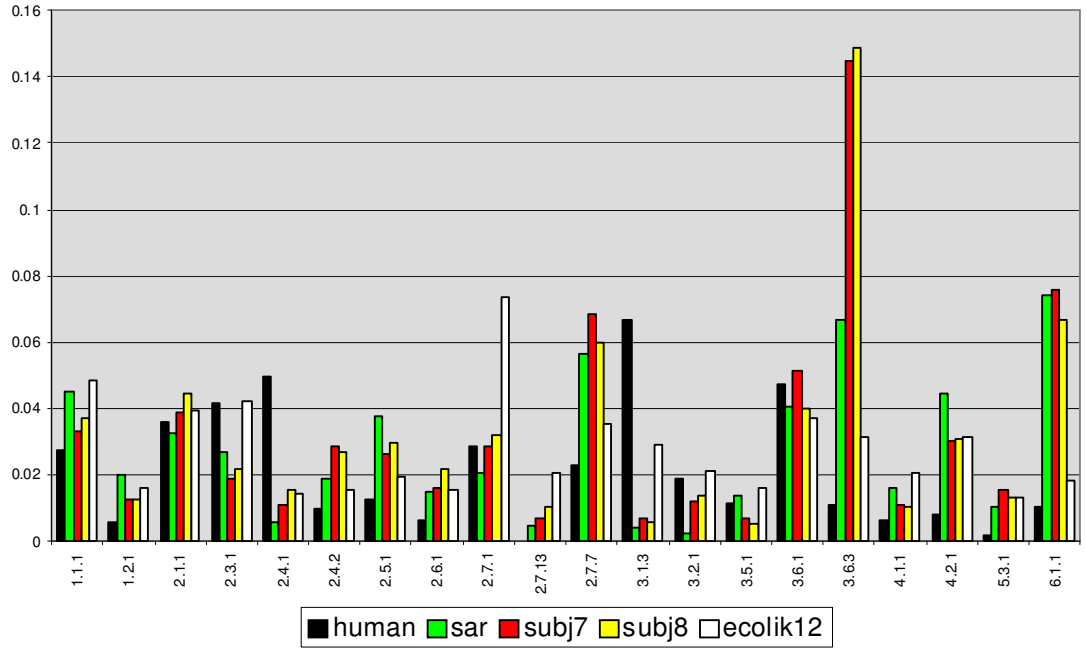


Figure 3.4: Comparison of enzymatic profiles based on the 20 leading categories of E. Coli.

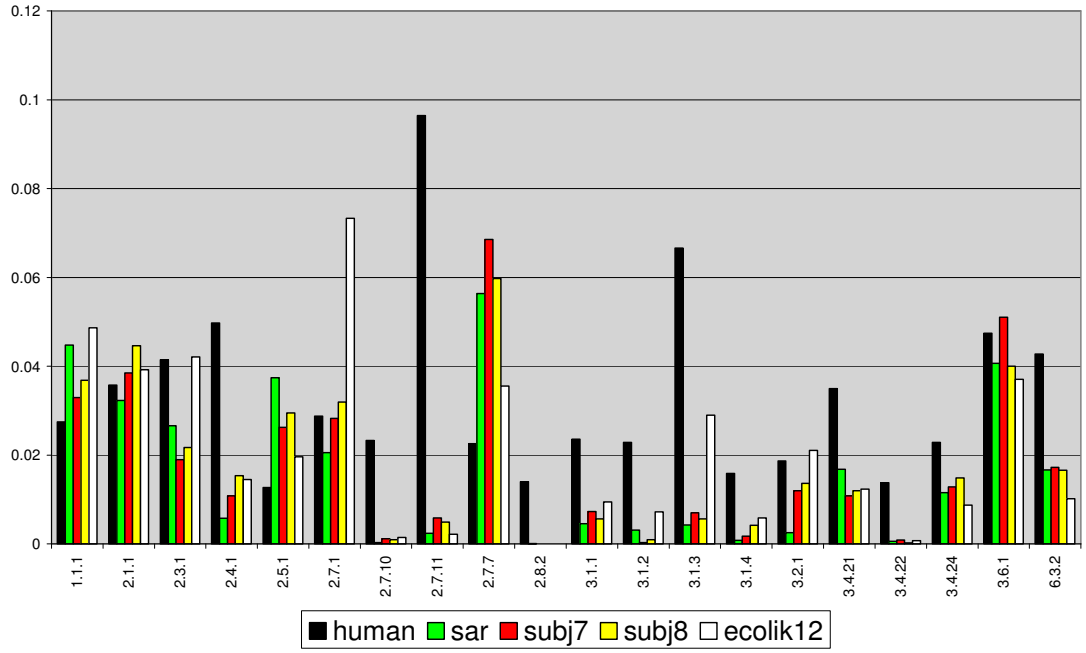


Figure 3.5: Comparison of enzymatic profiles based on the 20 leading categories of human

It is quite evident that the weights (or numbers of different genes) of different EC categories change considerably from human to E. coli to bacterial metagenomes. This implies that enzymatic profiles contain information that may be of value in future studies of novel genetic material.

Discussion

Using SPs it seems quite straightforward to perform data-mining of enzymes. There are however several provisos: a) although a majority of enzymes carry SPs, there exists a minority that does not; hence not all enzymes are expected to be discovered in a new dataset. b) SPs were substantiated on a training set, and their generalization carries with it some error, even on a test set composed of enzymes only. Errors may be due to a) changes in the official EC classification of an enzyme, or b) real biological changes such as evolutionary loss of an active site in a protein that resembles a known enzyme but has no catalytic function, or c) random appearance of SPs on proteins that have no catalytic activity. Errors due to reclassification of EC numbers cannot be controlled in any a-priori manner. The question of function loss can be partially checked through searching for the absence of annotated SPs in cases where such annotations may be expected for the enzyme in question. This demonstrates the importance of detailed corroboration of each individual prediction of the large-scale method studied here. The third source of errors, due to random appearance of SPs on proteins other than enzymes, has been taken into account by

limiting our predictions to consistent SP hits with minimal coverage length of 7, and specifying the L values of our predictions as a measure of their confidence.

DME is based on deterministic motifs only, i.e. strings with specific sequences of amino-acids. Comparing it with the well-known motif method of Prosite patterns (Bairoch et al, 1997), by using available information in Swiss-Prot, we find that the latter has precision of 97% and recall of only 47% on the Enzyme test set, thus falling short of DME predictions. When comparing DME to BLAST on the enzyme test-set we found that DME had comparable precision (98.4% vs 98%) while BLAST has much better recall (95% vs 70.0%). Note that this comparison was based on the 1st SP set of July 2006.

It should be appreciated that the comparative procedure based on the Enzyme test set has some bias in favor of BLAST, because the latter serves as one of the inputs to Swiss-Prot assignments. As a result, cases of remote homology which may be captured by DME could have been missed by BLAST-based assignments, as was demonstrated by [3.6] and by [3.8]. The SP-based search has two other advantages over BLAST: it is conceptually simpler, relying only on a look-up table, and it points to specific locations on the queried protein which may be relevant to the expected catalytic function of that enzyme. Hence it may have wide practical implications for enzyme research and development.

In spite of all the precautions outlined in the first paragraph, our predictions concerning the 10 organism test-set reported in this paper, do extremely well. Moreover, note that the recall quality of SPs on their training sets increased dramatically from 85% in 2006 to 94% in 2009 (see Table 3.1). This means that the minority of enzymes without SP hits diminishes with time. The reason is quite clear: MEX thrives on redundancy of patterns in the data. Therefore, the more proteins of the same family there are in the database, the better MEX will perform. As these lists fill up in the Swiss-Prot database, they can be better represented by simple SP motifs. Higher recall on the training set will undoubtedly reflect itself also as higher recall on future test sets, thus suggesting that the gap between the recall of BLAST vs DME will shrink with time. Indeed, carrying out a DME analysis, based on the 2nd SP set, of 19,849 enzymes that have been added to Swiss-Prot from July 28 to Sep 29, 2009, we find on this novel test set precision of 99.2% and recall of 92.4%. This is a considerable increase over the recall of 70% of the 1st SP set measured on the enzyme test set (see Table 3.1).

A straightforward peptide characterization of protein families seemed hopeless a decade or two ago, and hence necessitated the development of more sophisticated approaches such as BLAST, to quantify sequence similarities. Our analysis demonstrates that this has changed with time (and increasing amounts of data) so that nowadays the SP approach may be regarded as a useful tool, leading to valuable information. Such information, for three metagenomic data-sets, has been presented here as an example of the power of our novel methodology.

Conclusions

The requirement that SP occurrences on protein sequences has some minimal coverage length, e.g. $L \geq 7$ amino-acids in our analyses, leads to the novel tool of DME. It is applicable to large genomic and metagenomic data, and provides a good indicator for the enzymatic classification of the queried proteins, based on a look-up table only. A web tool identifying SP (and ASP) occurrences on any queried protein sequence, and providing the EC prediction of DME, is available online at http://adios.tau.ac.il/DME.

References

3.1. Angly, F.E., Felts B., Breitbart M., Salamon P., Edwards R.A., Carlson C, Chan A.M., Haynes M., Kelley S., Liu H., Mahaffy J.M., Mueller J.E., Nulton J., Olson R., Parsons R., Rayhawk S., Suttle C.A., Rohwer F. The Marine Viromes of Four Oceanic Regions. PLoS Biol 2006: 4(11), e368

3.2. Eisen, J.A. Environmental Shotgun Sequencing: Its Potential and Challenges for Studying the Hidden World of Microbes. PLoS Biol 2007: 5(3), e82

3.3. Tian, W. and Skolnick, J. How well is enzyme function conserved as a function of pairwise sequence identity? J. Mol. Biol. 2003: 333, 863-882.

3.4. Bork, P. and Koonin, E.V. Protein sequence motifs. Curr. Op. Structural Biology 1996: 6, 366-376.

3.5. Bairoch A, Bucher P, Hofmann K. Prosite. Nuc. Acids Res. 1997: 25, 217-221.

3.6. Kunik, V., Meroz, Y., Solan, Z., Sandbank, B., Weingart U., Ruppin E., Horn D. Functional representation of enzymes by specific peptides. PLOS Comp. Biol. 2007: 3(8), e167

3.7. Solan Z., Horn D., Ruppin E, Edelman S. Unsupervised learning of natural languages. Proc. Natl. Acad. Sci. USA 2005: 102, 11629-11634.

3.8. Meroz, Y. and Horn, D. Biological Roles of Specific Peptides in Enzymes. Proteins: Structure, Function, and Bioinformatics 2008: 72 (2), 606-612.

3.9. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J. Basic local alignment search tool. J. Mol. Biol. 1990: 215: 403-410.

3.10. Eddy S.R. Where did the BLOSUM62 alignment score matrix come from? Nat Biotechnol. 2004: 22(8), 1035-6.

3.11. Venter, J. C., Remington, K., Heidelberg, J.F., Halpern, A.L., Rusch D., Eisen J.A., Wu D.Y., Paulsen I., Nelson K.E., Nelson W., Fouts D.E., Levy S., Knap A.H., Lomas M.W., Nealson K., White O., Peterson J., Hoffman J., Parsons R., Baden-Tillson H., Pfannkoch C., Rogers Y.H., Smith H.O. Environmental Genome Shotgun Sequencing of the Sargasso Sea. Science 2004: 304, 66 – 74.

3.12. Watanabe, K., Nelson, J., Harayama, S. and Kasai, H. ICB database: the gyrB database for identification and classification of bacteria. Nucleic Acids Res. 2001: 29(1):344-5.

3.13. Gill, S.R., Pop, M., DeBoy, R.T., Eckburg, P.B., Turnbaugh, P.J., Samuel B.S., Gordon J.I., Relman D.A., Fraser-Liggett C.M., Nelson K.E. Metagenomic Analysis of the Human Distal Gut Microbiome . Science 2006: 312, 1355-1359.

3.14. Tringe, S.G., Von Mering, C., Kobayashi, A. , Salamov, A.A., Chen K., Chang H.W., Podar M., Short J.M., Mathur E.J., Detter J.C., Bork P., Hugenholtz P., Rubin E.M. Comparative Metagenomics of Microbial Communities. Science 2005: 308, 554-557.

3.15. von Mering, C., Hugenholtz, P., Raes, J., Tringe, S.G., Doerks T., Jensen L.J., Ward N., Bork P. Quantitative Phylogenetic Assessment of Microbial Communities in Diverse Environments. Science 2007: 315, 1126-1130.

4. SP Scaffolding of Genomes

4.1 Application of SP analysis to Genomic data

An interesting application of the DME methodology is its utilization in the study of genomic sequences of bacteria and archaea. Results of this analysis can generate EC predictions for the coding sequences of genes. However, beyond EC prediction for genes, we can leverage SPs to derive two additional types of information:

1. Calculate the beginning and end of a gene in a process we label "SP Scaffolding"
2. Utilize SP hits with consistent ECs in adjacent genes to trace genomic evolution

4.2 Prediction of Enzymatic Function for Genes

Using as input the complete genome of an organism, the methodology consists of translating the whole sequence of nucleotides of the genome into each of the six translation frames and searching for all SPs separately in each frame. The SP set used is SP set v2.3 which consists of SPs with a length of 7 or more amino-acids. Our assumption here is that the beginning and end of each of the genes in the genome of the organism are known. Thus, we make a meaningful EC prediction at the gene level. However, this is not a necessary requirement, and below we discuss the generation of delimiters using SP hits.

SP hits on a gene are required to be on the same translated frame in order to be considered contributors to consistent SP coverage of a Coding Data Sequence. Coverage-length of an SP is defined, as in previous references in this work, as the number of amino-acids exactly matching the sequence queried. We can look at the coverage of SP hits on the gene in the same translation frame, and, based on it, an EC prediction can be made for the gene. If the coverage of SPs on a certain frame exceeds our predetermined threshold (usually 9 amino-acids), the gene is assigned the highest EC level of all consistent SPs hitting the gene.

Below, we discuss the choice of L=9 as a coverage threshold to generate a reliable prediction, as it differs from the selection of L=7 used in the proteomic case.

The DME methodology (Weingart et al. 2009, chapter 3 in this thesis)) is based on employing SPs with individual length (i.e. number of amino-acids) L=7 or more. This has also been adopted by the SPSR methodology of (Weingart et al. 2010, chapter 5 below). Here we wish to test its sensitivity on the E. coli genome. The latter contains 4,639,675 nucleotides. We convert it in six possible ways to a long string of amino-acids and search for SP hits on them. We find 20,073 such records. The latter are compared with known NCBI and Swiss-Prot EC annotations.
SP hits on genic regions are classified as true-positive (TP) or false-positive (FP) according to the expert EC annotations. SP hits on intergenic regions serve as a convenient 'negative set' to define random false-positive hits. We find only 60 hits on intergenic sections, with 53 of SP length L=7, 6 of L=8, and 1 of L=10.

The total size of intergenic regions = 4,639,675 (Total Genome Length) -4,132,557 (Total CDS) =507,118 nucleotides.

Analysis of the intergenic regions:
L=7: 53 hits. 53/507118=1.05e-4 hits/nucleotide
L=8: 6 hits, leading to1.2e-5 hits/nucleotide
L=10: 1 hit, meaning 1.97e-6 hits/nucleotide

Analysis of the genic regions, presented as FP/TP (based on expert annotations):
L=7: 965/7485= 0.13
L=8: 237/4235= 0.056
L=9: 100/2346= 0.043
L=10: 68/1361= 0.05
Using the intergenic hits as a valid random model, we expect the following errors to occur in the genic regions:
L=7: 1e-4*4132557=413 false hits
L=8: 1.2e-5*4132557= 49 false hits
L=10: 2e-6*4132557= 8 false hits
Comparing with the larger number of FP predictions, we conclude that the remaining errors are false annotations. In other words, we expect 4-5% of the annotations to be wrong, assuming the SP errors are correctly estimated by the intergenic region analysis. Our main conclusion from this analysis is that, using SPs of length L=9 and 10 we should expect errors of less than 1%.

We gauge the quality of DME predictions using precision and recall metrics, comparing our predictions to Swissprot annotations for the corresponding coding sequences.
In addition, we can look at the impact of using different coverage length threshold on the quality of our predictions, by reviewing precision and recall figures.

When we focus on high accuracy predictions we use L=9, but for general exploration of enzymatic gene structures we are content with L=7 and use that threshold.

4.3 SP Scaffolding

Beyond predicting the functional annotations of genes, we use the SP hits as a scaffolding prediction mechanism to determine approximately the beginning and end of the gene set in a process we label “SP Scaffolding.”
The goal of SP scaffolding is enzymatic gene discovery, which includes annotation of known genes, known pseudo genes and unknown genes.
The last category means genes that were not recognized as such but we find their traces with SP hits.
We define the coding sequence as the domain bounded by the first Met after the 'Stop' that occurs to the left of the left-most SP and the final 'Stop' on the amino-acid sequence on which we observed all SP hits.
As in the cases before, all consistent SP hits must be consistent not only for EC but must be in the same reading frame. Obviously, accidental SP hits in the inspected frame and SP hits in other frames within the same region need to be disregarded.
Our method has the inherent drawback that neither EC predictions nor scaffolding can be done for DNA regions that are not enzymatic. Another obstacle we encounter to

generate good predictions consists of the fact that in some cases, because of the structure of the training set we did not generate SPs applicable to the checked sequences even though they were enzymatic.

SP Scaffolding is demonstrated below using a random domain of H. Pylori 26995, selected so that the consistent SP coverage in the same frame exceeds 50 amino acids. The following table shows the selected domain.
Eight SP hits with EC=1.1.1.267 contribute to a total coverage of 55 amino acids all in frame 1. First SP hit with EC=1.1.1.267 is at nucleotide 224,696 and last SP hits at nucleotide 225,248.

