Functional Representation of Enzymes by Specific Peptides

Vered Kunik[1], Yasmine Meroz[2], Zach Solan[2], Ben Sandbank[1], Uri Weingart[2], Eytan Ruppin[1,3], David Horn[2*]

1 School of Computer Science, Tel Aviv University, Tel Aviv, Israel, 2 School of Physics and Astronomy, Tel Aviv University, Tel Aviv, Israel, 3 Sackler School of Medicine, Tel Aviv University, Tel Aviv, Israel

Predicting the function of a protein from its sequence is a long-standing goal of bioinformatic research. While sequence similarity is the most popular tool used for this purpose, sequence motifs may also subserve this goal. Here we develop a motif-based method consisting of applying an unsupervised motif extraction algorithm (MEX) to all enzyme sequences, and filtering the results by the four-level classification hierarchy of the Enzyme Commission (EC). The resulting motifs serve as specific peptides (SPs), appearing on single branches of the EC. In contrast to previous motif-based methods, the new method does not require any preprocessing by multiple sequence alignment, nor does it rely on over-representation of motifs within EC branches. The SPs obtained comprise on average 8.4 ± 4.5 amino acids, and specify the functions of 93% of all enzymes, which is much higher than the coverage of 63% provided by ProSite motifs. The SP classification thus compares favorably with previous function annotation methods and successfully demonstrates an added value in extreme cases where sequence similarity fails. Interestingly, SPs cover most of the annotated active and binding site amino acids, and occur in active-site neighboring 3-D pockets in a highly statistically significant manner. The latter are assumed to have strong biological relevance to the activity of the enzyme. Further filtering of SPs by biological functional annotations results in reduced small subsets of SPs that possess very large enzyme coverage. Overall, SPs both form a very useful tool for enzyme functional classification and bear responsibility for the catalytic biological function carried out by enzymes.

Citation: Kunik V, Meroz Y, Solan Z, Sandbank B, Weingart U, et al. (2007) Functional representation of enzymes by specific peptides. PLoS Comput Biol 3(8): e167. doi:10.1371/journal.pcbi.0030167

Introduction

One of the major efforts of computational research in molecular biology is to predict the function and spatial structure of proteins from the protein sequence of amino acids [1,2]. Conventional approaches to function prediction rely on sequence [3] or structure [4] similarity with proteins whose functions are known. This is sometimes misleading [4–6]. Alternatively, one may use motif approaches [7–12], trying to extract from the data subsequences that are responsible for particular functions. Motifs can be deterministic sequences of amino acids, regular expressions that allow various alternatives for specific locations within the motif, or stochastic structures specifying the probability of an amino acid at every location. This work aims to uncover deterministic sequence motifs, and considers their relationships with protein functionality. We focus on enzymes, whose functions are classified by the Enzyme Commission (EC) four-level hierarchy which is represented by four integers, n1.n2.n3.n4, corresponding to the different levels of classification. For example, the oxidoreductases class corresponds to n1 = 1, one of the six main divisions. For this class, n2 (subclass) specifies electron donors, n3 (sub-subclass) specifies electron acceptor, and n4 indicates the exact enzymatic activity.

Conventional sequence motif searches in enzymes are performed in a supervised fashion, using sequences of proteins that are known to have the same function and looking for (deterministic, regular-expression, or stochastic) motifs that are over-represented in this group of proteins. The motifs in question should then subserve such functions as [9] phosphorylation of protein kinases; metal binding sites for calcium, zinc, copper, and iron; enzyme active sites, etc. With the advent of studies of protein–protein interactions, interest grew in finding sequence motifs that are responsible for them, and span an "interaction space" [13,14].

Here we perform a large-scale search for deterministic sequence motifs without specifying a priori their exact functional roles, using the unsupervised motif extraction (MEX) algorithm [15]. We have used one functional guidance: MEX was separately applied to each one of the six major EC classes. The same motifs may also appear in other classes, yet many of them turn out to occur in only one class, and belong to a specific EC branch. The latter (see Figure 1A) are termed specific peptides (SPs). By representing some 50,000 enzymes (of average length of 380 amino acids) in terms of about the same number of SPs (of average length 8.4), we obtain a largely compressed functional representation and an EC classification with 93% accuracy.

This may be compared with other methods based on e-motifs [16], sequence similarity [17], or physicochemical properties of the amino acids contained in the sequence [18,19]. Our results compare favorably with such methods, as

Editor: Yanay Ofran, Columbia University, United States of America

Received January 17, 2007; **Accepted** July 10, 2007; **Published** August 24, 2007

A previous version of this article appeared as an Early Online Release on July 11, 2007 (doi:10.1371/journal.pcbi.0030167.eor).

