**Specific Peptides Predict Protein Classification.**

David Horn* and Uri Weingart

School of Physics and Astronomy, Tel Aviv University, Tel Aviv 69978, Israel

*corresponding author

**Abstract**

The methodology of Specific Peptides (SP) has been introduced within the context of enzymes. It is based on an unsupervised machine leaning (ML) tool for motif extraction, followed by supervised annotation of the motifs. In the case of enzymes, the classifier is the Enzyme Classification (EC) number. Here we demonstrate that this method reaches precision of 96.5% and recall of 89.1% on presently available protein sequences. We also apply this method to two other protein families, GPCR and ZF, find their corresponding SPs, and provide the code for searching any protein sequence for its classification under any such family.

## 1. Introduction.

Genes were perceived well before they have been determined to exist on chromosomes. In hindsight, it seems quite a surprise to find that they are just stretches of nucleotides within much larger sequences of DNA, often also interspersed by non-coding sections (introns). The identity of genes comes to life after being transcribed into RNA molecules, and translated into proteins, the important components of the machinery of living cells.

Proteins are molecular chains of amino acids. They are being studied by investigating the linear composition of amino-acid sequences, or their folding structures, or their functional properties, as revealed by their interactions with other molecules. In this paper, we discuss a different perspective of their structures, resulting from amino acid motifs, which are observed to be common to many proteins having the same function, or belonging to homolog genes of different species.

We follow the methodology developed and tested in [1-3], pointing out the existence of **Specific Peptides (SP**s**)** which are motifs of length $\geq 7$ amino acids, occurring on enzymes only. We reanalyze all enzymes using the updated Enzyme

Classification (EC) labelling, employed by SwissProt [4]. This analysis demonstrates the high predictive power of enzymatic SPs which will be labelled ESPs. This is followed by analyzing GPCR proteins, including the large group of non-OR proteins, expanding the results of a previous study of OR proteins [5].  We then continue to analyze Zinc Finger proteins and find their relevant SPs. Thus we end up with SPs defined for proteins belonging to all these families.

The analysis starts with the motif extraction method MEX [6], which is an unsupervised algorithm finding motifs with high occurrence in a given text. In enzyme classification this text contains 90% of all enzymes in the data set [4], to be labelled as the training set Ptrain. Next we use a supervised methodology: labelling all EC motifs according to the EC assignments of proteins in Ptrain, and discarding motifs which appear in Ntrain, which contains 90% of the non-enzymatic proteins in the data. The prediction accuracy is finally tested on the remaining 10% of the data, Ptest and Ntest.

Clearly assignments of SPs should be considered within the context in which they were derived. They are not supposed to annotate a free peptide, but only the motif appearing within a protein sequence. Still, as such, they can help identifying and annotating novel proteins, and may turn out to be very useful for artificial protein engineering [7] and for medical research and development [8].

## 2. Results
### 2.1 Enzyme Specific Peptides

The SwissProt entry [4] (version 2021_01) contains 564,227 proteins of many species. In order to enable a training and testing procedure we divided randomly the enzymes which had a single EC annotation into two sets: 227,488 were designated to a training set (Ptrain) and 25,309 enzymes were designated to a test set (Ptest). In parallel we also constructed non-enzymatic training and test sets, Ntrain and Ntest, containing 264,739 and 29,416 proteins correspondingly. This validation set serves to discard motifs which are not specific to enzymes.

Using the Enzyme Classification (EC) nomenclature, enzymes are classified into seven classes, EC1 to EC7, and within each EC class they are grouped into a hierarchy of four levels.

Some are classified just into the first level, numbered by the class, some at levels 2 or 3, but most at level 4, which is often associated with homologs of the same gene in different species. Proteins which have enzymatic regions belonging to two different EC classes were discarded from the training set.

Following [4] we restricted our MEX search to motifs of length ≥7 amino acids. Details of our procedure of analysis are explained in the Methods section.

Our procedure leads to a set of 286,755 specific peptides which we label as ESPs. They are provided as a Json list in our github entry [9] which also includes the code for searching a protein for the occurrence of such ESPs.