SP Hit Location	EC Of SP	Frame Hit	SP
224,570	2.7.7.41	1	HGGVLDR
224,696	1.1.1.267	1	GSTGSIG
224,930	1.1.1.267	1	SNLVLNAIVGVAGL
225,008	1.1.1.267	1	LALANKE
225,014	1.1.1.267	1	LANKESL
225,152	1.1.1.267	1	ASGGAFRD
225,215	1.1.1.267	1	ALKHPNW
225,224	1.1.1.267	1	HPNWSMG
225,248	1.1.1.267	1	KITIDSA
236,738	1.8.1	1	IGGGSGG

Table 4.1: SP hits for sample a domain in H Pylori 26995 with consistent coverage ≥ 50 with EC=1.1.1.267. Shown also the SP hits in the same frame prior and after the consistent hits domain.

K E T A V F L G D **Stop** **Met** V V L **G S T G S I G** K N A L K I A K
K F G I E I E A L S C G K N I A L I N E Q I Q V F K P K K V A
I L D P S D L N D L E P L G A E V F V G L E G I D A **Met** I E E
C T **S N L V L N A I V G V A G L** K A S F K S L Q R N K K **L A L**
A N K E S L V S A G H L L D I S Q I T P I D S E H F G L W A L
L Q N K T L K P K S L I I S **A S G G A F R D** T P L E F I P I Q
N A Q N **A L K H P N W S Met G** S **K I T I D S A** S **Met** V N K L F
E I L E T Y W L F G A S L K I D A L I E R S S I V H A L V E F
E D N S I I A H L A S A D **Met** Q L P I S Y A I D P K L A S L S
A S I K P L D L Y A L S A I K F E P I S **Met** E R Y T L W C Y K
D L L L E N P K L G V V L N A S N E V A **Met** E K F L N K E I A
F G G L I Q T I S Q A L E S Y D K **Met** P F K L S S L E E V L E
L D K E V R E R F K N V A G V **Stop**

Figure 4.1: Translation of frame 1 of area subsequence starting at nucleotide 224,682 in H. Pylori 26995.

We define the coding sequence as the domain bounded by the first Met after the 'Stop' that occurs to the left of the left-most SP and the final 'Stop' on the amino-acid sequence on which we observed all SP hits.
We predict therefore the start of the coding sequence at 224,692 for a length of 1,104 nucleotides. Our prediction is in agreement both with NCBI's location of the start, length of the gene and its EC annotation.

4.4 Analysis of a full genome: H. pylori

All predictions for H. Pylori are included in 4.6 below.
There are several erroneous predictions using coverage-length L=9. Six of them consist of enzymes where the EC in the Swissprot annotation was modified recently from one that agreed with the DME prediction to a new EC. Extended details of two such cases are described as follows:

HP0086 (Swissprot O24913):
Swissprot annotates this gene as "Malate dehydrogenase [quinone]" with an EC=1.1.5.4. The DME prediction very strong signal with a consistent coverage at EC level 4 of 155 amino acids with an EC=1.1.99.16. Our research shows that EC=1.1.99.16 was transferred by Swissprot into EC=1.1.5.4.

HP1059 (Swissprot O25699)
Swissprot annotates this gene as "Holliday junction ATP-dependent DNA helicase ruvB" and an EC=3.6.4.12, which translates to DNA helicase.
DME predicts EC=3.6.1, which translates to "Hydrolases, acting on acid anhydrides in phosphorous-containing anhydrides" with a very strong signal of L3=150 generated by 25 SP hits in frame 1.
Research of this entry shows that on the 13th of July 2010, the annotation for this entry was modified by Swissprot from EC=3.6.1.- to EC=3.6.4.12.

The examples shown above exemplify the dependency of our training set on Swissprot. This dependency causes DME to be adversely impacted whenever Swissprot renumbers ECs or whenever Swissprot re-annotates proteins after the Training Set of the SPs was generated. A partial remedy to fix this problem is to generate new SP sets as frequent as possible.

Similarly to the proteomic environment, the chances of DME to provide real novelties are the cases of remote homology, with low SP coverages that exceed the threshold to determine a prediction, typically in the region of L3= 9 or 10 amino acids.

87% out of the 225 correct predictions are at EC level 4 and the rest are at EC level 3. The reason for the preponderance of EC level 4 predictions as compared to other test datasets we have studied is that we are conducting SP searches of large areas of complete nucleotide sequences that are rich in enzymatic cDNA sequences.

Figure 4.2 below shows the histogram of all correct predictions by SP coverage at EC level 4, for Helicobacter pylori 26695.

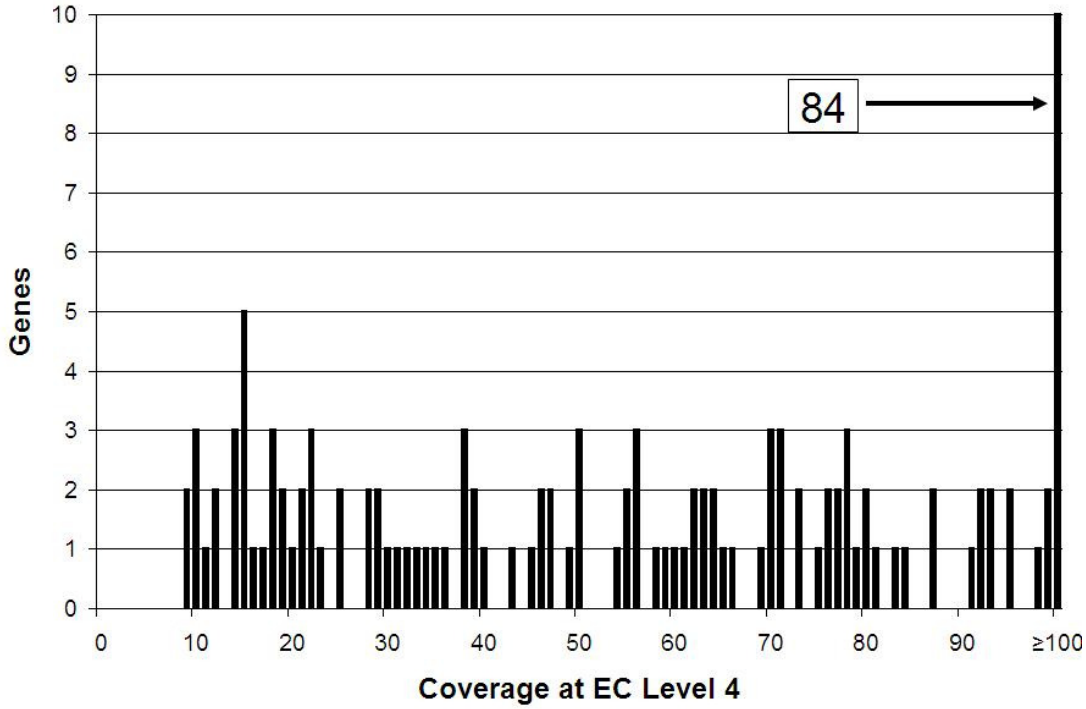


Figure 4.2: Histogram of gene predictions at EC level 4 for H. Pylori 26995

4.5 Locating traces of genetic evolution

Another interesting application of the analysis of SP hits is detection of shifts in nucleotide sequences caused either by addition or deletion of single or few nucleotides in a coding sequence. These are cases of genes that have very high EC consistent coverage. The adjacent gene is hit by a single SP the same consistent EC but in a different frame.

One possible explanation to this phenomenon is that the original large gene included part of the adjacent gene, an indel event occurred, and the boundaries of the genes have been shifted. The set of genes HP0760 and HP0761 in H. Pylori serves as an example of this phenomenon.
Figure 4.3 below is a diagrammatic description of SP hits on genes HP0760 and HP0761 of H. Pylori 26995.

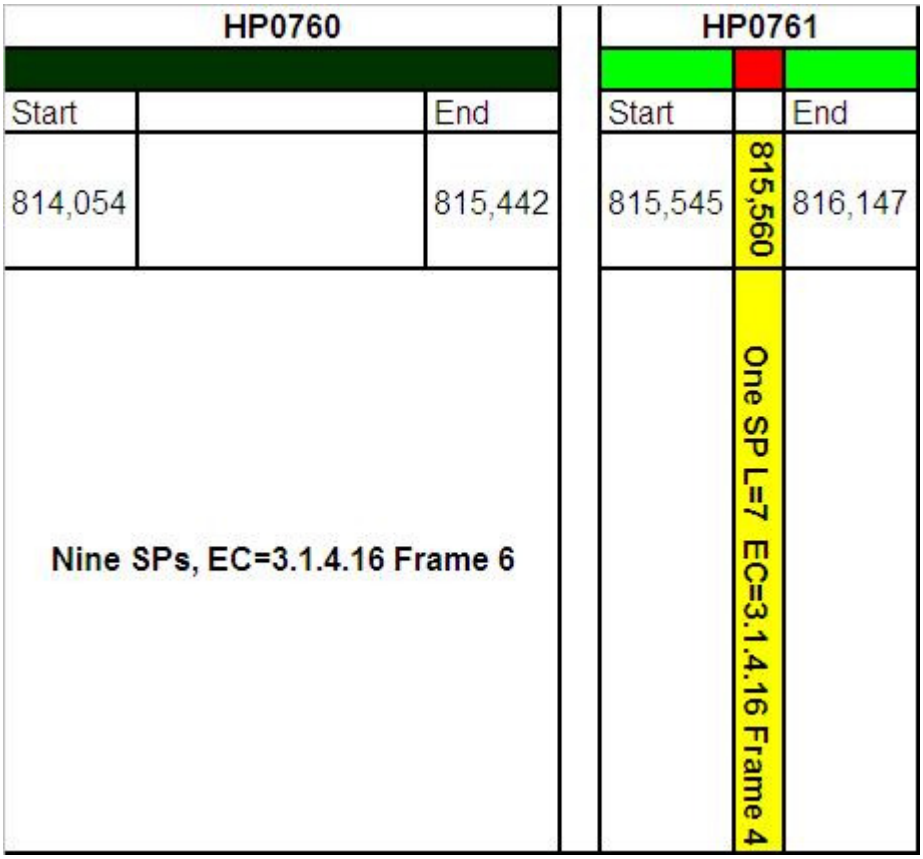


Figure 4.3: Diagrammatic description of SP hits on genes HP0760 and HP0761 of H. Pylori 26995

Gene HP0760 (Swissprot Accession ID O25455) is annotated with an EC=3.1.4.16 (2',3'-cyclic-nucleotide 2'-phosphodiesterase) and spans the locations between 814,054 and 815,442 nucleotides in the complement strand (L=1,388 nucleotides). It is hit by 9 SPs in frame six with EC=3.1.4.16 with a total SP coverage of 113 amino-acids. The adjacent gene, HP0761 (Swissprot Accession ID O25456) spans the locations between 815,545 and 816,147 nucleotides in the complement strand (L=602 nucleotides) and is hit by a single SP, "GLIYISLEV," in frame four with EC=3.1.4.16 as shown in table 4.2 below.

Gene	**HP0760**	**Start Gene**	**814,054**	**End Gene**	815,442
	SP	**EC**	**SP Hit Location**	**Frame**	
	GVEKAYA	3.1.4.16	814,237	6	
	ECASVCAADALSAGRPGARRKSDEEYAKRMQALEEIAL	3.1.4.16	814,267	6	
	AHHGHEE	3.1.4.16	814,393	6	
	LGVEVCKR	3.1.4.16	814,450	6	
	LHDIGKA	3.1.4.16	814,507	6	
	KLARRAG	3.1.4.16	814,531	6	
	LIRRYEK	3.1.4.16	815,086	6	
	MLNYMAYTK	3.1.4.16	815,170	6	
	LKHLEAQHKEFVRDEKRYLEKEK	3.1.4.16	815,299	6	
Gene	HP0761	Start Gene	815,545	End Gene	816,147
	SP	EC	SP Hit Location	Frame	
HP0761	GLIYISLEV	3.1.4.16	815,610	4	

Table 4.2: SP hits of genes HP0760 and HP0761 of H. Pylori 26995
[
We can predict that there was a SNP between the end of gene HP0760, 815442, and the beginning of SP hit GLIYISLEV at 815,560.

4.6 List of predictions for H. Pylori

We used SPs to generate enzymatic predictions and FSPs to generate gene predictions for the whole genome of H. Pylori 26995.
Aggregation of prediction results are presented in Table 4.3:

Pred. Type	**P|P**	**P|DP**	**P|NP**	**NP|P**	**Precision**	**Recall**	**Putative Novelties**
EC	225	7	40	83	97.0%	71.4%	14.7%
Gene	232	2	5	116	99.1%	66.3%	2.1%

Table 4.3: Summary of DME predictions vs. Swissprot annotations for H. Pylori.

Table 4.4 below shows all results except all cases resulting in NP|NP. L3 and L4 are SP coverages on the gene at EC Level 3 and EC Level 4.
Cases that have hits with L3<9 do not generate a DME enzymatic prediction, therefore they are classified in the NP|P or NP|NP category, depending on whether they have a Swissprot annotation.

Table 4.4: List of predictions vs. annotations for H. Pylori 26995:

Locus Tag	EC DME Prediction	EC Swissprot Annotation 1	EC Swissprot Annotation 2	DME Gene Prediction	Swissprot Gene Annotation	L3	L4
HP0002	2.5.1.9	2.5.1.9		RISB	RISB	80	80
HP0003	2.5.1.55	2.5.1.55		KDSA	KDSA	110	110
HP0004		4.2.1.1			CYNT		
HP0005	4.1.1.23	4.1.1.23		PYRF	PYRF	29	29
HP0006	6.3.2.1	6.3.2.1		PANC	PANC	62	62
HP0012		2.7.7			PRIM		
HP0026		2.3.3.1			CISY		
HP0027	1.1.1.42	1.1.1.42		IDH	IDH	87	87
HP0029		6.3.3.3			BIOD		
HP0034	4.1.1.11	4.1.1.11		PAND	PAND	40	40
HP0044	4.2.1.47					34	34
HP0051		2.1.1.37				7	7
HP0054		2.1.1.37					
HP0072	3.5.1.5	3.5.1.5		URE1	URE1	258	258
HP0073	3.5.1.5	3.5.1.5		URE23	URE23	106	106
HP0074	3.4.23.36	3.4.23.36		LSPA	LSPA	32	32
HP0075	5.4.2.10	5.4.2.10		GLMM	GLMM	201	201
HP0086	1.1.99.16	1.1.5.4		MQO	MQO	155	155
HP0089		3.2.2.9			MTNN		
HP0098		4.2.3.1			THRC		
HP0105	4.4.1.21	4.4.1.21		LUXS	LUXS	62	62
HP0106	2.5.1.48	2.5.1.48		METB	METB	9	9
HP0107	2.5.1.47	2.5.1.47			CYSM	16	16
HP0116	5.99.1.2	5.99.1.2		TOP1	TOP1	64	64
HP0121	2.7.9.2	2.7.9.2		PPSA	PPSA	71	71
HP0123	6.1.1.3	6.1.1.3		SYT	SYT	247	235
HP0132		4.3.1.17			SDHL	7	7
HP0134	2.5.1.54					25	25
HP0136		1.11.1.15			BCP		
HP0154	4.2.1.11	4.2.1.11		ENO	ENO	227	227
HP0157	2.7.1.71	2.7.1.71		AROK	AROK	66	66
HP0160		3.5.2.6			HCPD		
HP0163	4.2.1.24	4.2.1.24		HEM2	HEM2	22	22
HP0175	5.2.1.8				Y175	11	11
HP0176	4.1.2	4.1.2.13			ALF	25	0
HP0179	3.6.3					15	0
HP0180		2.3.1			LNT		
HP0182	6.1.1.6	6.1.1.6		SYK	SYK	184	174
HP0183	2.1.2.1	2.1.2.1		GLYA	GLYA	189	189
HP0191		1.3.99.1			FRDB		
HP0192		1.3.99.1			FRDA		
HP0194	5.3.1.1	5.3.1.1		TPIS	TPIS	104	104
HP0195		1.3.1.9			FABI		
HP0196	2.3.1	2.3.1		LPXD	LPXD	56	7
HP0197	2.5.1.6	2.5.1.6		METK	METK	241	241

Locus Tag	EC DME Prediction	EC Swissprot Annotation 1	EC Swissprot Annotation 2	DME Gene Prediction	Swissprot Gene Annotation	L3	L4
HP0198	2.7.4.6	2.7.4.6		NDK	NDK	71	71
HP0201		2.3.1.n2			PLSX		
HP0202	2.3.1	2.3.1.180		FABH	FABH	113	0
HP0211		3.5.2.6			HCPA		
HP0212	3.5.1.18	3.5.1.18		DAPE	DAPE	144	144
HP0215		2.7.7.41			CDSA	7	7
HP0216	1.1.1.267	1.1.1.267		DXR	DXR	55	55
HP0220	2.8.1.7	2.8.1.7			ISCS	38	38
HP0224	1.8.4.12	1.8.4.11	1.8.4.12	MSRA, MSRB - Dual function	MSRAB	28	28
HP0230	2.7.7.38	2.7.7.38		KDSB	KDSB	109	109
HP0235		3.5.2.6			HCPE		
HP0237	2.5.1.61	2.5.1.61		HEM3	HEM3	158	158
HP0238	6.1.1.15	6.1.1.15		SYP	SYP	289	289
HP0239	1.2.1.70	1.2.1.70		HEM1	HEM1	102	102
HP0244		2.7.13.3				7	7
HP0247	3.6.1					61	0
HP0250	3.6.3					40	0
HP0255	6.3.4.4	6.3.4.4		PURA	PURA	185	182
HP0258	3.4.24	3.4.24			Y258	21	0
HP0259	3.1.11.6	3.1.11.6		EX7L	EX7L	173	173
HP0281	2.4.2.29	2.4.2.29		TGT	TGT	160	160
HP0283	4.2.3.4	4.2.3.4		AROB	AROB	132	132
HP0286	3.4.24					10	0
HP0290	4.1.1.20	4.1.1.20		DCDA	DCDA	22	15
HP0294	3.5.1.4	3.5.1.4		AMIE	AMIE	180	180
HP0301	3.6.3					36	0
HP0306	5.4.3.8	5.4.3.8		GSA	GSA	196	196
HP0319	6.1.1.19	6.1.1.19		SYR	SYR	194	186
HP0321	2.7.4.8	2.7.4.8		KGUA	KGUA	29	29
HP0328	2.7.1.130	2.7.1.130		LPXK	LPXK	80	80
HP0329	6.3.1.5	6.3.1.5		NADE	NADE	77	77
HP0330	1.1.1.86	1.1.1.86		ILVC	ILVC	71	71
HP0336		3.5.2.6			HCPB		
HP0339		3.2.1.17					
HP0347		5.4.99			Y347	7	0
HP0349	6.3.4.2	6.3.4.2		PYRG	PYRG	171	171
HP0354	2.2.1.7	2.2.1.7		DXS	DXS	182	182
HP0360	5.1.3					9	0
HP0361	5.4.99.12	5.4.99.12		TRUA	TRUA	93	93
HP0363	2.1.1.77	2.1.1.77			PIMT	15	15
HP0364		1.17.4.1			RIR2		
HP0370	6.4.1.1					15	15
HP0372	3.5.4.13	3.5.4.13		DCD	DCD	77	77
HP0374		2.1.1			RSME		
HP0376	4.99.1.1	4.99.1.1		HEMH	HEMH	95	95