Abbreviations: EC, Enzyme Commission; MSA, multiple sequence alignment; SP, specific peptides; SVM, support vector machine

* To whom correspondence should be addressed. E-mail: horn@tau.ac.il

Author Summary

Sequence motifs are known to provide information about functional properties of proteins. In the past, many approaches have looked for deterministic motifs in protein sequences, by searching for functionally over-represented k-mers, with moderate levels of success. Here we revisit and renew the utility of deterministic motifs, by searching for them in a partially unsupervised and context-dependent manner. Using a novel motif extraction algorithm, MEX, deterministic sequence motifs are extracted from Swiss Prot data containing more than 50,000 enzymes. They are then filtered by the Enzyme Commission classification hierarchy to produce sets of specific peptides (SPs). The latter specify enzyme function for 93% of the data, comparing well with existing approaches for enzyme classification. Importantly, SPs are found to have biological significance. A majority of all known active and binding sites of enzymes are covered by SPs, and many SPs are found to lie within spatial pockets in the neighborhood of the active sites. Both these results have extremely high statistical significance. A user-friendly tool that displays the hits of SPs for any protein sequence that is presented as a query, together with the EC assignments due to these SPs, is available at http://adios.tau.ac.il/SPSearch.

will be shown below, yet our approach differs in several respects: we use a largely unsupervised motif extraction method, we perform a comprehensive study of all enzymes, and we put major emphasis on the biological relevance of the SPs themselves.

Importantly, in comparison with the large-scale and popular motif database ProSite [8], our approach displays a wide-margin advantage, their motifs coverage extending only to 63% of all enzymes in the database.

Results

The Specific Peptides

SPs, as defined above, are MEX motifs that are specific to a single branch of the EC hierarchical classification. Most belong to single branches of the fourth level of the hierarchy, to be denoted as SPs of level 4 (SP4) (see Figure 1A). SPs of higher hierarchy, SP3, SP2, and SP1, appear in more than one lower EC level. Thus, if a peptide is shared by two or more level 4 groups that belong to the same third EC level, and appears nowhere else, it is assigned to SP3. The SPs were further screened to eliminate any peptide that includes within it another peptide carrying the same SPN (N = 1,2,3,4) label.

The majority of SPs found at level 4 of the EC hierarchy (Table 1) are probably due to the high homology within this level, that often includes many orthologous genes. Thousands of SPs occur at higher levels of hierarchy, reflecting functional similarity among enzymes with lower sequence similarity. The occurrence of any one SP on the sequence of an enzyme specifies its EC functionality according to the specific branch N of its SPN. For example, enzyme P45048 (see Figure 1B) contains SSAATYG, an SP3 specific to 5.1.3, and LNVYGYSK, an SP4 specific to 5.1.3.20. The relationship of these SPs to the EC hierarchy of SP families is shown in Figure 1A.

Table 1 shows that the SPs cover (i.e., appear on the sequence of) most enzymes in the dataset. The coverage columns display the cumulative coverage of all SPs to their left. Coverage is a measure of the success of the SP approach. Thus, from the sixth column one can deduce that functional classification at the third level of EC is specified by 45,819 peptides of SP3 and SP4, covering 89.8% of the data.