In order to test the usefulness of ESPs in predicting the EC labelling of a protein, we ran it on the test sets Ptest and Ntest. An SP hit on Ptest is regarded as true positive (TP) if the Swissprot EC assignment of the enzyme appears on the EC tree of the SP. If no SP hits an enzyme, it is labelled as false negative (FN). If an SP hits a protein in Ntest, the latter is declared as false positive (FP). If no SP hits a protein in Ntest, it is regarded as true negative (TN).

The results are presented in Table 1:

| TP | FP | FN | TN | Precision | Recall |
|---|---|---|---|---|---|
| 22,283 | 806 | 2,726 | 28,910 | 96.5% | 89.1% |

Table 1. Classification of Enzymes according to ESPs.

We use conventional definitions of Precision=TP/(TP+FP) and Recall=TP/(TP+FN). The sizes of Ptest and Ntest are 25,309 and 29,416 correspondingly. The 806 FP events include 300 from Ptest (with mismatched EC assignments) and 506 from Ntest.

## 2.2 GPCR

G protein coupling receptors (GPCR) play dominant roles in olfaction, vision and many other cellular functions.

Olfactory Receptors (OR) were studied in [5] using motifs of length ≥5 derived by the MEX methodology. They [5] have demonstrated how the resulting motifs can be employed in providing the sketch of an evolutionary tree of species, and have provided a web-service for OR protein assignment on the basis of these motifs.

We extend our analysis to all Swissprot GPCRs. After motif extraction we start with human GPCRs, and exhibit the specificity of all motifs of length $\geq 7$ to either OR proteins, or to non-OR (NOR) proteins within all GPCRs. There exist 156 OR motifs of length $\geq 7$ with hits on the 469 human ORs, and 2896 NOR motifs hitting the 148 human NOR proteins. There is no overlap between these lists, i.e. they are specific to either OR or NOR proteins.

While the number of human OR proteins (469) is larger than the NOR proteins (148), the number of the NOR motifs is much larger (2896 vs 156 of length $\geq 7$) because of the large variety of modalities which are served by NOR proteins. Turning to all Swissprot GPCRs, for all organisms, we expect to find a clear separation between OR and NOR, as well as discover a very large number of protein-SP biclusters in the NOR GPCRs. We find 562 OR proteins, leading to 367 OR motifs with length $\geq 7$, and 2481 NOR proteins with 3710 corresponding motifs. Once again, the two sets of motifs are specific to the two sets of proteins. We then proceed to search for protein-SP biclustering of the NOR data. The large number of NOR proteins and motifs allows for their sorting into clusters, as listed in Table 2. Both OR and NOR SPs will be referred to as GSPs. Their lists are provided as Json files in our github entry [9].