Locus Tag	EC DME Prediction	EC Swissprot Annotation 1	EC Swissprot Annotation 2	DME Gene Prediction	Swissprot Gene Annotation	L3	L4
HP0380	1.4.1.4	1.4.1.4			DHE4	79	34
HP0387		3.6.4			PRIA	7	0
HP0388	2.1.1	2.1.1		CMOA	CMOA	73	7
HP0389	1.15.1.1	1.15.1.1			SODF	19	19
HP0390		1.11.1			TPX	7	0
HP0392	2.7.13.3					11	11
HP0400	1.17.1.2	1.17.1.2		ISPH	ISPH	123	123
HP0401	2.5.1.19	2.5.1.19		AROA	AROA	111	111
HP0402		6.1.1.20			SYFB	7	7
HP0403	6.1.1.20	6.1.1.20		SYFA	SYFA	134	122
HP0409	6.3.5.2	6.3.5.2		GUAA	GUAA	230	230
HP0417	6.1.1.10	6.1.1.10		SYM	SYM	65	58
HP0419	2.1.1	2.1.1		CMOB	CMOB	115	0
HP0422	4.1.1.19	4.1.1.19				19	15
HP0440	5.99.1.2					9	9
HP0475	3.6.3.31			POTA		26	9
HP0476	6.1.1.17	6.1.1.17			SYE1	171	164
HP0493	2.7.8.13	2.7.8.13		MRAY	MRAY	119	119
HP0494	6.3.2.9	6.3.2.9		MURD	MURD	50	50
HP0496		3.1.2			Y496	8	0
HP0500		2.7.7.7			DPO3B		
HP0501	5.99.1.3	5.99.1.3		GYRB	GYRB	121	121
HP0510	1.3.1.26	1.3.1.26		DAPB	DAPB	93	93
HP0512	6.3.1.2	6.3.1.2			GLNA	31	31
HP0515	3.4.25	3.4.25		HSLV	HSLV	89	0
HP0527	3.5.1					14	0
HP0549	5.1.1.3	5.1.1.3		MURI	MURI	100	99
HP0550	3.6.1	3.6.4		RHO	RHO	68	0
HP0552		2.1.1			RSMI		
HP0557	6.4.1.2	6.4.1.2		ACCA	ACCA	127	127
HP0558	2.3.1					22	7
HP0566	5.1.1.7	5.1.1.7		DAPF	DAPF	120	120
HP0570	3.4.11.1	3.4.11.1		AMPA	AMPA	171	171
HP0572	2.4.2.7	2.4.2.7		APT	APT	75	75
HP0576		3.4.21.89			LEP	7	7
HP0577		1.5.1.5	3.5.4.9		FOLD		
HP0581	3.5.2.3	3.5.2.3		PYRC	PYRC	105	105
HP0598		2.3.1.47	2.3.1.29		BIKB	7	7
HP0604	4.1.1.37	4.1.1.37		DCUP	DCUP	129	129
HP0615	6.5.1.2	6.5.1.2		DNLJ	DNLJ	92	92
HP0617	6.1.1.12	6.1.1.12		SYD	SYD	261	245
HP0618	2.7.4.3	2.7.4.3		KAD	KAD	81	81
HP0620		3.6.1.1			IPYR	8	8
HP0623	6.3.2.8	6.3.2.8		MURC	MURC	114	114
HP0625	1.17.7.1	1.17.7.1		ISPG	ISPG	158	158
HP0643	6.1.1.17	6.1.1		SYE2	SYE2	182	175
HP0646	2.7.7.9					10	10
HP0648	2.5.1.7	2.5.1.7		MURA	MURA	171	171

Locus Tag	EC DME Prediction	EC Swissprot Annotation 1	EC Swissprot Annotation 2	DME Gene Prediction	Swissprot Gene Annotation	L3	L4
HP0649	4.3.1.1	4.3.1.1		ASPA	ASPA	76	76
HP0653		1.16.3.1			FTN	8	8
HP0658	6.3.5	6.3.5		GATB	GATB	145	0
HP0661	3.1.26.4	3.1.26.4		RNH	RNH	36	36
HP0662	3.1.26.3	3.1.26.3		RNC	RNC	95	95
HP0663	4.2.3.5	4.2.3.5		AROC	AROC	150	150
HP0665	1.3.99.22	1.3.99.22		HEMN	HEMN	21	21
HP0680	1.17.4.1	1.17.4.1			RIR1	14	14
HP0683		2.7.7.23	2.3.1.157		GLMU		
HP0690	2.3.1.9					41	16
HP0691	2.8.3.5	2.8.3.5			SCOA	18	18
HP0692	2.8.3.5	2.8.3.5			SCOB	23	23
HP0700		2.7.1.107			KDGL		
HP0701	5.99.1.3	5.99.1.3		GYRA	GYRA	120	120
HP0705	3.6.3				UVRA	15	0
HP0707	2.1.1	2.1.1		MRAW	RSMH	123	0
HP0715	3.6.3.25			LPTB		25	9
HP0723	3.5.1.1	3.5.1.1			ASPG	34	10
HP0728		6.3.4			TILS		
HP0734	2.3.1				RIMO	14	0
HP0736		2.6.1			Y736		
HP0738	6.3.2.4	6.3.2.4		DDL	DDL	56	56
HP0742	2.7.6.1	2.7.6.1		KPRS	KPRS	43	43
HP0745	5.4.99	5.4.99			Y745	17	0
HP0747	2.1.1.33	2.1.1.33		TRMB	TRMB	45	38
HP0748	3.6.3					25	0
HP0760	3.1.4.16	3.1.4.16		CNPD	CNPD	113	113
HP0761	3.1.4.16					9	9
HP0774	6.1.1.1	6.1.1.1		SYY	SYY	80	73
HP0776		2.7.7.6			RPOZ	8	8
HP0777	2.7.4.22	2.7.4.22		PYRH	PYRH	83	83
HP0779	4.2.1.3	4.2.1.3		ACON2	ACON2	47	47
HP0791		3.6.3.3	3.6.3.5		HMCT	7	0
HP0793	3.5.1.88	3.5.1.88		DEF	DEF	73	73
HP0794	3.4.21.92	3.4.21.92		CLPP	CLPP	117	117
HP0799		2.7.7.n5			MOG		
HP0802	3.5.4.25	3.5.4.25		RIBA	RIBA	39	39
HP0804	4.1.99.12	4.1.99.12			RIBB	25	25
HP0808	2.7.8.7	2.7.8.7		ACPS	ACPS	46	46
HP0822	1.1.1.3	1.1.1.3			DHOM	9	9
HP0825		1.8.1.9			TRXB		
HP0829	1.1.1.205	1.1.1.205		IMDH	IMDH	63	63
HP0830	6.3.5	6.3.5		GATA	GATA	205	0
HP0831		2.7.1.24			COAE		
HP0832	2.5.1.16	2.5.1.16		SPEE	SPEE	99	92
HP0843	2.5.1.3	2.5.1.3		THIE	THIE	22	22
HP0844		2.7.1.49	2.7.4.7		THID		
HP0845	2.7.1.50	2.7.1.50		THIM	THIM	69	69

Locus Tag	EC DME Prediction	EC Swissprot Annotation 1	EC Swissprot Annotation 2	DME Gene Prediction	Swissprot Gene Annotation	L3	L4
HP0846	3.1.21.3					14	14
HP0854	1.7.1.7	1.7.1.7		GUAC	GUAC	161	161
HP0857	5.3.1	5.3.1		GMHA	GMHA	38	0
HP0858		2.7.1	2.7.7		HLDE	7	7
HP0859	5.1.3.20			HLDD		12	12
HP0860		3.1.3			GMHB		
HP0862		2.7.1.33			COAX		
HP0865	3.6.1.23	3.6.1.23		DUT	DUT	39	39
HP0867	2.4.1.182	2.4.1.182		LPXB	LPXB	126	126
HP0871	3.6.1.26	3.6.1.26		CDH	CDH	60	60
HP0875	1.11.1.6	1.11.1.6		CATA	CATA	113	113
HP0877	3.1.22.4	3.1.22.4		RUVC	RUVC	76	76
HP0883	3.6.1	3.6.4.12		RUVA	RUVA	73	0
HP0886	6.1.1.16	6.1.1.16		SYC	SYC	245	245
HP0888	3.6.3				Y888	12	0
HP0919	6.3.5.5	6.3.5.5		CARB	CARB	199	199
HP0921	1.2.1.12	1.2.1.12		G3P	G3P	27	12
HP0922	1.1.1					14	7
HP0924		5.3.2			Y924		
HP0926	5.4.99	5.4.99		TRUD	TRUD	79	0
HP0927	3.4.24	3.4.24		HTPX	HTPX	77	0
HP0928	3.5.4.16	3.5.4.16		GCH1	GCH1	65	65
HP0930	3.1.3.5	3.1.3.5		SURE	SURE	103	103
HP0941	5.1.1.1	5.1.1.1		ALR	ALR	157	157
HP0949	2.1.1	2.1.1		RLMH	RLMH	50	0
HP0950	6.4.1.2					24	24
HP0955	2.4.99	2.4.99		LGT	LGT	88	0
HP0956		5.4.99			Y956	7	0
HP0960	6.1.1.14	6.1.1.14		SYGA	SYGA	150	150
HP0961	1.1.1.94	1.1.1.94		GPDA	GPDA	98	98
HP0972	6.1.1.14	6.1.1.14		SYGB	SYGB	208	201
HP0974	5.4.2.1	5.4.2.1		GPMI	GPMI	99	99
HP0975		6.3.5			GATC		
HP0976	2.6.1.62	2.6.1.62		BIOA	BIOA	14	11
HP0980	3.4.24					14	0
HP0981	3.1.11.6			EX7L		19	19
HP1010	2.7.4.1	2.7.4.1		PPK	PPK	50	50
HP1011	1.3.3.1	1.3.5.2		PYRD	PYRD	131	131
HP1013	4.2.1.52	4.2.1.52		DAPA	DAPA	129	129
HP1019	3.4.21					23	0
HP1020		2.7.7.60	4.6.1.12		ISPDF		
HP1026	3.4.21.53					9	9
HP1036		2.7.6.3			HPPK		
HP1038	4.2.1.10	4.2.1.10		AROQ	AROQ	22	22
HP1045	6.2.1.1	6.2.1.1		ACSA	ACSA	243	236
HP1050	2.7.1.39	2.7.1.39		KHSE	KHSE	38	35
HP1052	3.5.1	3.5.1		LPXC	LPXC	57	0
HP1058	2.1.2.11	2.1.2.11		PANB	PANB	115	115

Locus Tag	EC DME Prediction	EC Swissprot Annotation 1	EC Swissprot Annotation 2	DME Gene Prediction	Swissprot Gene Annotation	L3	L4
HP1059	3.6.1	3.6.4.12		RUVB	RUVB	150	0
HP1063	2.1.1	2.1.1		RSMG	RSMG	53	7
HP1068	2.1.1	2.1.1		PRMA	PRMA	20	0
HP1069	3.4.24	3.4.24			FTSH	75	0
HP1071		2.7.8.8			PSS		
HP1072	3.6.3.4	3.6.3.4			COPA	189	162
HP1082	3.6.3.43					21	14
HP1084	2.1.3.2	2.1.3.2		PYRB	PYRB	131	131
HP1088	2.2.1.1					29	22
HP1098		3.5.2.6			HCPC		
HP1100	4.2.1.12	4.2.1.12		EDD	EDD	32	18
HP1101	1.1.1.49	1.1.1.49			G6PD	18	18
HP1102		3.1.1.31			6PGL	7	7
HP1103	2.7.1.2	2.7.1.2		GLK	GLK	122	122
HP1112	4.3.2.2	4.3.2.2		PUR8	PUR8	15	15
HP1121		2.1.1.37					
HP1123		5.2.1.8			SLYD		
HP1132	3.6.3.14	3.6.3.14		ATPB	ATPB	291	262
HP1134	3.6.3.14	3.6.3.14		ATPA	ATPA	276	252
HP1141	2.1.2.9	2.1.2.9		FMT	FMT	117	117
HP1148	2.1.1.31	2.1.1.31		TRMD	TRMD	47	47
HP1153	6.1.1.9	6.1.1.9		SYV	SYV	156	134
HP1155	2.4.1.227	2.4.1.227		MURG	MURG	183	183
HP1158		1.5.1.2			P5CR		
HP1160	3.4.24	3.4.24			Y1160	41	0
HP1166	5.3.1.9	5.3.1.9		G6PI	G6PI	172	172
HP1171	3.6.3.31					24	10
HP1178	2.4.2.1	2.4.2.1		DEOD	DEOD	17	17
HP1179	5.4.2.7	5.4.2.7		DEOB	DEOB	148	148
HP1189	1.2.1.11	1.2.1.11		DHAS	DHAS	12	12
HP1190	6.1.1.21	6.1.1.21		SYH	SYH	216	216
HP1198	2.7.7.6	2.7.7.6		RPOBC	RPOBC	1268	1246
HP1206	3.6.3					9	0
HP1210	2.3.1.30	2.3.1.30			CYSE	21	21
HP1213	2.7.7.8	2.7.7.8		PNP	PNP	87	87
HP1218		6.3.4.13			PUR2		
HP1220	3.6.3.25					12	9
HP1221	2.5.1.31	2.5.1.31			UPPS	33	33
HP1228	3.6.1	3.6.1		RPPH	RPPH	62	0
HP1229	2.7.2.4	2.7.2.4		AK	AK	19	19
HP1237	6.3.5.5	6.3.5.5		CARA	CARA	54	46
HP1238	3.5.1.49	3.5.1.49		AMIF	AMIF	201	201
HP1241	6.1.1.7	6.1.1.7		SYA	SYA	372	372
HP1248	3.1.13				RNR	11	0
HP1249	1.1.1.25	1.1.1.25		AROE	AROE	78	78
HP1253	6.1.1.2	6.1.1.2			SYW	27	20
HP1257	2.4.2.10	2.4.2.10		PYRE	PYRE	78	78
HP1259	3.5.1	3.5.1			NPD	9	0

Locus Tag	EC DME Prediction	EC Swissprot Annotation 1	EC Swissprot Annotation 2	DME Gene Prediction	Swissprot Gene Annotation	L3	L4
HP1260		1.6.99.5					
HP1261	1.6.99.5	1.6.99.5			NUOB	50	50
HP1262		1.6.99.5				7	7
HP1263	1.6.99.5	1.6.99.5		NUOD	NUOD	131	131
HP1267	1.6.99.5	1.6.99.5		NUOH	NUOH	70	70
HP1268	1.6.99.5	1.6.99.5		NUOI	NUOI	78	78
HP1270		1.6.99.5			NUOK		
HP1275	5.4.2.10			GLMM		10	10
HP1277	4.2.1.20	4.2.1.20		TRPA	TRPA	139	139
HP1278	4.2.1.20	4.2.1.20			TRPB	117	117
HP1279	4.1.1.48	4.1.1.48	5.3.1.24		TRPC	25	25
HP1280	2.4.2.18	2.4.2.18		TRPD	TRPD	79	79
HP1281		4.1.3.27			TRPG		
HP1282	4.1.3.27	4.1.3.27		TRPE	TRPE	70	70
HP1293	2.7.7.6	2.7.7.6		RPOA	RPOA	56	49
HP1299		3.4.11.18			AMPM		
HP1323	3.1.26.4	3.1.26.4		RNH2	RNH2	115	115
HP1325	4.2.1.2	4.2.1.2		FUMC	FUMC	145	145
HP1335	2.8.1	2.8.1		MNMA	MNMA	104	0
HP1337		2.7.7.18			NADD		
HP1345	2.7.2.3	2.7.2.3			PGK	45	45
HP1347	3.2.2	3.2.2.27		UNG	UNG	61	0
HP1348		2.3.1.51			PLSC		
HP1355		2.4.2.19			NADC		
HP1356		2.5.1.72			NADA	8	8
HP1357	4.1.1.65	4.1.1.65		PSD	PSD	54	54
HP1362		3.6.4.12			DNAB	7	0
HP1364		2.7.13.3					
HP1375	2.3.1.129	2.3.1.129		LPXA	LPXA	91	91
HP1376	4.2.1	4.2.1		FABZ	FABZ	55	0
HP1379	3.4.21.53	3.4.21.53			LON	56	56
HP1385	3.1.3.11	3.1.3.11		F16PA	F16PA	63	56
HP1386	5.1.3.1	5.1.3.1			RPE	10	10
HP1394	2.7.1.23	2.7.1.23		PPNK	PPNK	126	119
HP1399		3.5.3.1					
HP1406		2.8.1.6			BIOB		
HP1413	1.7.1.13	1.7.1.13		QUEF	QUEF	63	63
HP1415	2.5.1.8	2.5.1.75		MIAA	MIAA	120	120
HP1418		1.1.1.158			MURB		
HP1420	3.6.3.14	3.6.3.14			FLII	15	15
HP1422	6.1.1.5	6.1.1.5		SYI	SYI	243	229
HP1428	2.1.1	2.1.1		RLMN	RLMN	135	0
HP1431	2.1.1	2.1.1		KSGA	RSMA	92	0
HP1441		5.2.1.8			PPIA		
HP1443	2.7.1.148	2.7.1.148		ISPE	ISPE	30	30
HP1448	3.1.26.5	3.1.26.5		RNPA	RNPA	38	38
HP1459		5.4.99			Y1459		
HP1460	2.7.7.7	2.7.7.7		DPO3A	DPO3A	70	70