| # | Function | # proteins | #SPs | #motifs |
|---|----------|-----------|------|---------|
| 1 | 5-hydroxytryptamine receptor | 93 | 98 | 151 |
| 2 | Adhesion G protein-coupled receptor | 49 | 92 | 134 |
| 3 | Alpha  adrenergic receptor | 52 | 62 | 92 |
| 4 | Angiotensin II receptor | 22 | 22 | 31 |
| 5 | Beta adrenergic receptor | 49 | 78 | 116 |
| 6 | Blue-sensitive opsin  - Green-sensitive opsin  -  Rhodopsin | 156 | 160 | 269 |
| 7 | Cadherin EGF LAG seven-pass G-type receptor | 9 | 30 | 62 |
| 8 | Chemokine-like receptor | 146 | 108 | 177 |
| 9 | Dopamine receptor | 43 | 49 | 76 |
| 10 | Frizzled | 53 | 117 | 119 |
| 11 | G protein-coupled receptor kinase | 12 | 19 | 40 |
| 12 | Galanin receptor type | 11 | 3 | 10 |
| 13 | Gamma-aminobutyric acid type B receptor subunit | 4 | 1 | 5 |
| 14 | Gastric inhibitory polypeptide receptor | 7 | 3 | 10 |
| 15 | Gastrin/cholecystokinin type B receptor | 10 | 9 | 14 |
| 16 | Golgi pH regulator | 9 | 13 | 18 |
| 17 | Gonadotropin-releasing hormone receptor | 17 | 11 | 20 |
| 18 | G-protein coupled bile acid receptor | 5 | 3 | 4 |
| 19 | G-protein coupled receptor | 163 | 160 | 273 |
| 20 | Growth hormone-releasing hormone receptor | 11 | 6 | 13 |
| 21 | Histamine   receptor | 20 | 41 | 60 |
| 22 | Hydroxycarboxylic acid receptor | 6 | 2 | 4 |
| 23 | Latrophilin Cirl | 10 | 66 | 83 |
| 24 | Leukotriene B4 receptor | 4 | 2 | 4 |
| 25 | Lutropin-choriogonadotropic hormone receptor | 12 | 13 | 28 |
| 26 | Lysophosphatidic acid receptor | 17 | 11 | 17 |
| 27 | Medium-wave-sensitive opsin | 27 | 50 | 44 |
| 28 | Melanin-concentrating hormone receptor | 6 | 4 | 5 |
| 29 | Melanocortin receptor | 19 | 12 | 21 |
| 30 | Melanocyte-stimulating hormone receptor | 81 | 118 | 146 |
| 31 | Melanopsin | 10 | 9 | 13 |
| 32 | Melatonin-related receptor | 24 | 10 | 23 |
| 33 | Metabotropic glutamate receptor | 45 | 100 | 146 |
| 34 | Muscarinic acetylcholine receptor | 35 | 73 | 108 |
| 35 | Mu-type opioid receptor | 13 | 5 | 25 |
| 36 | N-arachidonyl glycine receptor | 5 | 3 | 4 |
| 37 | Neuromedin  receptor -- Neuropeptide  receptor | 47 | 29 | 50 |
| 38 | N-formyl peptide receptor | 15 | 16 | 24 |
| 39 | Nociceptin receptor | 5 | 6 | 8 |
| 40 | Orexin receptor type | 10 | 17 | 25 |
| 41 | Oxytocin receptor | 13 | 10 | 22 |
| 42 | P2Y purinoceptor | 29 | 23 | 39 |
| 43 | Parathyroid hormone/parathyroid hormone-related peptide receptor | 12 | 16 | 28 |
| 44 | Pituitary adenylate cyclase-activating polypeptide type I receptor | 4 | 1 | 7 |
| 45 | Platelet-activating factor receptor | 8 | 7 | 10 |
| 46 | Prokineticin receptor | 8 | 7 | 11 |
| 47 | Prostaglandin   receptor | 29 | 26 | 39 |
| 48 | Proteinase-activated receptor | 18 | 11 | 19 |
| 49 | Proto-oncogene Mas | 5 | 1 | 2 |
| 50 | Relaxin receptor | 7 | 5 | 8 |
| 51 | Serpentine receptor class | 18 | 8 | 14 |
| 52 | Short-wave-sensitive opsin | 11 | 11 | 22 |
| 53 | Smoothened homolog | 4 | 12 | 13 |
| 54 | Somatostatin receptor type | 21 | 13 | 23 |
| 55 | Sphingosine 1-phosphate receptor | 16 | 14 | 19 |
| 56 | Substance-K receptor | 10 | 5 | 19 |
| 57 | Taste receptor member | 49 | 99 | 141 |
| 58 | Thromboxane A2 receptor | 5 | 3 | 4 |
| 59 | Thyrotropin receptor | 14 | 24 | 34 |
| 60 | Trace amine-associated receptor | 44 | 41 | 58 |
| 61 | Urotensin-2 receptor | 5 | 4 | 6 |
| 62 | Vasoactive intestinal polypeptide receptor | 19 | 5 | 16 |
| 63 | Vasopressin  receptor | 12 | 14 | 23 |
| 64 | Vomeronasal type-1 receptor | 26 | 22 | 39 |

Table 2. 64 Protein clusters among NOR GPCR. #SPs refers to those which are specific to the cluster, while #motifs refers to other GSPs occurring in other NOR clusters as well.

Next we run all the GSPs against the Enzymes of our EC study. We find 63 hits of these motifs on 3 EC classes, thus providing EC identifications of 3 NOR classes. They are listed in Table 3:

| EC classification | NOR classification |
|---|---|
| 2.7.11.14   Rhodopsin kinase | Blue-sensitive opsin, Green-sensitive opsin, Rhodopsin |
| 2.7.11.15   [Beta-adrenergic-receptor] kinase | Beta adrenergic receptor |
| 2.7.11.16   [G-protein-coupled receptor] kinase | G protein-coupled receptor kinase |

Table 3. Three NOR classes which belong to three EC classes.

There exist 20 other sporadic hits of NOR GSPs on EC proteins, which are consistent with expected noise.