Locus Tag	EC DME Prediction	EC Swissprot Annotation 1	EC Swissprot Annotation 2	DME Gene Prediction	Swissprot Gene Annotation	L3	L4
HP1468	2.6.1.42	2.6.1.42			ILVE	14	14
HP1470	2.7.7.7	2.7.7.7			DPO1	28	28
HP1474	2.7.4.9	2.7.4.9		KTHY	KTHY	84	84
HP1475	2.7.7.3	2.7.7.3		COAD	COAD	59	59
HP1476		4.1.1			PAAD	7	0
HP1478	3.6.1					24	0
HP1480	6.1.1.11	6.1.1.11		SYS	SYS	177	177
HP1482		3.1.11.6			EX7S		
HP1494	6.3.2	6.3.2.13			MURE	17	8
HP1495	2.2.1.2	2.2.1.2		TAL	TAL	55	55
HP1497	3.1.1.29	3.1.1.29		PTH	PTH	64	64
HP1503	3.6.3					14	0
HP1509		2.3.1.n3			PLSY		
HP1513		2.9.1.1			SELA		
HP1523	3.6.1	3.6.4.12			RECG	21	0
HP1532	2.6.1.16	2.6.1.16			GLMS	61	61
HP1533		2.1.1.148			THYX		
HP1540		1.10.2.2				7	7
HP1541	3.6.1	3.6.4			MFD	36	0
HP1547	6.1.1.4	6.1.1.4		SYL	SYL	286	286
HP1563	1.11.1.15	1.11.1.15			TSAA	14	14
HP1576	3.6.3.28	3.6.3		METN	METN	111	10
HP1582	2.6.99.2	2.6.99.2		PDXJ	PDXJ	138	138
HP1583	1.1.1.262	1.1.1.262		PDXA	PDXA	127	127
HP1584	3.4.24.57	3.4.24.57		GCP	GCP	142	142

Transformation between SP hit location and corresponding nucleotide

Transformation formulae relating the location of the SP hit to the corresponding nucleotide location in the genomic sequence for every frame.

Using the following notation:
G: = Length of the genome
X: = Location of SP hit (in amino acids)
L: = Length of hitting SP
F: = Frame number {1,2,3,4,5 or 6}

S: = Start location of SP hit (in nucleotides)
E: = End location of SP hit (in nucleotides)

	S = Start location of SP hit (in nucleotides)	E=End location of SP hit (in nucleotides)
Forward strand	3(X-1) +F	3(X -1) + 3L+ (F-1)
Reverse strand	G- 3(X-1) – 3L + (5-F)	G -3(X-1) +(4-F)

Table 4.6: Transformation formulae relating the location of the SP hit to the corresponding nucleotide location in the genomic sequence for every frame

5. Deriving enzymatic and taxonomic signatures of metagenomes from short reads

Our article, “Deriving enzymatic and taxonomic signatures of metagenomes from short read data” published in BMC Bioinformatics 2010, 11:390doi:10.1186/1471-2105-11-390 is included in its entirety as part of this work.

5.1 Background

Characterizing complex microbial ecosystems remains a challenge for metagenomics. Environments such as soil, containing many thousands of species require massive sequencing power to obtain a reasonable coverage of the microbial community. In practice this means that such studies may suffer from highly incomplete sampling, see for example Tringe et al. [5.1]. The so called "deep sequencing" technologies offer hope due their tremendously high-throughput – the Illumina Genome analyzer (Illumina) and the SOLiD 3 (Life Technologies) can currently produce over 10 Gb, and up to 40 Gb of high quality reads, respectively. However these fantastic capacities come with a price – a short read length that currently stands at 100 bases or lower for both these technologies. For a recent review of experimental and computational achievements and challenges in metagenomics see Wooley et al. [5.2].

Unlike a bacterial genome, where short reads can be compensated for by using paired ends and relying on assembly, a highly complex metagenome will often not enable such assembly, and the short individual reads will therefore constitute the data from which information has to be extracted. Of course, getting significant BLAST hits with queries of 100 nucleotides or below is challenging, which results in no match that can be assigned a putative function for the vast majority of sequence reads. In the seminal paper by Dinsdale and coworkers [5.3] using reads of 105 bases and below, most of the biomes investigated yielded less than 20% BLAST hits, many of which could not be ascribed a function.

Conventionally, one first tries to reconstruct a long contig from short reads. The contigs are then analyzed for open reading frames (ORFs) which may be translated into putative proteins. The function of the putative proteins can be deduced by comparing them with known proteins whose sequence similarity is high enough (e.g. very low BLAST e-values) to warrant such predictions. This can be improved by combining various methods such as studying both phylogeny and function [5.4]. The problems of handling and analyzing these environmental data have been recently discussed by Raes and Bork [5.5].

We propose to forego some of the stages used in conventional analysis and consider the multitude of available short reads directly. This can allow us to gather ***inclusive*** information. We use this term to imply functional information on the aggregate of all data rather than the ***exclusive*** information specifying what are the exact genes present and to which species these genes belong. Here we present such a tool employing peptide-based enzymatic signatures and demonstrate its application to quality control and functional investigation of metagenomic data.

Extending the peptide-based approach, we can also derive taxonomic signatures from metagenomic short reads. Current technologies for estimating microbial phylogenetic diversity of metagenomes involve calculation of similarity between sequences encoding rRNAs to database entries such as the ones available in the Ribosomal Database Project, RDP [5.6]. This procedure requires the expensive operation of assembly of contigs, and is based on the premise that 16S rRNA sequences provide a suitable basis for taxa-separations, defining operational taxonomic units (OTUs) [5.7]. Our approach differs from this conventional method in two respects: first we deal directly with short reads, second we do not employ the 16S rRNA as the taxonomic indicator. Instead we use SPs of aminoacyl tRNA synthetases (aaRS) for taxonomic indication.

Recently, the algorithm of CARMA [5.8] was introduced to provide phylogenetic classification directly from short reads. It is composed of two components: detection of Pfam domain and protein family fragments (EGTs) that are conserved in an environmental sample and reconstruction of a phylogenetic tree for each matching Pfam family. The authors state that environmental gene tags as short as 27 amino acids can accurately be classified with high specificity. We provide an accurate alternative to this approach, based on peptides of lengths 7 amino acids and higher, and therefore more suitable for short read data.

The workflow of our paper is the following:

a. Based on the concept of Specific Peptides (SPs) we propose their direct application to short read (SR) analysis.
b. We derive factors that reflect the ratio between counts of SPs, corresponding to a specific EC category, on a set of SRs of a genome or a metagenome, and the numbers of enzyme sequences carrying the same EC annotation on the genome or the metagenome. This is exemplified first on *Escherichia coli* data and further developed on artificial metagenomes of known bacteria, relying on their genomic sequences and enzymatic annotations of their proteins in Uniprot.
c. We develop the concept of TSPs, taxa-specific SPs, using amino-acyl tRNA synthetases that are known to appear only once per species. The determinations of which SPs are taxon-specific, and their associated factors, are derived from all enzymatic data of Swiss-Prot.

The methodology is explained in detail in the following section, and then exemplified and tested in the Results sections.

Methods

5.2 The Specific Peptides Approach.

Kunik et al. [5.9] have extracted very short (~8aa) deterministic motifs, named Specific Peptides (SPs), whose presence in the protein sequence is a good marker for enzymatic functions. The use of motifs has a long history in bioinformatics [5.10; 5.11]. It is only recently, however, that the increasing amounts of annotated protein data, combined with novel motif-extraction techniques [5.12], allowed extracting short SPs and using them with good precision and recall values**.** SPs are strings of amino-acids, extracted from enzyme sequences using the motif extraction algorithm

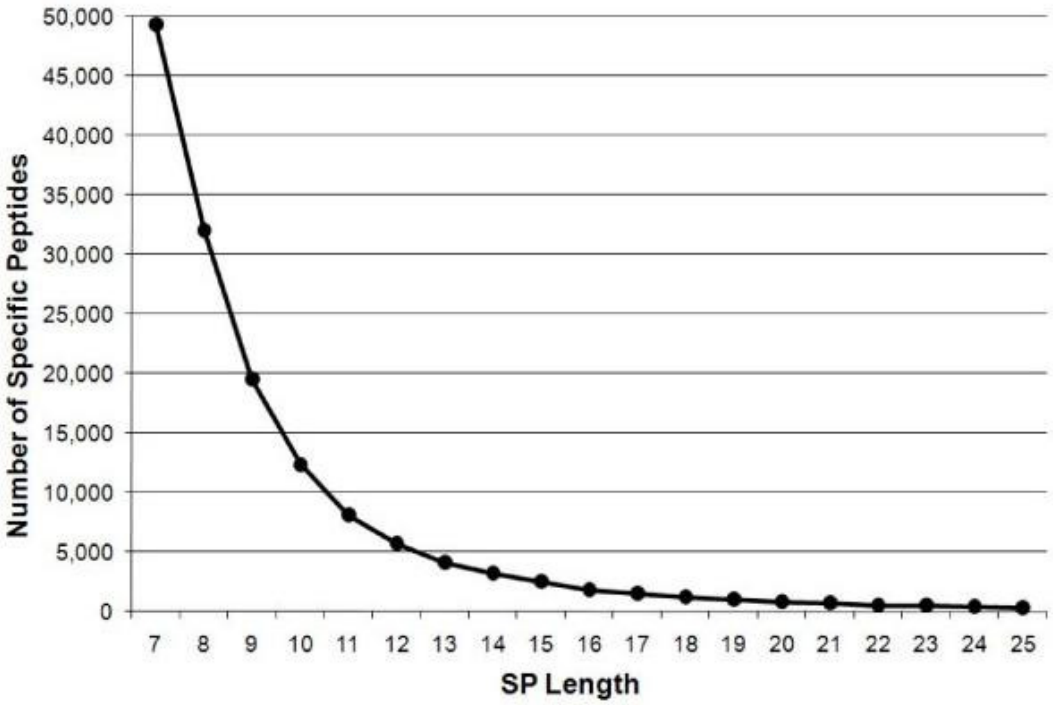
MEX [5.12]. They are selected for their specificity to levels of the Enzyme Commission (EC) 4-level functional hierarchy. Weingart et al [5.13] have demonstrated how SPs can be employed for Data Mining of Enzymes (DME) on any given ensemble of protein sequences. Their methodology relies on coverage length (L, overall number of amino-acids) of SP hits that carry the same EC assignments. In their analysis, L≥7 has led to highly accurate results. They have also updated the SP list, extracting them from a training set of Swiss-Prot data dated July 27th, 2009. This set includes 257,598 SPs of length ≥ 7 with labels corresponding to EC levels 3 and 4. The latter are further filtered for redundancy to discard any SP that contains within it a shorter SP with the same EC specification. This leaves us with a final set of 148,395 SPs that we use in our analysis. Testing the DME approach on a set of 19,849 enzymes that were integrated into Swiss-Prot from July 28th until September 22nd 2009, Weingart et al [5.13] obtained precision of 99.2% and recall of 92.4%, thus vouching for the high quality of DME predictions at the 3rd level of the EC hierarchy.

Here we propose using an SP search on raw Short Read (SR) data, independent of gene reconstruction. Available reads of k nucleotides, where $50 \leq k \leq 200$, may be turned into peptide candidates in six possible ways, counting 3 possible ORFs and 2 possible strands. Each of these pseudo-peptides is checked for SP hits. The latter are required to reside completely within the pseudo-peptides and have a length of 7 amino-acids or more. Ignoring shorter matches has proved to reduce considerably the number of false positive hits in various trial runs. This reliance on k=7 and higher k-mers agrees with the DME methodology of Weingart et al [5.13].

Given a set of short reads we try to obtain a prediction of the number of enzymes in the different EC categories that are expected to be found in the studied metagenome. For that we have to develop a method that relates the number of SP hits observed on a given ensemble of short reads to the expected number of related genes. We define this ratio as the raw-factor, RF(EC) = (number of SP hits)/(number of enzymatic genes) defined for each EC category. To explain this concept we will first illustrate it on a single organism and then proceed to derive it for suitable metagenomes.

5.3 The SPSR methodology: Training on Escherichia coli.

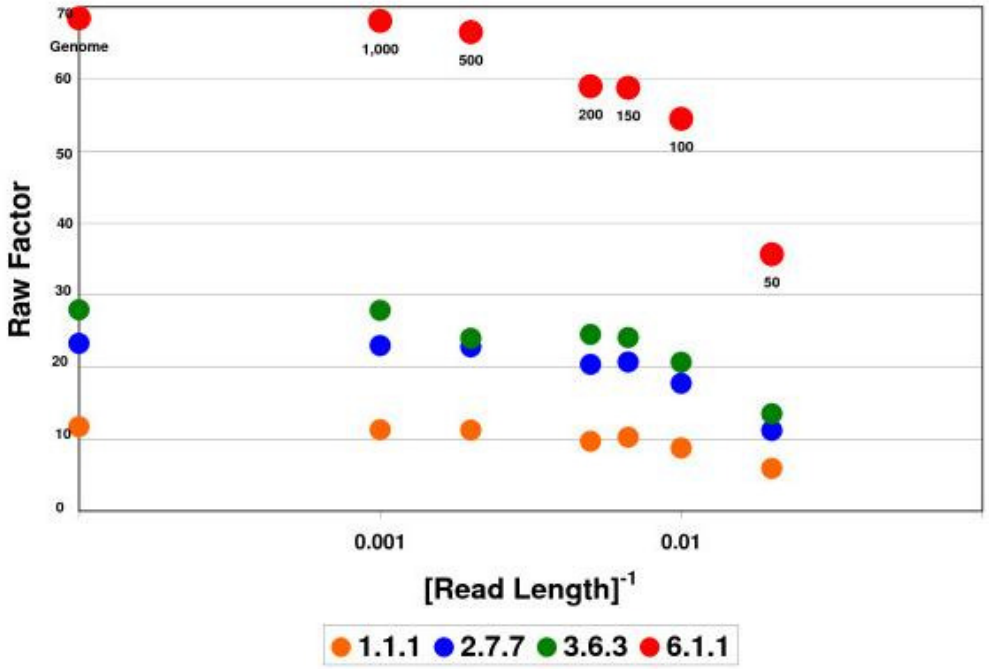
Here we study the derivation and meaning of factors on *E. coli*, making use of its well-studied genome and its well-annotated genes. We notice that if we insert the full genomic sequence instead of short reads in the evaluation of the RFs, these factors coincide with the average number of SP hits on an enzyme within each EC category. Given the genome, we generate SRs randomly, making sure we obtain a 5-fold coverage of the full genome. Calculating the raw-factors, we realize that they vary as we change the length of our SRs. The RFs for finite short read lengths are always lower than their asymptotic values, because SP lengths have to fit inside the lengths of the SRs. Figure 5.1 below displays the distribution of SP lengths for all EC categories. It allows us to estimate the reduced efficiency of SP detection according to the length of the SR.



Distribution of SP lengths. Distribution of numbers of SPs as function of their length (number of amino-acids).

Figure 5.1 – Histogram of SP lengths – represents the distribution of numbers of SPs as function of their length (number of amino-acids).

Thus for a 50 nucleotide short read, no SP hit is expected with length larger than 16 amino-acids. Given this geometrical constraint, the relative efficiency of observation of an SP with length L amino-acids, will be (17-L)/16, just by counting the number of times it can fit into a window of 16 amino-acids. Given the distribution in Figure 5.1 we estimate the total efficiency for a 50 nucleotide short read to be 0.48. Similarly, we estimate the efficiency for SRs of 100 and 200 nucleotides, to be 0.73 and 0.87 respectively. In practice the numbers may vary somewhat between EC categories, since their SP length distributions are not all equal to one another.
Testing this procedure on *E. coli*, we obtain the factors displayed in Figure 5.2 below, following the general trend explained above.



Raw factors. Factors defined as (number of SP hits)/(number of enzymatic genes) for *E. coli* on SR ensembles with different lengths (numbers of nucleotides). Displayed are four 3rd-level EC categories that possess the largest factors. The x-scale is the inverse of the read-length. The values of the different lengths are denoted within the figure. The extreme points at 1/read-length = 0 correspond to factors evaluated from the full genome length. The latter coincide with values derived from the full proteome.

Figure 5.2 - Raw factors.

The 3rd level EC category with the largest factor is 6.1.1, the aminoacyl tRNA synthetases (aaRS). Since all SPs are subject to similar constraints, we observe that if we measure the relative amounts of different EC categories, as shown in Figure 5.3, they remain approximately constant as we vary the SR length.