## 2.3 Zinc Finger proteins

We have analyzed 2582 Swissprot ZF proteins and extracted 1487 motifs of length $\geq 7$ which are declared to be ZSPs. 786 of all the proteins are human ZF proteins, and they display hits by 1412 of the SPs.

Since ZF proteins may contain several ZF domains, we encounter reappearance of motifs on different locations within the same protein. This is different from our previous studies of EC and GPCR proteins, where inter-protein multiple appearances were responsible for the generation of MEX motifs. Here we find that intra-protein recurrences play an important role.

To illustrate this fact, we display in Table 4 some of the "popular" ZSPs, which have 100 or more hits on all human ZF proteins. On the right of the table we provide the sum of all shown hits for each protein as displayed here, and compare it to the total number of ZSPs hitting each protein. On the bottom we provide analogous counts for each ZSP.

It should be realized that SPs of length n can be contained within SPs of length >n, as can be seen in this table, which serves as an examples rather than a summary. Summary of all ZSPs and their hits on ZF proteins is provided in our github entry [9].

Clearly the repetitive appearances of SPs on a given protein reflect the existence of many ZF regions on the same protein. The latter is usually larger than the number of repeats of a single SP, since different SPs may belong to different ZF regions.

| SP | CEECGKA | EKPYKCE | GEKPYKC | HKIIHTG | HKRIHTGEK | HTGEKPY | HTGEKPYK | IHTGEKP | IHTGEKPY | KCEECGK | KPYKCEE | PYKCEECGK | PYKCEECGKAF | RIHTGEK | YKCEECG | shown hits | total hits |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| length | 7 | 7 | 10 | 7 | 9 | 7 | 10 | 7 | 8 | 7 | 10 | 9 | 11 | 7 | 7 | | |
| A6NK75 | 9 | 7 | 7 | 3 | 2 | 5 | 5 | 5 | 5 | 9 | 8 | 8 | 8 | 2 | 10 | 93 | 169 |
| A6NN14 | 25 | 16 | 14 | 11 | 1 | 13 | 13 | 15 | 12 | 26 | 20 | 20 | 19 | 1 | 22 | 228 | 386 |
| A6NNF4 | 10 | 14 | 12 | 4 | 4 | 11 | 10 | 9 | 8 | 13 | 11 | 11 | 9 | 4 | 12 | 142 | 251 |
| A8MQ14 | 0 | 0 | 0 | 0 | 1 | 23 | 0 | 20 | 19 | 0 | 0 | 0 | 0 | 14 | 0 | 77 | 123 |
| A8MTY0 | 7 | 6 | 6 | 4 | 3 | 8 | 5 | 8 | 8 | 7 | 7 | 7 | 7 | 3 | 8 | 94 | 174 |
| A8MXY4 | 18 | 6 | 5 | 6 | 0 | 5 | 5 | 5 | 5 | 19 | 14 | 14 | 14 | 0 | 15 | 131 | 221 |
| O43345 | 19 | 24 | 20 | 1 | 4 | 18 | 17 | 18 | 17 | 28 | 23 | 25 | 11 | 4 | 27 | 256 | 418 |
| O75346 | 6 | 5 | 5 | 0 | 3 | 9 | 5 | 6 | 6 | 5 | 5 | 5 | 4 | 3 | 5 | 72 | 127 |
| O75373 | 7 | 9 | 8 | 2 | 6 | 6 | 6 | 7 | 6 | 8 | 6 | 6 | 5 | 6 | 8 | 96 | 169 |
| O75437 | 7 | 9 | 6 | 4 | 1 | 6 | 6 | 4 | 4 | 10 | 9 | 9 | 7 | 1 | 11 | 94 | 161 |
| O95780 | 6 | 5 | 4 | 0 | 4 | 6 | 4 | 5 | 5 | 5 | 4 | 4 | 4 | 4 | 5 | 65 | 126 |
| P0DKX0 | 11 | 6 | 6 | 1 | 3 | 4 | 4 | 6 | 4 | 14 | 7 | 7 | 6 | 4 | 12 | 95 | 167 |
| P0DPD5 | 8 | 7 | 5 | 1 | 7 | 6 | 5 | 6 | 6 | 9 | 7 | 7 | 6 | 7 | 9 | 96 | 190 |
| P17019 | 9 | 8 | 7 | 4 | 1 | 9 | 7 | 8 | 8 | 11 | 10 | 10 | 8 | 1 | 11 | 112 | 192 |
| P17038 | 14 | 9 | 5 | 2 | 2 | 7 | 6 | 5 | 5 | 13 | 10 | 11 | 11 | 3 | 13 | 116 | 207 |