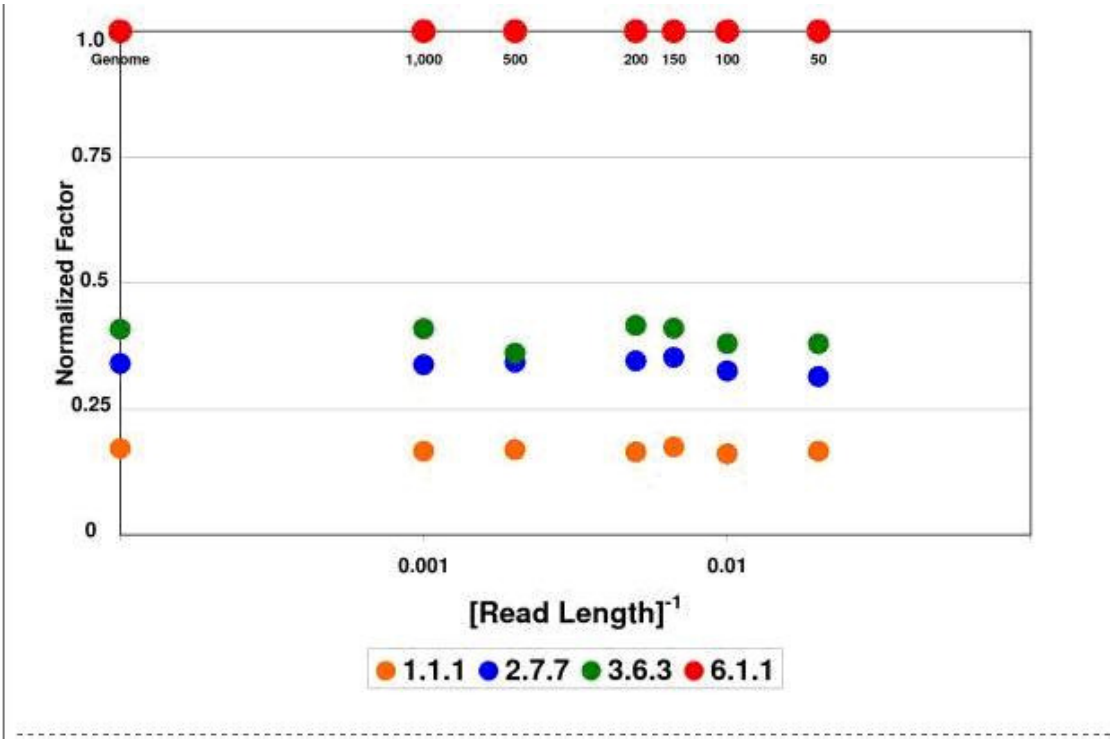


Figure 5.3 - Normalized factors.
Factors normalized to the 6.1.1. raw-factor for the same data as in Figure 5.2.

We will therefore normalize the raw factors by dividing them by the highest raw factor as follows (Figure 5.3): NF(EC)=RF(EC)/(RF(6.1.1)). The stability of the NFs will allow us to employ them in metagenomic studies of variable SR lengths.

5.4 The SPSR methodology: Training on 11 bacteria.

Next we use a set of 11 bacteria to serve as a training set, to provide factors that are suitable for metagenomic studies. The identities of the bacteria are displayed in Table 5.1 below, together with another set of 11 bacteria that will be used as a test set for the resulting factors. The bacteria were chosen from different phyla and classes to provide a balanced representation of the expected variance in metagenomic studies. Moreover, care was taken to choose species with well-studied genomes, having many EC-annotated enzymes. Proteomic information has been derived from Uniprot.

Organism Name	ID	Phyla	Total Uniprot Proteins	Total Uniprot Enzymes	Choice
Mycobacterium tuberculosis.	B01	Actinobacteria	5,971	1,371	Train
Mycobacterium bovis.	B02	Actinobacteria	3,986	1,253	Test
Sulfurihydrogenibium azorense	B03	Aquificae	1,708	486	Train
Aquifex aeolicus.	B04	Aquificae	1,556	368	Test
Cytophaga hutchinsonii	B05	Bacteroidetes	3,771	895	Test
Gramella forsetii	B06	Bacteroidetes	3,554	992	Train
Pelodictyon luteolum	B07	Chlorobi	2,078	496	Test
Chlorobium chlorochromatii	B08	Chlorobi	1,991	609	Train
Nostoc punctiforme	B09	Cyanobacteria	6,601	1,534	Train
Anabaena variabilis	B10	Cyanobacteria	5,643	1,362	Test
Synechocystis sp	B11	Cyanobacteria	3,529	575	Train
Bacillus cereus (strain ZK).	B12	Firmicutes	5,638	1,469	Test
Bacillus cereus (strain ATCC).	B13	Firmicutes	5,248	1,546	Train
Pseudomonas aeruginosa.	B14	Proteobacteria	9,091	848	Train
Rhizobium meliloti	B15	Proteobacteria	7,107	1,583	Test
Salmonella typhimurium.	B16	Proteobacteria	5,768	1,279	Train
Shigella flexneri.	B17	Proteobacteria	5,395	813	Test
Salmonella typhi.	B18	Proteobacteria	5,351	942	Test
Escherichia coli (K12).	B19	Proteobacteria	4,412	1,443	Train
Caulobacter crescentus	B20	Proteobacteria	3,852	1,238	Test
Leptospira biflexa	B21	Spirochaetes	3,730	957	Train
Thermotoga petrophila	B22	Thermotogae	1,784	411	Test

Table 5.1: Bacterial genomes used for training and testing the SPSR methodology.

Each genome on this list has been randomly divided into reads of length 50, with 5 fold coverage of each genome, and submitted to SP analysis. To gather statistics we have analyzed 15 combinations of 7 out of the 11 organisms of the training set. Each such set of 7 organisms served to define a super-organism (or artificial metagenome) with given annotated enzymes and SP counts. The resulting numbers of SP hits were then compared with the known numbers of enzyme-genes, leading to the desired factors for each EC category. Normalized factors of leading categories are presented in Table 5.2 below.

EC	Normalized factor	Standard Deviation
1.1.1	0.15	0.014
1.2.1	0.28	0.024
2.1.1	0.22	0.030
2.3.1	0.11	0.023
2.4.1	0.16	0.028
2.4.2	0.26	0.011
2.5.1	0.25	0.011
2.6.1	0.17	0.010
2.7.13	0.03	0.003
2.7.7	0.45	0.022
3.1.1	0.08	0.022
3.1.3	0.07	0.010
3.2.1	0.07	0.017
3.5.1	0.15	0.034
3.6.1	0.89	0.026
3.6.3	0.45	0.064
4.1.1	0.25	0.019
4.2.1	0.30	0.024
6.1.1	1.00	0.000

Table 5.2: Factors normalized to the 6.1.1 raw factor derived from an analysis of SRs with length of l=50 nucleotides belonging to15 combinations of 7 out of the 11 organisms of the training set listed in Table 5.1.

Technical details

We utilize the Knuth Morris Pratt algorithm to perform the search of SPs of length m amino-acids on the six-mode translations of short reads of length n bases. This leads to temporal complexity of order O(m+2n). Our system runs on a four-processors Intel(R) Xeon(R) CPU 2.33GHz Linux machine and performs a search of the full SP list on approximately 50,000 nucleotides per hour.
We provide an online web tool that processes short read files provided by users. The system can be accessed at http://horn.tau.ac.il/SPSR.

5.5 Taxon Specific Peptides

The SP methodology can be further developed to characterize taxon-specific SPs, to be denoted as TSPs. This is of interest for pervasive EC categories, some of which we will encounter in our metagenomic analysis. The idea is then, for a particular EC category (6.1.1, aminoacyl tRNA synthetases, aaRS) to filter the SPs according to whether they are specific to a given domain, given phylum or class. The training data

on the quoted EC categories are rich enough to allow separation into Archaea, Eukarya and Bacteria, and further specification of bacteria into Proteobacateria, Firmicutes, Cyanobacteria and Actinobacteria. The phylum Proteobacteria, being the largest in the data, allows for further filtering into alpha-, beta- and gammaproteobacteria.
We further concentrate on those aaRS EC numbers that are known to have a single protein per species. An analysis of all bacterial aaRS in Swiss-Prot leads to the statistics displayed in Table 5.3 below.

EC	# doublets	# triplets	# Proteins	% multiples	S61
6.1.1.1	18	0	474	3.80	
6.1.1.2	3	0	125	2.40	
6.1.1.3	1	0	616	0.16	x
6.1.1.4	2	0	703	0.28	x
6.1.1.5	10	0	524	1.91	x
6.1.1.6	60	3	527	12.52	
6.1.1.7	1	0	628	0.16	x
6.1.1.9	0	0	293	0.00	x
6.1.1.10	2	0	421	0.48	x
6.1.1.11	4	0	735	0.54	x
6.1.1.12	2	0	688	0.29	x
6.1.1.13	68	1	172	40.70	
6.1.1.14	276	0	825	33.45	
6.1.1.15	10	0	762	1.31	x
6.1.1.16	14	0	691	2.03	x
6.1.1.17	114	0	808	14.11	
6.1.1.18	0	0	139	0.00	x
6.1.1.19	4	0	675	0.59	x
6.1.1.20	251	0	877	28.62	
6.1.1.21	6	0	627	0.96	x
6.1.1.22	1	0	256	0.39	x

Table 5.3: Statistics of bacterial aaRS enzymes in Swiss-Prot data.
The column '%multiples' refers to the percentage of species that display multiple proteins with the same EC number. The sub-set S61, defined by x entries in the last column, is selected for taxonomic classification.

Confining ourselves to aaRS that have up to 2% multiple entries, we select the subgroup to be denoted S61 (single proteins in the 6.1.1 EC category), indicated on Table 5.3. It is this S61 set that we will employ for taxon classification. Eliminating aaRS categories with many multiples helps in reducing the margin of error in our predictions.

TSPs are selected for their phylum-level and class-level specificity, after scrutinizing the enzyme data-set of Swiss-Prot. We make use of the same data-set to determine the raw-factors that may be associated with the various TSPs. These would theoretically correspond to very large reads, and only their ratios should be trusted for short reads. Table 5.4 below represents the factors for the S61 subset of the EC category of 6.1.1.

	Taxon	# enzymes	# TSPs	# hits	factor
	Archaea	543	408	1807	3.33
	Eukaryota	259	150	260	1.00
	Bacteria	7752	8310	98556	12.71

Bacteria	Proteobacteria	4341	3768	34376	7.92
Bacteria	Firmicutes	1561	1130	7457	4.78
Bacteria	Cyanobacteria	328	175	541	1.65
Bacteria	Actinobacteria	494	392	1874	3.79
Bacteria	Tenericutes	193	25	72	0.37
Bacteria	Bacteroidetes	132	103	223	1.69
Bacteria	Spirochaetes	185	71	173	0.94
Bacteria	Thermotogae	81	9	22	0.27
Bacteria	Chlamydiae	114	140	383	3.36
Bacteria	Chlorobi	90	31	79	0.88
Archaea	Crenarchaeota	165	53	158	0.96
Archaea	Euryarchaeota	359	281	932	2.60

Proteobacteria	Gammaproteobacteria	2372	1624	13622	5.74
Proteobacteria	Alphaproteobacteria	870	675	3806	4.37
Proteobacteria	Betaproteobacteria	638	394	1950	3.06
Proteobacteria	Epsilonproteobacteria	223	178	430	1.93
Proteobacteria	Deltaproteobacteria	229	9	19	0.08
Firmicutes	Bacillales	614	327	1811	2.95
Firmicutes	Clostridia	374	142	567	1.52
Firmicutes	Lactobacillales	573	387	2543	4.44
Cyanobacteria	Chroococcales	114	64	184	1.61
Bacterioidetes	Bacteroidia	85	67	138	1.62

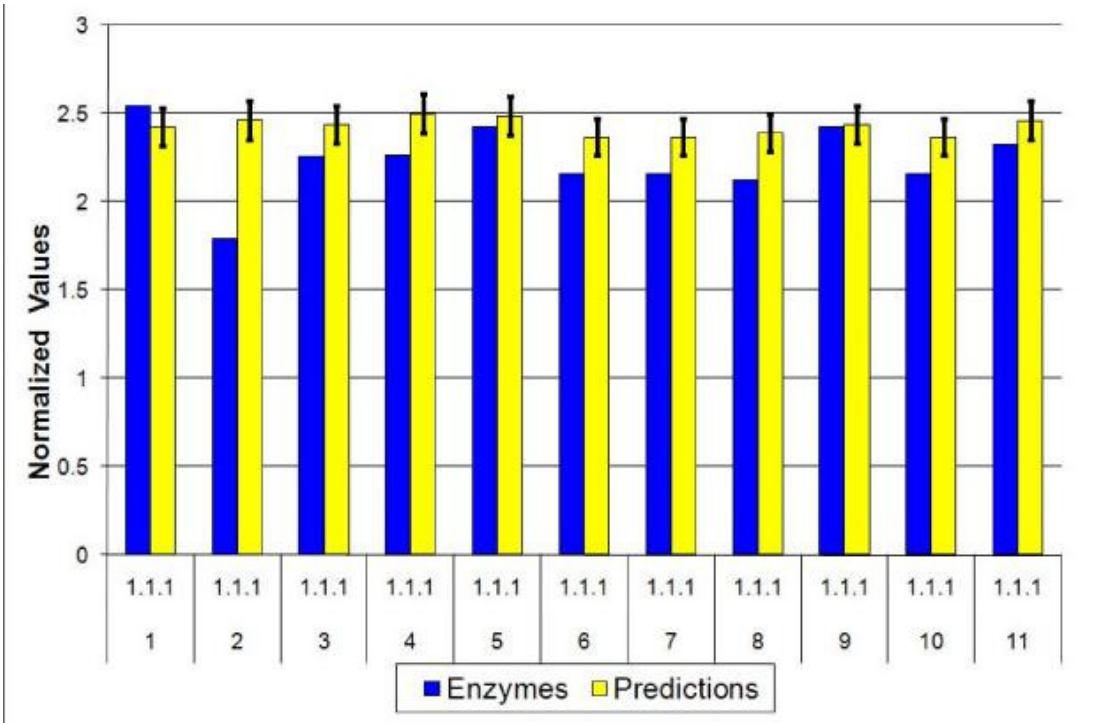
Table 5.4. Raw factors of TSPs corresponding to the S61 subset of EC category 6.1.1, as derived from TSP hits on proteomes in the Enzyme Swiss-Prot data-base.
Shown are all taxa that have more than 80 aaRS enzymes listed in this data-base, numbers of TSP associated with the S61 set corresponding to the relevant taxa, their hits and the deduced raw factors. For bacteria and archaea, predicted numbers of enzymes are assumed to be proportional to numbers of cells present in the sample. In eukaryotes one should apply a further reduction by 1.5, the average number of aaRS enzymes per cell known to be detected in Uniprot data.

We provide an online web tool that processes short read files queried by users, leading to a prediction of relative taxonomic mixtures of the presented data. The system can be accessed at http://horn.tau.ac.il/S61TSPSR.

Results: Analysis of the Methodology

5.6 Test of the SPSR methodology.

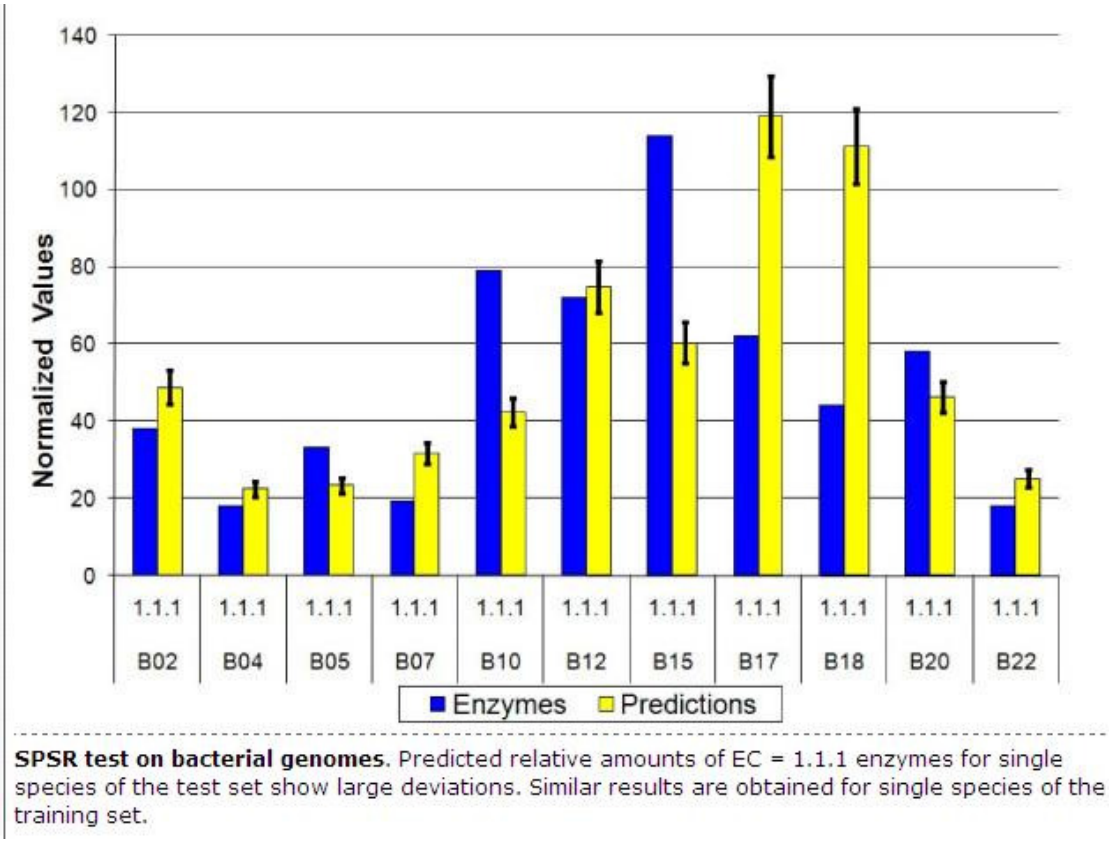
In the present section we test the factors derived from the artificial metagenomes (the super-organisms consisting of 7 out of the 11 training set organisms) on the test-set organisms listed in Table 5.1. Using the errors (standard-deviations) determined by the training procedure, we quote the quality of fits by using the chi-square test, which is expected to be of the order of the number of degrees of freedom, $E[(X-\mu)^2/\sigma^2]=N$ (where E is the expectation value, X is the variable whose average is μ and standard-deviation is σ, and N is the number of degrees of freedom). Overall, when the factors are applied to novel artificial 7 species metagenomes, the generalization errors are about the same as expected from the training set errors (Figure 5.4 below), with $E \sim 1.5\ N$.



SPSR test on artificial metagenomes. SPSR tests, based on normalized factors derived from the training set. Shown are predicted and known numbers of enzymatic genes relative to predicted and known aaRS enzymes (EC = 6.1.1). Error-bars reflect the standard deviations of the factors derived from 15 trials of the artificial super-organisms. Shown here is a comparison of predicted relative amounts of EC = 1.1.1 enzymes for 11 artificial metagenomes containing 7 species from the test set. EC = 1.1.1 is a leading EC category containing alcohol dehydrogenases with NAD^+ or $NADP^+$ as acceptor. The results imply a good generalization for metagenomes composed of 7 bacteria.

Figure 5.4 - SPSR test on artificial metagenomes.

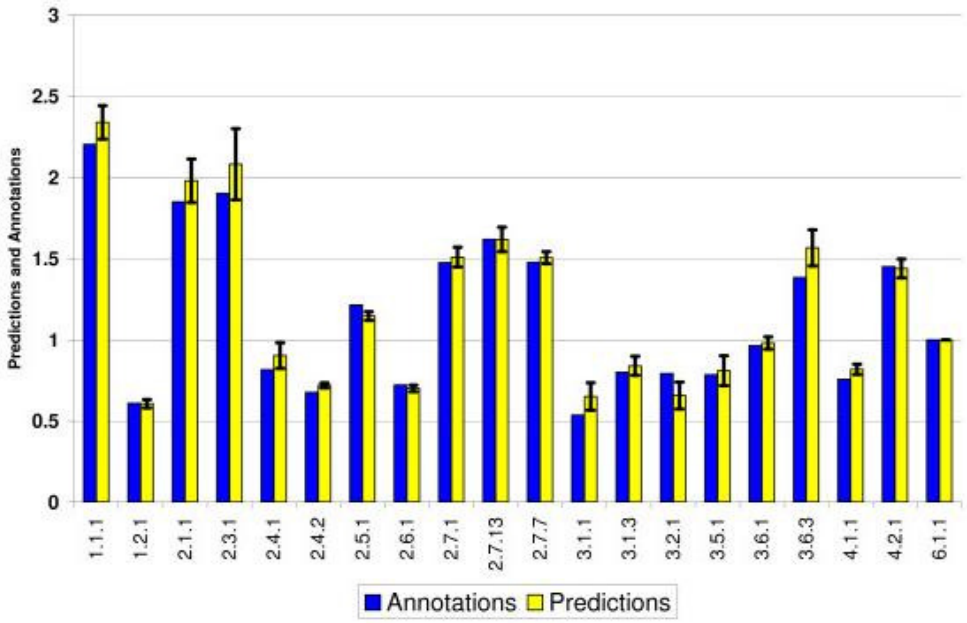
However, when the same factors are applied to single species predictions (Figure 5.5 below) the deviations are much larger.



SPSR test on bacterial genomes. Predicted relative amounts of EC = 1.1.1 enzymes for single species of the test set show large deviations. Similar results are obtained for single species of the training set.

Figure 5.5 - SPSR test on bacterial genomes.

The chi-squared test leads to E ~ 27N. Somewhat better fits are obtained for raw predictions, with E ~ 8N. The poor chi-square values reflect the fact that metagenomic averages smooth-out differences between single organisms. A similar behaviour is observed also for single species from the training set. Another aspect of the same effect is seen when larger metagenomes are considered, e.g. one composed of all 22 species, with predictions that are better than the training-set errors shown in Figure 5.6, where E ~ 0.4 N.



SPSR test on a metagenome of all 22 bacteria. Predictions of relative amounts of enzymes belonging to many EC categories, for a metagenome of all 22 bacteria, demonstrate a better fit than the training errors.

Figure 5.6 - SPSR test on a metagenome of all 22 bacteria.

5.7 Test of the TSPSR method at the phylum level.

We have applied the S61 TSPs to the 22 bacteria of Table 5.1. In each case we have calculated the TP (true-positive) signals (i.e. predicted numbers of enzymes associated with the correct phylum) and the FP (false-positive) ones. The results shown in Table 5.5 validate this methodology.

Organism	Phylum	Precision
Mycobacterium tuberculosis.	Actinobacteria	96%
Mycobacterium bovis.	Actinobacteria	96%
Sulfurihydrogenibium azorense	Aquificae	no prediction
Aquifex aeolicus.	Aquificae	no prediction
Cytophaga hutchinsonii	Bacteroidetes	74%
Gramella forsetii	Bacteroidetes	74%
Pelodictyon luteolum	Chlorobi	91%
Chlorobium chlorochromatii	Chlorobi	95%
Nostoc punctiforme	Cyanobacteria	81%
Anabaena variabilis	Cyanobacteria	89%
Synechocystis sp.	Cyanobacteria	96%
Bacillus cereus (strain ZK).	Firmicutes	94%
Bacillus cereus (strain ATCC 14579).	Firmicutes	95%
Pseudomonas aeruginosa.	Proteobacteria	94%
Rhizobium meliloti	Proteobacteria	97%
Salmonella typhimurium.	Proteobacteria	99%
Shigella flexneri.	Proteobacteria	100%

Salmonella typhi.	Proteobacteria	100%
Escherichia coli (K12).	Proteobacteria	99%
Caulobacter crescentus	Proteobacteria	88%
Leptospira biflexa serovar Patoc	Spirochaetes	72%
Thermotoga petrophila	Thermotogae	91%

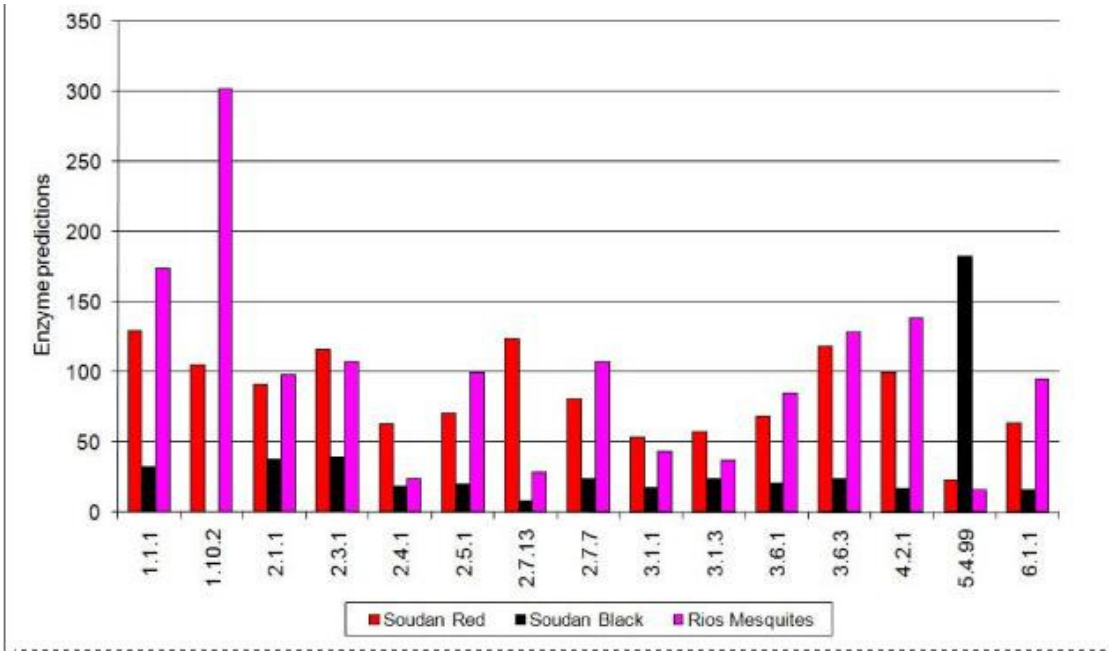
Table 5.5: Phylum predictions according to S61 TSPs for the 22 species of Table 5.1.

Although some of the data has been included in the training procedure, it should be emphasized that whereas training (i.e. assignment of TSPs) was carried out on all Swiss-Prot enzymes, the calculations of Table 5.5 are carried out on the full genomes of the 22 organisms, i.e. the procedure includes processing all genic and intergenic regions of these organisms. Precision is defined as TP/(TP+FP). No assignment has been made if the number of predicted enzymes, on the basis of TSPs, was less than 1. This was the case for the two species of Aquificae, for which we have no corresponding TSPs.

5.8 Results: Environmental Metagenomic Analysis

Enzymatic signatures of several metagenomes.

Figure 5.7 below displays our analysis of 3 metagenomes taken from Dinsdale et al



[5.3].

Figure 5.7 - Enzymatic annotations predicted for three metagenomes.
Predictions of number of enzymes in many EC categories for 3 metagenomes [Dinsdale et al. , 2008]

All of them comprise short reads, with average lengths of around 100 nucleotides. SPSR predictions are represented in absolute terms, i.e. predicted numbers of enzymes, using the l=100 raw factors, to exhibit common and different trends among these three examples.

The Rios Mesquites Stromatolites bacteria (to be denoted Rios Mesquitos henceforth) and the Soudan Mine Red biofilm data (to be denoted Soudan Red) have more than 60 predicted proteins in the EC category 6.1.1., while the Soudan Mine Black biofilm (Soudan Black) has only about 20 such proteins. Thus one would conclude that the total coverage content of the first two metagenomes is of the order of three cells or more, while that of the Soudan Black should be only of the order of one cell. In the two large metagenomes we find large contributions of EC category 1.1.1 (alcohol dehydrogenases with NAD^+ or $NADP^+$ as acceptor) and EC category 3.6.3 (hydrolases catalysing transmembrane movement of substances). The Rios Mesquitos has the strongest signal in EC 1.10.2, oxidoreductases acting on diphenols with cytochrome as acceptor.
It is clear from Figure 5.7 above that the Soudan Black metagenome is very different from the two others. In particular, it has a very strong signal for 5.4.99., intramolecular transferases. Follow up analysis indicates that this signal is due to 426 SRs that carry the EC 5.4.99.2 (methylmalonyl-CoA mutase) signature. Their identification is due to two SPs with this assignment, NSISISGYH occurring 276 times, and ISISGYHMQEAG occurring 185 times in these data. As these peptides partly overlap, there exist many short reads on which the two occur together. These results stand out for several reasons: their numerous counts outnumber all other enzyme classes by more than an order of magnitude; no other SP of the same EC category is observed in the data; extending these SRs by other partially overlapping short reads does not lead to considerably larger putative proteins. These lines of evidence hint that the Soudan Black data-set should be reexamined, as some artifact has likely been introduced at some point. While such an examination is outside the scope of this paper, we wish to emphasize that the SPSR methodology quickly highlights such anomalies and can therefore serve, among other purposes, also as a rapid quality assessment tool for metagenomic data.

5.9 Taxonomic analysis of metagenomes using TSPs.

Taxonomic analysis of the three metagenomes analyzed above has been carried out using S61 TSPs. In all of the metagenomes examined we conclude that Bacteria are the dominant kingdom (with small traces of Archaea in the Soudan mine data). Both Soudan Red and Rios Mesquitos show that, among Bacteria, there is an order of magnitude difference in the quantities of Proteobacteria *vs* Firmicutes. Soudan Black data have the same order of magnitude for both, but given the artifact we have noted, this estimate should be taken with a grain of salt. Predictions for classes of Proteobacteira in Soudan Red are shown in Table 5.6 below, where they are compared with the results of the 16S rRNA-based analysis of Edwards et al [5.14] and with a CARMA analysis [5.8] of the same data.

Class	Edwards	CARMA	S61TSP
Alphaproteobacteria	40%	37%	45%
Gammaproteobacteria	54%	40%	45%
Betaproteobacteria	2%	8%	8%
Epsilonproteobacteria	0%	2%	2%
Deltaproteobacteria	3%	13%	0%

Table 5.6: Comparison of class predictions within proteobacteria for the Soudan Red data.

The Edwards results were estimated from Figure 1 of their paper, and the CARMA analysis was carried out by us using their website. There is an overall agreement regarding relative abundance of alpha- and gamma-proteobacteria, but the details of the minor classes differ among the different methods. This may be because the three methods rely on three different aspects of the data.

A fourth metagenomic data-set to which we have applied our taxonomic analysis is that of DeLong et al. [5.15] who have studied metagenomes in the ocean at different depths, thus obtaining stratified microbial assemblages. The latter have been analyzed according to taxonomic groups, functional gene repertoires and metabolic potential. Their data were assembled into contigs of average length of 1000 nucleotides, and their taxonomic analysis has been carried out by comparing cumulative TBLASTX high-scoring sequence pairs bit scores of each depth against one another. The different depths were grouped into Photic Zone (10m, 70m and 130m) and Deep Water zone (500m, 770m and 4000m). Analysis of these data using S61 TSPs leads to the results displayed in Table 5.7 below.

Kingdom	Photic Zone	Deep Water
Archaea	3	3
Eukaryota	1	0
Bacteria	22	32

Phylum	Photic Zone	Deep Water
Proteobacteria	9	12
Firmicutes	1	2
Cyanobacteria	8	2
Actinobacteria	2	2

Table 5.7: Taxonomic predictions of DeLong data based on S61 TSPs.
Numbers signify expected numbers of S61 aaRS enzymes. The latter are proportional to the numbers of cells of the different taxa in the data.

Numbers shown are predicted numbers of enzymes in the data. Obviously the quantity of the data amounts to just a few cells in total of all depths. Data are dominated by bacteria although there are some traces of archaea and eukaryotes (with decrease of the latter in deep water). Among bacteria we find a relatively large abundance of Cyanobacteria at low depths (mostly 10m and 70m). Proteobacteria, whose fraction in the community is relatively stable as function of depth, may be further analyzed for

their breakdown into classes. We find the ratio of Alphaproteobacteria: Gammaproteobacteria: Betaproteobacteria to be 4:3:1 in the photic zone, and 3:4:1 in deep water, i.e. roughly stable with depth (not shown in Table 5.7).

DeLong et al. [5.15] have constructed large contigs (average length 1000 nucleotides) that can provide much more specific taxonomic information than our EC 6.1.1 based analysis. Nonetheless the latter is consistent with theirs. The advantage of the TSP analysis is that it allows one to obtain a rough taxonomic breakdown of the microbial community when short reads are the sole source of information.
It should be noted that the raw factors of the TSPs were determined by Swiss-Prot data. Since the latter may be richer in SP hits than yet unassigned proteins that are identified by our methodology, the absolute values quoted in Table 5.7, being based on these raw factors, should be regarded as lower bound estimates of the true taxonomic distribution.

Conclusions

The use of SPs allows deriving enzymatic information directly from short reads of genomic and metagenomic data. This is of great importance in view of the large amount of data-analysis performed with short read methods. It is of particular importance in metagenomic studies, where the organismal composition of the studied data is usually unknown and contig assembly is often impossible. Thus one may functionally study high complexity ecosystems, such as soil and seawater, overcoming the barrier of genome reconstruction, by deriving enzymatic signatures in a straightforward manner.

The enzymatic signatures obtained may serve for coarse grain functional characterization of the environment. Lapierre and Gogarten [5.16] have pointed out that "character genes" typical to taxonomic groups, such as methanogen-specific enzymes, may also inform us of the composition of the microbiome. We have shown that the use of TSPs for aaRSs, can serve as the basis for taxonomic analysis. Our SP signatures can also serve as indicators for novel functionalities and, in extreme cases, as indicators for the possible contamination of the data-set that is being analyzed.

We provide a webtool at http://horn.tau.ac.il/SPSR that analyzes sets of short reads, extracting all those that have SP hits, together with the indication of their EC categorization. These lists can be further processed, by the tools explained above, to provide enzymatic spectra, or to search for consistency of the analyzed data.

The aaRS super-family plays a special role in our analysis because of several reasons. The first is the large over-all similarity of aaRS enzymes throughout all kingdoms of life, leading to extraordinarily large numbers of 6.1.1 SPs derived from the training-set. This is reflected by the large factors of the 6.1.1 category in Figures 5.2 and 5.3. The second reason is their usefulness in discriminating among species, by providing a large number of TSPs. Finally, the fact that for many of the aaRS enzyme types there exists one corresponding protein on each bacterial genome, allows using this super-family as a suitable calibrating device.

Our use of the aaRS SPs as taxonomic measures can be compared to the phylogenetic classification based on Environmental Gene Tags (EGTs) introduced by Krause et al [8] in their CARMA tool. Their method is based on selecting DNA fragments of lengths of order 100 bases, i.e. short reads, and comparing them to Pfam profile HMMs. The identified short reads are defined as EGTs. Incorporating them into phylogenetic trees, the authors developed an algorithm that provides a taxonomic distribution with relatively high accuracy. The similarity between the two tools is that both depend on protein-markers rather than on 16S rRNA ones, which is the gold standard of prokaryotic taxonomy.

There are however many differences. First, they employ Pfam domains over many protein families, whereas we concentrate on SPs of aaRS enzymes only. This guarantees that their tool is more powerful, in the sense that its larger statistics allows for extension to lower taxonomic levels than ours. Second, their methodology relies on employing a battery of tools of the trade, such as BLASTX for sequence matching, pHMM for the Pfam generated ETGs, and PHYLIP for clustering phylogenetic trees. This is commonly regarded necessary, in order to take into account all the generated know-how in bioinformatics. We, on the other hand, rely on a simple look-up table of SPs that has been generated from the enzymes that exist on Swiss-Prot. Its advantage is its simplicity. Third, both methods suffer from biases, since their tools are constructed on existing labeled data. CARMA provides its final results by counting the number of EGTs correlated with each taxon. The analog in our case would have been to count the TSPs. Because of the simplicity of our approach we are aware of one explicit bias: TSP hits differ among taxa because of differences in the sizes of TSP pools. We are able to address this bias by correcting the numbers of TSP counts through the use of raw-factors, providing expected numbers of proteins that should be proportional to numbers of cells. Thus, without diminishing the value of tools like CARMA, we believe that our tool has some clear advantages, and should be used as an additional source of information.