| P35789 | 11 | 6 | 5 | 0 | 2 | 10 | 5 | 8 | 7 | 9 | 8 | 8 | 8 | 2 | 9 | 98 | 177 |
| P52742 | 0 | 0 | 0 | 0 | 0 | 13 | 0 | 8 | 8 | 0 | 0 | 0 | 0 | 8 | 0 | 37 | 100 |
| Q02386 | 0 | 7 | 5 | 0 | 0 | 13 | 6 | 3 | 3 | 6 | 5 | 6 | 0 | 3 | 7 | 64 | 110 |
| Q03923 | 5 | 7 | 5 | 6 | 2 | 12 | 6 | 12 | 12 | 6 | 5 | 5 | 3 | 2 | 6 | 94 | 154 |
| Q03924 | 7 | 4 | 3 | 1 | 2 | 4 | 2 | 4 | 4 | 6 | 4 | 4 | 4 | 2 | 6 | 57 | 111 |
| Q03936 | 9 | 7 | 4 | 5 | 1 | 6 | 5 | 6 | 6 | 9 | 8 | 8 | 8 | 1 | 9 | 92 | 163 |
| Q03938 | 8 | 4 | 3 | 1 | 5 | 4 | 2 | 3 | 3 | 8 | 6 | 7 | 7 | 5 | 7 | 73 | 132 |
| Q05481 | 21 | 14 | 12 | 6 | 7 | 18 | 13 | 15 | 15 | 21 | 14 | 14 | 14 | 7 | 21 | 212 | 400 |
| Q14593 | 5 | 5 | 3 | 3 | 2 | 6 | 3 | 4 | 4 | 5 | 4 | 5 | 5 | 2 | 6 | 62 | 120 |
| Q5SXM1 | 3 | 5 | 4 | 0 | 9 | 10 | 4 | 10 | 10 | 4 | 4 | 4 | 2 | 10 | 6 | 85 | 160 |
| Q68DY1 | 9 | 4 | 4 | 2 | 1 | 5 | 4 | 4 | 4 | 9 | 6 | 8 | 8 | 2 | 9 | 79 | 138 |
| Q6ZN08 | 4 | 7 | 7 | 5 | 3 | 9 | 7 | 9 | 9 | 8 | 8 | 8 | 4 | 3 | 9 | 100 | 164 |
| Q6ZR52 | 12 | 10 | 6 | 2 | 2 | 5 | 5 | 6 | 5 | 11 | 10 | 10 | 9 | 2 | 11 | 106 | 185 |
| Q86V71 | 7 | 7 | 6 | 0 | 2 | 7 | 4 | 7 | 7 | 6 | 6 | 6 | 6 | 3 | 6 | 80 | 150 |
| Q8IW36 | 6 | 5 | 3 | 0 | 2 | 4 | 3 | 4 | 4 | 6 | 6 | 6 | 4 | 4 | 6 | 63 | 117 |
| Q8IYB9 | 10 | 3 | 2 | 1 | 4 | 9 | 2 | 10 | 9 | 6 | 3 | 3 | 3 | 4 | 6 | 75 | 138 |
| Q8IYN0 | 6 | 5 | 5 | 1 | 1 | 6 | 4 | 6 | 6 | 6 | 5 | 5 | 5 | 2 | 6 | 69 | 120 |
| Q8N7Q3 | 11 | 13 | 9 | 4 | 3 | 8 | 8 | 8 | 8 | 13 | 12 | 12 | 8 | 3 | 13 | 133 | 225 |
| Q8N972 | 0 | 0 | 0 | 0 | 0 | 13 | 0 | 3 | 3 | 0 | 0 | 0 | 0 | 3 | 0 | 22 | 107 |
| Q8TAQ5 | 2 | 2 | 2 | 0 | 1 | 15 | 2 | 12 | 12 | 2 | 2 | 2 | 2 | 11 | 2 | 69 | 153 |
| Q8TD23 | 9 | 6 | 4 | 2 | 3 | 5 | 5 | 5 | 5 | 8 | 6 | 8 | 8 | 3 | 8 | 85 | 155 |
| Q8TF20 | 7 | 7 | 5 | 0 | 3 | 13 | 5 | 16 | 13 | 7 | 5 | 5 | 5 | 11 | 7 | 109 | 201 |
| Q8TF32 | 8 | 6 | 5 | 5 | 3 | 6 | 5 | 7 | 6 | 7 | 4 | 4 | 4 | 3 | 6 | 79 | 140 |
| Q96IR2 | 0 | 5 | 5 | 0 | 0 | 17 | 4 | 9 | 9 | 0 | 0 | 0 | 0 | 5 | 0 | 54 | 108 |
| Q96N22 | 9 | 4 | 3 | 3 | 3 | 8 | 3 | 8 | 8 | 5 | 4 | 4 | 4 | 3 | 5 | 74 | 134 |
| Q96N38 | 9 | 5 | 4 | 2 | 2 | 5 | 4 | 6 | 5 | 8 | 4 | 4 | 4 | 2 | 6 | 70 | 130 |
| Q96SE7 | 0 | 0 | 0 | 0 | 0 | 13 | 0 | 12 | 12 | 0 | 0 | 0 | 0 | 3 | 0 | 40 | 100 |
| Q9H7R5 | 0 | 0 | 0 | 0 | 1 | 13 | 0 | 13 | 12 | 0 | 0 | 0 | 0 | 7 | 0 | 46 | 114 |
| Q9H8G1 | 4 | 6 | 4 | 4 | 2 | 7 | 4 | 7 | 7 | 5 | 4 | 4 | 4 | 3 | 5 | 70 | 119 |
| Q9HCG1 | 0 | 0 | 0 | 0 | 0 | 13 | 0 | 10 | 9 | 0 | 0 | 0 | 0 | 5 | 0 | 37 | 107 |
| Q9P255 | 8 | 6 | 6 | 3 | 2 | 5 | 4 | 5 | 5 | 8 | 7 | 7 | 7 | 2 | 9 | 84 | 154 |
| Q9UII5 | 9 | 10 | 6 | 5 | 3 | 12 | 7 | 12 | 12 | 11 | 7 | 7 | 5 | 3 | 9 | 118 | 201 |
| Q9Y2Q1 | 9 | 8 | 6 | 4 | 1 | 6 | 6 | 6 | 5 | 9 | 8 | 8 | 7 | 1 | 8 | 92 | 165 |
| shown hits | 364 | 309 | 246 | 108 | 114 | 436 | 230 | 385 | 365 | 386 | 306 | 316 | 267 | 187 | 376 | | |
| total hits | 473 | 433 | 322 | 127 | 192 | 2136 | 327 | 1352 | 1158 | 509 | 364 | 391 | 303 | 966 | 477 | | |

Table 4. Number of hits by different SPs, displayed on different human ZF proteins. Large numbers correlate with the fact that many ZF regions can be found on the same protein.

There exist some proteins which act as enzymes and possess zing fingers. One outstanding example is PRDM9. This protein serves recombination hotspots during meiosis by binding nucleotides with its zinc fingers. The annotations of the human version of this protein are provided by https://www.uniprot.org/uniprot/Q9NQV7. They contains 14 ZF regions. The first starts at location 388 and has length of 24 amino acids. The other 13 start at 524 and are of length 23 each. In Fig. 1 we display, in color code, the loci of hits by all ZSP and ESP motifs of length $\geq 7$.

Fig. 1 shows that the last 12 ZF domains display very high similarity and an exact periodicity. Comparing with Uniprot annotations we find that all ZF domains have the structure YVCRECxxxxxxxxHQRTHT, where the additional 8 amino acids,

replaced by x, vary according to the nucleotide targeted by the ZF domain. In Fig. 1 we note prevalent occurrence of the structure HQRTHTGEKPYVCRECGRGF which includes the suffix of a previous ZF domain and the prefix of the next ZF domain. The colors reflect occurrences of hits by ZSPs and ESPs. Between the suffix of one structure and the prefix of the next we find the quintet GEKPY which fills the gap between 23, the length of the ZF motif, and 28, the regular periodicity observed in this protein over the range of its last 12 ZF regions.