We provide a taxon-search webtool at http://horn.tau.ac.il/S61TSPSR. Upon submission of a list of short reads, it extracts taxonomic distributions at levels of kingdoms, bacterial phyla, and bacterial classes.

References

5.1. Tringe, S.G., Von Mering, C., Kobayashi, A. , Salamov, A.A., Chen K., Chang H.W., Podar M., Short J.M., Mathur E.J., Detter J.C., Bork P., Hugenholtz P., Rubin E.M. Comparative metagenomics of microbial communities. Science 2005: 308:554-557.

5.2. Wooley J C, Godzik A and Friedberg I. A Primer on Metagenomics. PLOS Comp. Bio. 2010: 6 (2), e1000667.

5.3. Dinsdale, E. A., Edwards R.A., Hall D., Angly F., Breitbart M., Brulc J.M., Furlan M., Desnues C., Haynes M., Li L., McDaniel L., Moran M. A., Nelson K.E., Nilsson C., Olson R., Paul J., Brito B.R., Ruan Y., Swan B. K., Stevens R., Valentine D. L., Thurber R. V., Wegley L., White B. A., Rohwer F. Functional metagenomic profiling of nine biomes. Nature 2008: 452, 629-632.

5.4. Riesenfeld, C.S., P.D. Schloss, and J. Handelsman, Metagenomics: genomic analysis of microbial communities. Annu Rev Genet, 2004. 38: p. 525-52.

5.5. Raes, J and Bork, P. Molecular eco-systems biology: towards an understanding of community function. Nature Reviews Microbiology 2008. 6: 693-699

5.6. Cole, J.R., Wang Q., Cardenas E., Fish J., Chai B., Farris R. J., Kulam-Syed-Mohideen A. S., McGarrell D. M., Marsh T., Garrity G. M., and Tiedje J. M. The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. Nucleic Acids Res. 2009: 37(Database issue), D141–D145.

5.7. Schloss P. D. and Handelsman J.. Introducing SONS, a Tool for Operational Taxonomic Unit-Based Comparisons of Microbial Community Memberships and Structures. App. Env. Microb. 2006. 72: 6773-6779.

5.8. Krause L., Diaz N. N., Goesmann A, Kelley S.,Nattkemper T. W., Rohwer F., Edwards R. A., Stoye,J. Phylogenetic classification of short environmental DNA fragments. Nucleic Acids Research, 2008. 36: 2230–2239

5.9. Kunik, V., Meroz, Y., Solan, Z., Sandbank, B., Weingart U., Ruppin E., Horn D. Functional representation of enzymes by specific peptides. PLOS Comp. Biol. 2007:3(8), e167

5.10. Bork, P. and Koonin, E.V. Protein sequence motifs. Curr. Op. Structural Biology 1996. 6, 366-376.

5.11. Bairoch A, Bucher P, Hofmann K. Prosite. Nuc. Acids Res. 1997: 25, 217-221.

5.12. Solan Z., Horn D., Ruppin E, Edelman S. Unsupervised learning of natural languages. Proc. Natl. Acad. Sci. USA 2005: 102, 11629-11634.

5.13. Weingart U, Lavi Y and Horn D. Data Mining of Enzymes using Specific Peptides. BMC Bioinfomratics 2009: 10, 446

5.14. Edwards R. A., Rodrigues-Brito B., Wegley L., Haynes M., Breitbart M., Peterson D., Saar.M., Alexander.S., Alexander.E.C., Rohwer,F.. Using pyrosequencing to shed light on deep mine microbial ecology. BMC Genomics, 2006. 7: 57-70.

5.15. DeLong E. F., Preston C. M., Mincer T., Rich V., Hallam S. J., Frigaard N-U., Martinez A., Sullivan M. B., Edwards R., Brito B. R., Chisholm S. W., Karl D. M. Community Genomics Among Stratified Microbial Assemblages in the Ocean's Interior. Science 2006: 311, 496 – 503.

5.16. Lapierre P. and Gogarten J.P. Estimating the size of the bacterial pangenome. Trends in Genetics 2009: doi:10.1016/j.tig.2008.12.004

6. Summary

The work presented here expands previous studies of SPs into many different venues of the Bioinformatic research: analysis of single proteins, metaproteomes, single bacterial and archaea genomes, metagenomes and analysis of short reads.
We presented the construction and structure of enzymatic SPs sets, selection of training sets, pros and cons of utilization of supervised and unsupervised methods and optimization of the SP datasets. In addition to the production SP set we constructed specialized subsets:
ASPs: Annotated SPs: Used to predict Active, Metal and Binding properties
GSPs: Gene ontology SPs used to predict Gene Ontologies
FSPs: Family SPs used to predict gene names
TSPs: Taxon SPs used to predict taxonomic lineage

We have developed the methodology of employing SPs for data mining of enzymes (DME). In particular we have shown in Chapter 3 that the requirement that SP occurrences on protein sequences has some minimal coverage length, e.g. $L \geq 7$ amino-acids in our analyses, leads to the novel tool of DME. It is applicable to large genomic and metagenomic data, and provides a good indicator for the enzymatic classification of the queried proteins, based on a look-up table only. We have successfully applied it to Sargasso Sea Data and the Human Gut Metagenome, presenting our results as enzymatic profiles of these data. A web tool identifying SP (and ASP) occurrences on any queried protein sequence, and providing the EC prediction of DME, has been made available online at http://adios.tau.ac.il/DME.

Using the same spirit we have developed the SP scaffolding method, applicable to full genomes. This is based on pinpointing SP hits on the genomic strands, and concluding from them where enzymatic genes occur and what their EC assignments are. The capability of this method has been demonstrated on the full genome of H. Pylori 26995. In Chapter 4 we have used for this application our FSP sets, providing not only EC classification of enzymatic genes but also the gene names of the families to which they belong.

We proceeded to describe the expansion of the DME methodology to derive enzymatic and taxonomic information directly from short-reads of genomic and metagenomic data. This is of great importance in metagenomic studies, where the organismal composition of the studied data is usually unknown and contig assembly is often impossible. Utilizing our methodology we can overcome the barrier of genome reconstruction, by deriving enzymatic signatures and taxonomic information in a straightforward manner. We provide a web tool at http://horn.tau.ac.il/SPSR that analyzes sets of short reads, extracting all those that have SP hits, together with the indication of their EC categorization. These lists can be further processed, by the tools explained in Chapter 5, to provide enzymatic spectra, or to search for consistency of the analyzed data.

We have extended the short read analysis to provide also taxonomic signatures of metagenomic data. This has been done by employing SPs belonging to aaRS enzymes. Applying it to metagenomic results we came up with predictions that were consistent with other authors who employed different methodologies. We provide a

taxon-search web tool at http://horn.tau.ac.il/S61TSPSR. Upon submission of a list of short reads, it extracts taxonomic distributions at levels of kingdoms, bacterial phyla, and bacterial classes.

An important and exciting direction for future expansion of this work will be generation of Protein Family SPs: Only one fifth of proteins are enzymes, therefore the scope of predictivity power of Protein Family SPs will be far greater than that of enzymatic SPs. Construction of Protein Family SPs will be conducted selecting groups of families of proteins, running MEX against each group and implementing specificity criteria, the final result being a set of Protein Family SPs. A major contribution of this work is laying the methodological and software foundation for this future expansion.

SPs possess excellent predictivity power, accuracy, versatility and ease of use for large volumes of Bioinformatic data, and are therefore a very powerful instrument in the Bioinformatics toolkit.

7. Web tools

In parallel of batch processing for large volume DME predictions, we have developed a number of web applications.

7.1: Data Mining of Enzymes - Peptide Search

http://adios.tau.ac.il/DME/

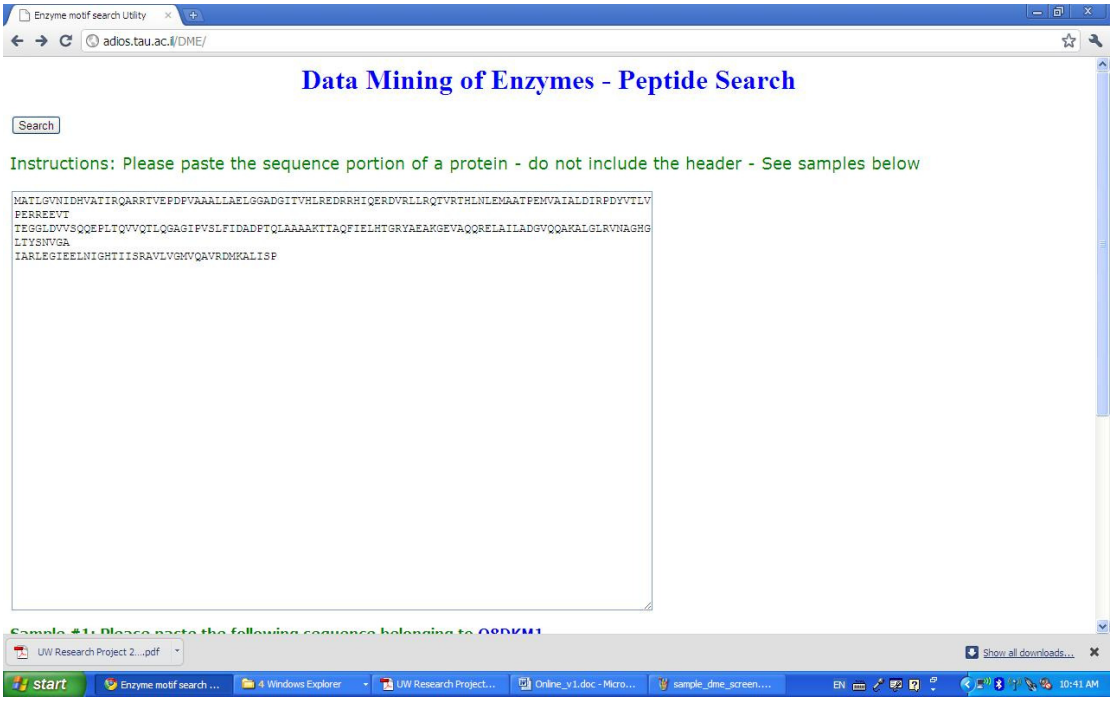


Figure 7.1: http://adios.tau.ac.il/DME/ - Input screen

Processing consists of search of each one of the SPs belonging to the Production SP set V2.3. Hits are analyzed and an enzymatic prediction for the sequence is generated as shown in figure 7. 2 below. The application shows metal, binding and active sites hit by the Annotated SPs. Default L3 threshold for a prediction is 7 amino-acids. Analysis is real-time and typical response time is sub-second for sequences of two to three hundred amino-acids.
As an example we display the analysis of Swissprot enzyme Q8DKM1 (PDXJ_THEEB), for Thermosynechococcus elongatus (strain BP-1). This enzyme is annotated in Swissprot with an EC=2.6.99.2 (Pyridoxine 5'-phosphate synthase).

Figure 7.2: Results of online inquiry for enzyme Q8DKM1 in http://adios.tau.ac.il/DME/

Data Mining for Enzymes Search Utility

Active, Metal and Binding Site Annotations based on Training Swissprot Dataset

Specific Peptide	EC	Function	Location of SP in Protein	Act Site	Act Site Desc	Metal Site	Metal Site Desc	Binding Site	Binding Site Desc
LGVNIDH	2.6.99.2;	Pyridoxine 5'-phosphate synthase.	4	-	-	-	-	4	3-amino-2-oxopropyl phosphate.
TVEPDPV	2.6.99.2;	Pyridoxine 5'-phosphate synthase.	20	-	-	-	-	-	-
EPDPVAAA	2.6.99.2;	Pyridoxine 5'-phosphate synthase.	22	-	-	-	-	-	-
HLREDRRH	2.6.99.2;	Pyridoxine 5'-phosphate synthase.	43	1	Proton acceptor (By similarity	-	-	3	1-deoxy-D-xylulose 5-phosphate
LLRQTVR	2.6.99.2;	Pyridoxine 5'-phosphate synthase.	58	-	-	-	-	-	-
TVRTHLNLEMA	2.6.99.2;	Pyridoxine 5'-phosphate synthase.	62	9	Proton acceptor (By similarity	-	-	-	-
NLEMAAT	2.6.99.2;	Pyridoxine 5'-phosphate synthase.	68	3	Proton acceptor (By similarity	-	-	-	-
PDYVTLVPE	2.6.99.2;	Pyridoxine 5'-phosphate synthase.	86	-	-	-	-	-	-
VPERREE	2.6.99.2;	Pyridoxine 5'-phosphate synthase.	92	-	-	-	-	-	-
EVTTEGG	2.6.99.2;	Pyridoxine 5'-phosphate synthase.	98	-	-	-	-	3	1-deoxy-D-xylulose 5-phosphate
TTEGGLD	2.6.99.2;	Pyridoxine 5'-phosphate synthase.	100	-	-	-	-	1	1-deoxy-D-xylulose 5-phosphate
VSLFIDA	2.6.99.2;	Pyridoxine 5'-phosphate synthase.	128	-	-	-	-	-	-
SLFIDAD	2.6.99.2;	Pyridoxine 5'-phosphate synthase.	129	-	-	-	-	-	-
FIELHTG	2.6.99.2;	Pyridoxine 5'-phosphate synthase.	149	-	-	-	-	-	-

ELHTGRYA	2.6.99.2;	Pyridoxine 5'-phosphate synthase.	151	-	-	-	-	-	-
VNAGHGL	2.6.99.2;	Pyridoxine 5'-phosphate synthase.	187	5	Proton donor (By similarity).	-	-	6	3-amino-2-oxopropyl phosphate;
AGHGLTY	2.6.99.2;	Pyridoxine 5'-phosphate synthase.	189	3	Proton donor (By similarity).	-	-	4	3-amino-2-oxopropyl phosphate;
EELNIGHTI	2.6.99.2;	Pyridoxine 5'-phosphate synthase.	208	-	-	-	-	-	-
AVRDMKAL	2.6.99.2;	Pyridoxine 5'-phosphate synthase.	228	-	-	-	-	-	-

Mapping of the SPs in the Protein

Red characters denote the location of the Specific Peptide Matches

```
---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+
MATLGVNIDHVATIRQARRTVEPDPVAAALLAELGGADGITVHLREDRRHIQERDVRLLRQTVRTHLNLEMAATPEMVAIALDIRPDYVTLVPERREEVT
TEGGLDVVSQQEPLTQVVQTLQGAGIPVSLFIDADPTQLAAAAKTTAQFIELHTGRYAEAKGEVAQQRELAILADGVQQAKALGLRVNAGHGLTYSNVGA
IARLEGIEELNIGHTIISRAVLVGMVQAVRDMKALISP
```

DME EC Prediction for this protein is: 2.6.99.2

Check another protein

Figure 7.2: Results of online sample analysis of amino-acid sequence

7.2: Derivation of Enzymatic Signatures from Short Read Data

http://adios.tau.ac.il/SPSR.html

This application accepts as input a file containing short reads and generates as output the enzymatic signatures of the short reads provided.

Figure 7.3: Derivation of enzymatic signatures from short reads data: Input screen.

Processing flow: Processing is on-line, real time. The file provided by the user is uploaded onto the server and each short read is translated into the six possible translation frames generating six records of pseudo peptides for every short read. All SPs from SP set V2.3 are searched within each and one of the pseudo peptides. Results of the SP hits are displayed in an output screen.

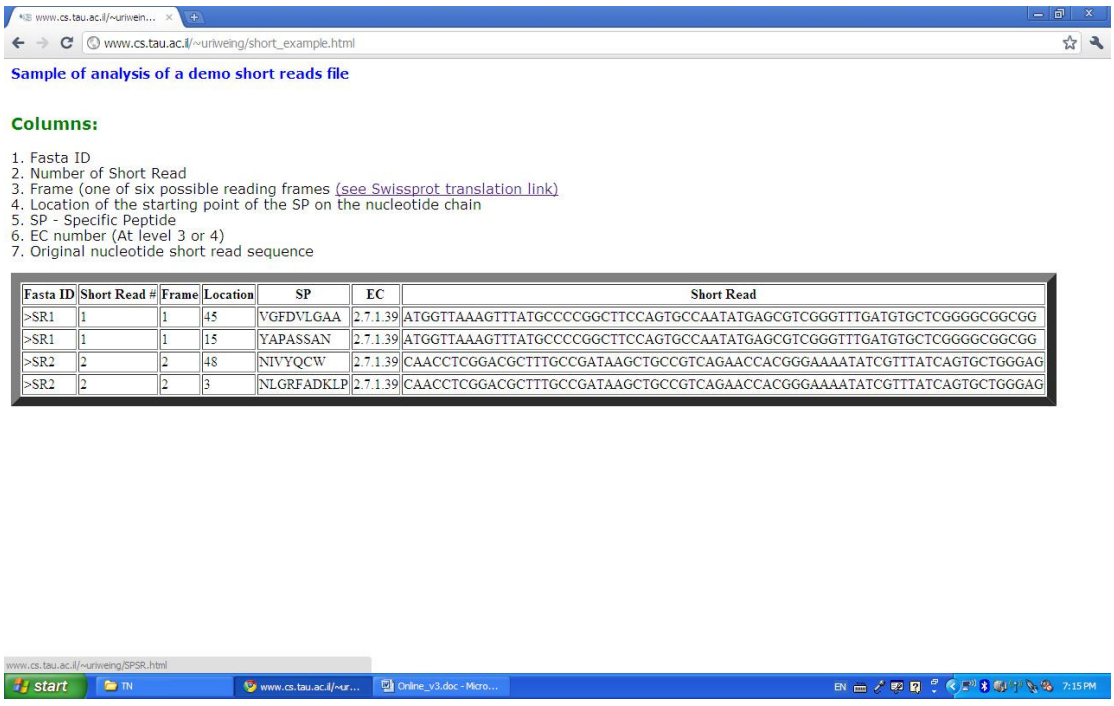


Figure 7.4: Sample output results of online utility to derive enzymatic signatures from short reads.