MSPEKSQEESPEEDTERTERKPMVKDAFKDISIYFTKEEWAEMGDWEKTR
YRNVKRNYNALITIGLRATRPAFMCHRRQAIKLQVDDTEDSDEEWTPRQQ
VKPPWMALRVEQRKHQKGMPKASFSNESSLKELSRTANLLNASGSEQAQ
KPVSPSGEASTSGQHSRLKLELRKKETERKMYSLRERKGHAYKEVSEPQD
DDYLYCEMCQNFFIDSCAAHGPPTFVKDSAVDKGHPNRSALSLPPGLRIGP
SGIPQAGLGVWNEASDLPLGLHFGPYEGRITEDEEAANNGYSWLITKGRN
CYEYVDGKDKSWANWMRYVNCARDDEEQNLVAFQYHRQIFYRTCRVIR
PGCELLVWYGDEYGQELGIKWGSKWKKELMAGREPKPEIHPCPSCCLAFS
SQKFLSQHVERNHSSQNFPGPSARKLLQPENPCPGDQNQEQQYPDPHSRN
DKTKGQEIKERSKLLNKRTWQREISRAFSSPPKGQMGSCRVGKRIMEEESR
TGQKVNPGNTGKLFVGVGISRIAKVKYGECGQGFSVKDVITHQRTHTGEK
LYVCRECGRGFSWKSHLLIHQRIHTGEKPYVCRECGRGFSWQSVLLTHQR
THTGEKPYVCRECGRGFSRQSVLLTHQRRHTGEKPYVCRECGRGFSRQSV
LLTHQRRHTGEKPYVCRECGRGFSWQSVLLTHQRTHTGEKPYVCRECGR
GFSWQSVLLTHQRTHTGEKPYVCRECGRGFSNKSHLLRHQRTHTGEKPYV
CRECGRGFRDKSHLLRHQRTHTGEKPYVCRECGRGFRDKSNLLSHQRTHT
GEKPYVCRECGRGFSNKSHLLRHQRTHTGEKPYVCRECGRGFRNKSHLLR
HQRTHTGEKPYVCRECGRGFSDRSSLCYHQRTHTGEKPYVCREDE

Color code:
CRECGRGF  is an ESP

HQRTHTGEKPYVC  is a ZSP
HQRTHTGEKPYVC  are compound hits by a ZSP and an ESP

Fig. 1. The sequence of PRDM9_HUMAN Histone-lysine N-methyltransferase (Q9NQV7) and color coded display of hits by ESPs of EC 2.1.1.43 and ZSPs, which may partially overlap each other.

## 3. Summary and Discussion.

Our methodology is based on machine learning (ML) practices: MEX is an unsupervised tool for motif extraction; these motifs are then searched on protein sequences using supervised annotation to classify the results. In the case of enzymes, the classifier is the Enzyme Classification which is defined in terms of seven classes and four levels in each class.

ESPs are specific peptides whose presence on the amino acid sequence of the protein indicates its EC number, as well as the tree associated with it. This methodology was introduced in 2007 [1]. Other ML studies appeared in the meantime, trying to solve the same (or related) problems using various ML tools. Many neglected to notice that SPs can do the required EC prediction quite well, often even better than the new tools.

Some examples of recent ML methodologies are DeepEC [10] and MAHOMES [11]. DeepEC employs 3 deep convolutional neural networks and a homology analysis tool to the study of enzyme sequences. When applying it to a test set which uses 201 enzymes they obtained precision = 0.92 and recall = 0.455 (quoted from Table 2 in [10]). This is considerably worse than our results in Table 1, which were based on a much larger (25K) test set. Other five ML methods which they [9] compared themselves to, were even worse.

MAHOMES[11] uses a decision-tree ML model, which is structure-based, employing physicochemical features specific to catalytic activity. Their main aim is to classify metals bound to proteins as enzymatic or non-enzymatic, and they succeed doing it with precision=0.922 and recall=0.901. Comparing to sequence-based technologies, they find that DeepEC scores on their tasks are precision = 0.905 and recall = 0.596. They find that another homology method, EFICAz2.5 [12] (which lost to DeepEC according to [10]), had better statistics

(precision=0.922 and recall=0.901) but still falls short of their own [11]. For an older review of ML studies of enzymes see [13].