7.3 Derivation of Taxonomic Signatures from Short Read Data

http://www.cs.tau.ac.il/~uriweing/tspSPSR.html

This application accepts as input a file containing short reads and generates as output the taxonomic signatures of the short reads provided.
Processing consists of translation of each one of the short reads into the 6 reading frames and generating pseudo-peptides.
Search is conducted to find all Taxonomic SPs with EC=6.1.1 within each one of the pseudo peptides. Results are aggregated at the level of super kingdom, phyla and class.
This application differs from DME and SPSR because of its more sophisticated architectural design: To enable processing of large files, process is asynchronous: Each short read file is uploaded to the server, time stamped and an independently processed by a started task which scans the request queue a few times a day. Results are provided to the user via a notification email providing him with a link to the web page containing the results.

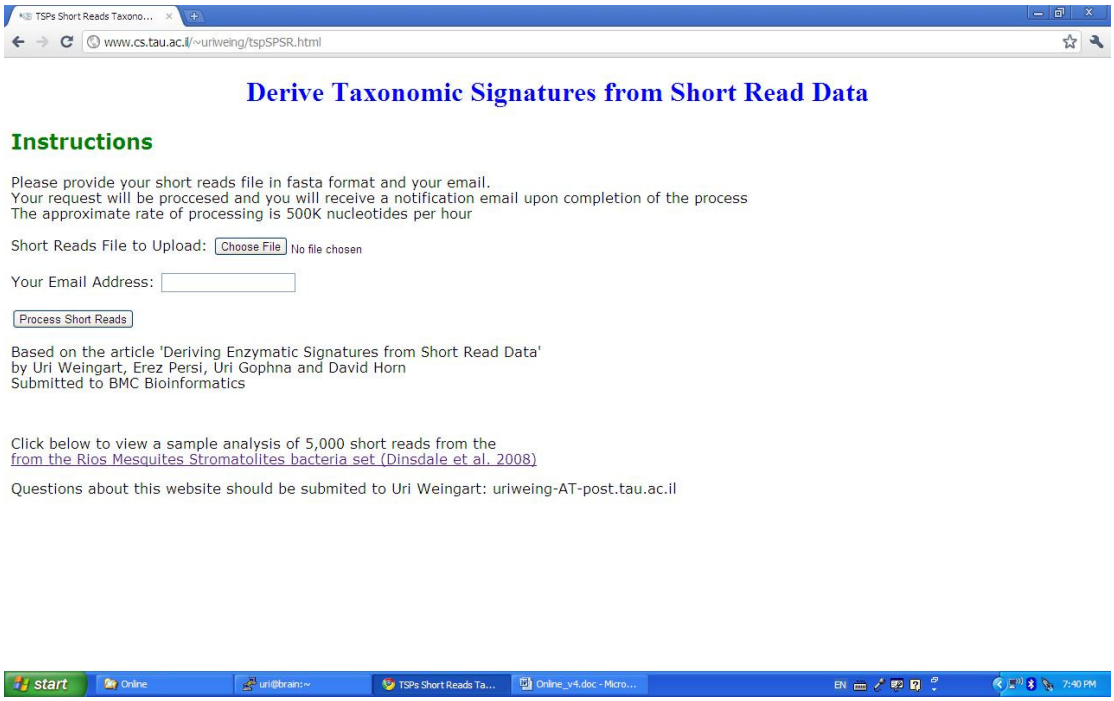
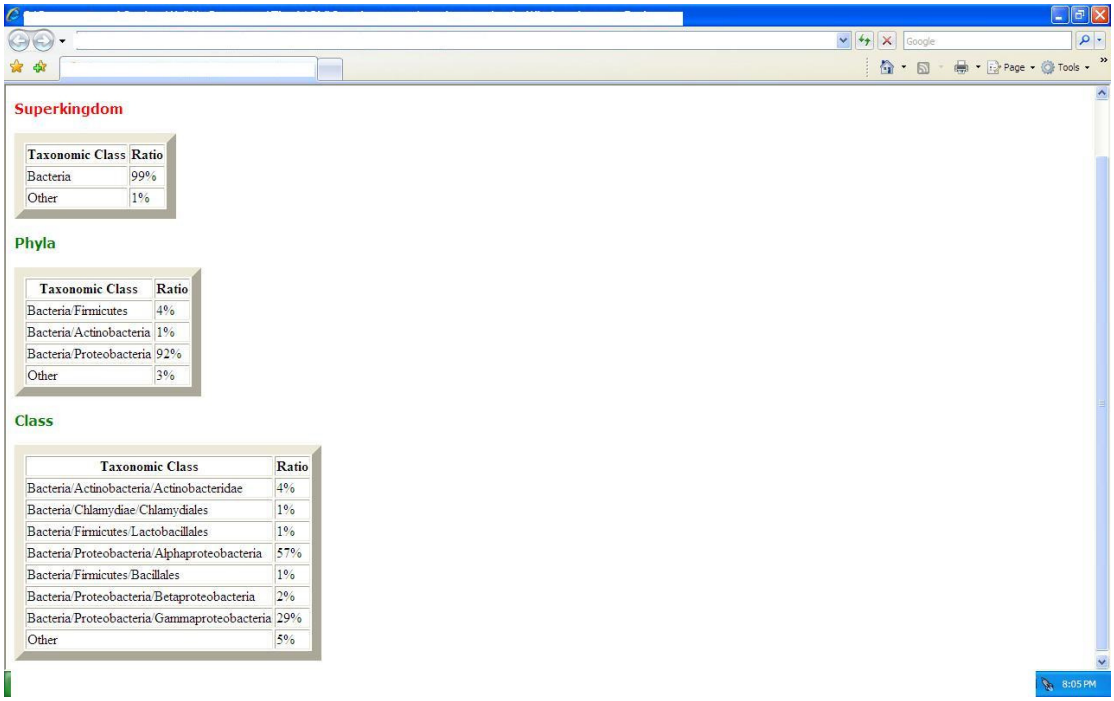


Figure 7.5: Input screen to derive Taxonomic Signatures from Short Read Data application.

Sample of results from this application for the Rios Mesquites Stromatolites bacteria set (Dinsdale et al. 2008) are shown in figure f_8.06 below:



Superkingdom

Taxonomic Class	Ratio
Bacteria	99%
Other	1%

Phyla

Taxonomic Class	Ratio
Bacteria/Firmicutes	4%
Bacteria/Actinobacteria	1%
Bacteria/Proteobacteria	92%
Other	3%

Class

Taxonomic Class	Ratio
Bacteria/Actinobacteria/Actinobacteridae	4%
Bacteria/Chlamydiae/Chlamydiales	1%
Bacteria/Firmicutes/Lactobacillales	1%
Bacteria/Proteobacteria/Alphaproteobacteria	57%
Bacteria/Firmicutes/Bacillales	1%
Bacteria/Proteobacteria/Betaproteobacteria	2%
Bacteria/Proteobacteria/Gammaproteobacteria	29%
Other	5%

Figure 7.6: Derivation of taxonomic analysis of Rios Mesquites Stromatolites bacteria set (Dinsdale et al. 2008)

Appendix 1

Abbreviation List

Abbreviation	Description
ASP	Annotated Specific Peptide
DME	Data Mining of Enzymes
DP	Different Prediction
EC	Enzyme Commission functional classification
Enzome	The conglomerate of all enzymes
FP	False Positive
FSP	Family Specific Peptides
GSP	Gene Ontology based Specific Peptides
HSP	High-scoring segment pairs (Blast)
MEX	Motif Extraction
NP	No Prediction
SP	Specific Peptide
SP V1.0	The first production Specific Peptides dataset – 87,017 SPs
SP V2.3	The second production Specific Peptides datasets – 148K SPs
SPSR	Specific Peptide Short Reads Method, used to generate enzymatic functionality predictions
TP	True Positive
TSP	Taxon Specific Peptide
TSPSR	Taxon Specific Peptides Short Read method used to generate taxonomic lineage predictions

Appendix 2

Table A2.1: List of DME predicted single EC annotations of proteins in Sargasso-Sea data.

This table can be downloaded from http://adios.tau.ac.il/DME_Additional_Material/

Table A2.2: List of DME predicted double EC annotations of proteins in Sargasso-Sea data

Entries ann1 and ann2 refer to the two DME predicted EC annotations. Entries cov1 and cov2 refer to coverage length L4 of the relevant annotations.

id	ann1	cov1	ann2	cov2
1084002109508	1.1.1.205	28	1.7.1.7	9
1084002025518	1.1.1.205	22	1.7.1.7	9
1084001361194	1.1.1.205	21	1.7.1.7	9
1084000026840	1.1.1.205	19	1.7.1.7	9
1084000038344	1.1.1.205	19	1.7.1.7	9
1084001646054	1.1.1.205	19	1.7.1.7	9
1084000060030	2.7.1.25	18	2.7.7.4	8
1084002379210	2.7.7.4	43	2.7.1.25	39
1087009015737	2.7.7.4	43	2.7.1.25	31
1087011115979	2.7.7.4	43	2.7.1.25	29
1087011915511	2.7.7.4	36	2.7.1.25	28
1087008719463	2.7.7.4	31	2.7.1.25	29
1087009908931	3.5.4.25	40	4.1.99.12	15
1084002372452	3.5.4.25	36	4.1.99.12	21
1087012809981	3.5.4.25	36	4.1.99.12	17
1087011308203	3.5.4.25	33	4.1.99.12	15
1084000038672	3.5.4.25	25	4.1.99.12	14
1084000045336	3.5.4.25	25	4.1.99.12	14
1084001104054	3.5.4.25	23	4.1.99.12	15
1084002413820	3.5.4.25	23	4.1.99.12	15
1087009815615	3.5.4.25	23	4.1.99.12	15
1087010814081	3.5.4.25	23	4.1.99.12	15
1087009611321	3.5.4.25	21	4.1.99.12	9
1084000846178	3.5.4.25	15	4.1.99.12	14
1084001467546	3.5.4.25	15	4.1.99.12	14
1087009614331	3.5.4.25	15	4.1.99.12	9
1087011319241	3.5.4.25	15	4.1.99.12	14
1084001304578	3.5.4.25	10	4.1.99.12	9
1084002015760	3.6.3.44	40	2.7.1.130	14
1087009217043	3.6.3.44	40	2.7.1.130	32
1087012119815	3.6.3.44	40	2.7.1.130	34
1087008920589	3.6.3.44	37	2.7.1.130	26
1087012113699	3.6.3.44	32	2.7.1.130	26
1084001244192	3.6.3.44	23	2.7.1.130	14
1084001940220	4.1.99.12	28	3.5.4.25	10
1084002357612	4.1.99.12	26	3.5.4.25	25
1084002364102	4.1.99.12	25	3.5.4.25	17
1084001118292	4.1.99.12	23	3.5.4.25	10
1084001028776	4.1.99.12	19	3.5.4.25	10
1084001420468	4.1.99.12	17	3.5.4.25	10
1084001241634	4.1.99.12	15	3.5.4.25	9
1084002074160	4.1.99.12	15	3.5.4.25	9
1087011012271	4.1.99.12	15	3.5.4.25	9

.

Table A2.3:

List of DME predicted triple EC annotations of proteins in Sargasso-Sea data.

Triple enzymatic predictions for Sargasso-Sea data.
Entries ann1 to ann3 refer to the three DME predicted EC annotations. Entries cov1 to cov3 refer to coverage lengths L3 or L4 as appropriate to the relevant annotations.

id	ann1	cov1	ann2	cov2	ann3	cov3
1084001889232	3.6.3.42	22	2.7.1.130	17	3.6.3.43	14
1087009516643	2.7.1.130	38	3.6.3.42	22	3.6.3.43	14
1087009912363	2.7.1.130	37	3.6.3.42	22	3.6.3.43	14
1087010109485	2.7.1.130	49	3.6.3.42	22	3.6.3.43	14
1087010219179	3.6.3.20	16	3.6.3.28	14	3.6.3.25	11
1084000851830	3.6.3.36	17	3.6.3.25	13	3.6.3.28	10
1087011411429	2.7.1.130	21	3.6.3.43	14	3.6.3.42	8
1084001061496	3.6.3	17	6.3.2	9	4.2.1.11	7
1084001195166	1.1.1.205	14	1.7.1.7	9	1.1.1.158	7
1084001469430	1.1.1.85	23	1.1.1.41	17	1.1.1.42	7
1084002305812	2.7.9.1	74	6.1.1.14	35	4.2.1.11	7
1084002325204	3.5.4.25	19	4.1.99.12	9	3.6.3.14	7
1084002328128	3.5.4.25	15	4.1.99.12	9	3.6.3.14	7
1084002368428	2.7.1.130	26	3.6.3.44	13	6.1.1.7	7
1084002409024	1.2.4.2	76	4.1.1.71	20	2.7.4.6	7
1084002411944	2.3.1	32	5.4.3.8	9	5.1.1	7
1087010109959	1.1.1.85	23	1.1.1.41	17	1.1.1.42	7
1087011014211	2.7.1.130	21	3.6.3.43	14	3.5.1	7

תקציר

תזה זו מתמקדת בבנית ותאור השימוש בפפטידים יחודיים לצורך נבוי הפונקציונליות האנזימאטית ופונקציות ביולוגיות נוספות של חלבונים יחידים, מטה-פרוטאומים, גנומים ומטה-גנומים.
שני הפרקים הראשונים מתארים בפרוטרוט את יצירת קבוצות האנזימים היחודיים ואת תת הקבוצות של האנזימים היחודיים המשמשים לצורך יצירת תחזיות ביולוגיות, כגון אתרים פעילים, אתרי קישור, אתרי מתכות, אונטולוגיה של גנים ויוחסין טאקסונומי של רצפים.

הפרק השלישי מוקדש לכרית נתונים אנזימאטיים, שכוונתו שיטה להפעלת הפפטידים היחודיים על נתונים רב-חלבוניים כדי לקבל נבויים אנזימאטיים. אנו מציגים ומבססים את המתודולוגיה ומסיקים את הנבויים האנזימאטיים עבור שלושה מטה-גנומים גדולים, אחד מהם (המטה-גנום של ים סארגאסו) כולל בתוכו מעל למליון חלבונים.אנו מציגים עבורו את כל הנבויים האנזימאטיים, שכוללים מספר חלבונים בעלי פעילות אנזימאטית כפולה או משולשת. אנו מציגים את המושג של פרופיל אנזימאטי של מדגם מטה-גנומי ומדגימים אותו בעזרת הנתונים שחקרנו.

פרק ארבע מציג את הניתוח של גנומים ומציג את היכולות של הפפטידים היחודיים לחשוף גנים אנזימאטיים על פני גנום שלם ללא ידע מוקדם על גבולות הגנים. אנו מציגים את המושג של "פיגום גנים בעזרת פפטידים יחודיים" המיועד לאמוד בקירוב התחלה וסוף של גנים. גנים אנזימאטיים מאופינים הן על ידי הפפטידים היחודיים והן על ידי פפטידים יחודיים משפחתיים, איתם אפשר לנבא שמות של גנים (או משפחות חלבונים). אנו מציגים את היכולת לגלות תזוזות ברצפים נוקלאוטידיים שנוצרים כתוצאה של תוספת או מחיקה של נוקלאוטיד יחיד או של מספר נוקלאוטידים ברצף קידוד, תוך כדי גילוי עקבות שנויים אבולוציונים גנטיים. אנו מציגים בתור דוגמא את הניתוח של ה. פילורי 26995, תוך כדי הדגמת התחזיות האנזימאטיות עבור כל הגנום.

פרק חמש מוקדש ליישום המתודולוגיה שתוארה בפרקים הקודמים לנתוח של short-reads של מטה-גנומים, בלי הצורך להפעיל שחזור של גנים או ייצור contigs.
אנו מציגים את שיטת הפפטידים היחודיים לקריאות קצרות ליצור נבויים טאקסונומיים. אנו מיישמים את המושגים האלה על כמה מטה-גנומים ומסיקים את החותמות הטאקסונומיות והאנזימאטיות שלהם.

בעוד שמירב הדגש של המחקר הוקדש על עיבוד נפח גדול של נתונים ביו-אינפורמטיים, יצרנו מספר כלים ברשת כדי להראות מושגים רבים שפותחו כחלק מהתזה באופן מקוון. פרק שבע מציג את הכלים האלה.

הפשטות והגמישות הרבה של הפפטידים היחודיים מוצגת בפרקים השונים של התזה: השימוש בפפטידים יחודיים כרוך אך ורק בחיפוש פשוט של תת-רצף (פפטיד יחודי) בתוך רצף (חלבון או תרגום של 6 מסגרות הנוקלאוטידים או נתונים שמקורם בקריאות קצרות) שאחריו חישוב כיסוי ואנליזה של התוצאות. אנו מראים בתזה שהפשטות והגמישות האלה מאפשרים ליצור מיגוון רב של תחזיות ביולוגיות תוך כדי שימוש במתודולוגיות מאד דומות.

TEL AVIV UNIVERSITY אוניברסיטת תל-אביב
RAYMOND AND BEVERLY SACKLER FACULTY OF EXACT SCIENCES
SCHOOL OF PHYSICS AND ASTRONOMY
הפקולטה למדעים מדויקים ע"ש ריימונד ובברלי סאקלר
בית-הספר לפיזיקה ולאסטרונומיה

חקר האנזום בעזרת פפטידים יחודיים

תזה מוגשת לקראת

דוקטור לפילוסופיה

על ידי

אורי וינגרט

הוגש לסנאט של אוניברסיטת תל-אביב

ינואר 2011

עבודה זו התבצעה בהנחייתו

של פרופ׳ דוד הורן