Our precision/recall results attest to the usefulness of the MEX unsupervised methodology in discovering relevant and unique motifs, the specific peptides (SPs). Our approach is not limited to enzyme studies. We have demonstrated this flexibility by investigating GPCR and Zinc-finger proteins, leading to a wealth of novel SPs. We provide in [9] a documented python code which allows for SP searches of all the functionalities which we have studied. It contains the lists of 2,002 NOR GSPs, 351 OR GSPs and 1,482 ZSPs in addition to the 286,755 ESPs.

## Methods

Building the list of ESPs,

In order to run the Motif Extraction program (MEX) [6], we divided the enzymes training set into batches grouped by joint level 2 assignments, and batches of enzymes with single level 1 assignments. Following [4] we restricted our MEX search to motifs of length ≥7 amino acids. The analysis led to 307,989 motifs. All motifs were then annotated after collecting the information of the IDs of enzymes hit by a particular motif (i.e. occurring in full on the amino acid chain of the enzyme) and how many times was a particular enzyme hit by a particular motif.

The EC number description, indicating both class and level, can be viewed as an inverted tree with a maximum depth of 4. For every motif, we map the EC numbers of the enzymes it hits on the training set onto a single EC tree. Starting from level 4 and moving upwards, we search for the first level which is a unique descendent of a higher level. The EC number of this unique descendant is assigned to the motif.

In order to remove motifs which may occur also on non-enzymatic proteins, we search for hits of all motifs on the non-enzymatic Ntrain set. Such motifs are removed from the list of specific peptides. Thus, to summarize, a motif of length ≥ 7 amino acids is labeled as an Enzyme Specific Peptide (ESP) if:

   -it hits (i.e. appears in full on the amino acid chain of) enzymes belonging to only a single EC class of Ptrain

-and it does not hit any protein in Ntrain

This procedure leads to the reduction of the set of motifs to 286,755 specific peptides which we label as ESPs. They are provided as a Json list in our github entry [9] which also includes the code for searching the sequence of a protein for the occurrence of such ESPs on its amino acid chain..

## References

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

2. Yasmine Meroz and David Horn (2008) Biological roles of specific peptides in enzymes.
Proteins: Structure, Function, and Bioinformatics 72 (2), 606-612, 2008

3. Uri Weingart, Yair Lavi and David Horn (2009). Data mining of enzymes using specific peptides. BMC Bioinformatics 2009, 10:446
DME website: http://adios.cs.tau.ac.il/DME/

4. Swissprot, as provided by Uniprot.
https://ftp.uniprot.org/pub/databases/uniprot/previous_major_releases/release-2021_01/knowledgebase/UniProtKB_SwissProt-relstat.html

5. Assaf Gottlieb, Tsviya Olender, Doron Lancet and David Horn (2009). Common peptides shed light on evolution of Olfactory Receptors
BMC Evolutionary Biology 2009, 9:91. Olfactory Receptors Search Website: http://adios.cs.tau.ac.il/ORSearch/

6. Zach Solan, David Horn, Eytan Ruppin and Shimon Edelman (2005). Unsupervised learning of natural languages
Proc. Natl. Acad. Sc. 102 (2005) 11629-11634

7. Rusk, N. Protein circuit engineering. *Nat Methods* **15,** 860 (2018).

8. Jolene L. Lau, Michael K. Dunn (2018) Therapeutic peptides: Historical perspectives, current development trends, and future directions.
Bioorganic and Medical Chemistry 26 (2018) 2700-2707.

9. Uri Weingart. Github entry providing SP code and lists: https://github.com/uri-weingart/SPs

10. Jae Yong Ryua, Hyun Uk Kimc, and Sang Yup Lee (2019). Deep learning enables high-quality and high throughput prediction of enzyme commission numbers. PNAS 116 (2019) 13996-14001.
11. Feehan, R., Franklin, M.W. & Slusky, J.S.G. Machine learning differentiates enzymatic and non-enzymatic metals in proteins. *Nat Commun* **12,** 3712 (2021).
12. Kumar N and Skolnick J (2012). EFICAz2.5: application of a high-precision enzyme function predictor to 396 proteomes. Bioinformatics. 28(20): 2687–2688.
13. Mohammed A and Guda C (2011) Computational Approaches for Automated Classification of Enzyme Sequences. J Proteomics Bioinform. 2011 4: 147–152.