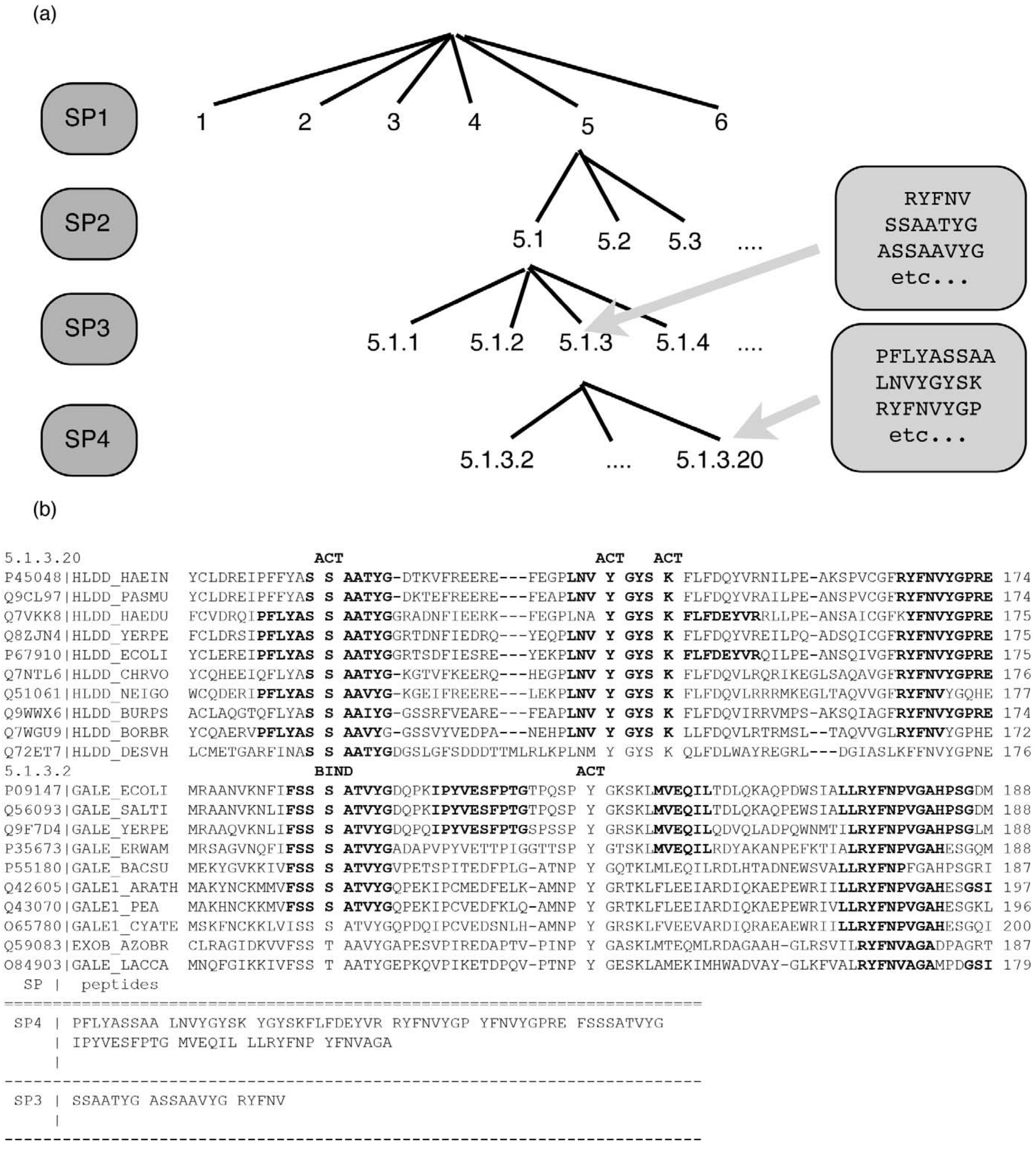
Information about the separate coverage of each SPN group is provided in Table S1. The length distribution of SPs is displayed in Figure S1 for all enzyme classes. No SP exists with a length shorter than four amino acids. The average SP length is 8.4 (s.d. 4.5). The distribution of the number of SPs occurring on enzymes is given in Figure S2. It is very flat. On average, 15.6 SPs appear on each enzyme and the standard deviation is 16. Enzyme sequences that share long SPs are highly similar, while sharing short SPs indicates smaller sequence similarity. This is displayed for short (smaller than nine amino acids) and medium length (between nine and 12 amino acids) SPs in Figures S3 and S4: most enzyme pairs that share SPs of length larger than 12 amino acids possess sequence identity of over 90%.

Prediction of Enzyme Classes

The SwissProt 48.3 dataset contains 260 enzymes that have more than one annotation, and, therefore, have been excluded from the training set (see Methods). Using them as a test set, we find 849 hits of SPs on 157 of these enzymes. 711 of the 849 hits agree with one of the given annotations and 138 do not, thus obtaining an accuracy of 84%. The results are displayed in Table S2, comparing the Swiss-Prot EC annotations with SP predictions. For example, the first protein on the list has Swiss-Prot EC annotations of 2.7.2.4 and 1.1.1.3. Its sequence matches two SPs, one SP1 of class 1 and one SP4 of 2.7.2.4. This is counted as two correct matches. An analysis of Table S2 shows that predictions based on a single SP hit may be erroneous, while those based on more than two SPs whose EC assignments are consistent with one another are correct.

We have tested the generalization quality of our SP-based enzyme classification by running MEX on the Swiss-Prot 45 release (October 2004) and testing its predictions on 10,000 novel enzymes that are listed in the Swiss-Prot 48.3 release (for the relation between these two sets see Figure S5 and Table S3). Generalization quality is assessed in Table 2 by recall (matching SPs extracted from the 45 data on novel enzymes) and precision (correctness of the "45" EC assignment according to "48.3" annotations). Precision can be defined at the SP level, i.e., to what extent did the EC of this SP match the true EC of the enzyme that it hits. Precision can also be defined at the enzyme level: how many enzymes are correctly identified by all SPs that hit them. In other words, demanding the EC assignments of all SPs to be consistent with one another as well as with the "48.3" annotation of the enzyme. Overall recall is 84%. Precision at the SP level is almost perfect, 98.7%; nonetheless, at the enzyme level it reduces to 81.7%. The reason is that usually there are many SPs hitting each enzyme, and the small error at the SP level is magnified by the requirement that the EC labels of all SPs on the same enzyme should be consistent with each other.

This generalization test suffers from bias, i.e., there exist enzymes in the test set that have high sequence similarity to some enzymes in the training sets. In conventional machine-learning analysis of sequence to function classification [2], one often tries to eliminate bias by avoiding high sequence similarity between proteins in the test set and proteins in the



```
5.1.3.20                        ACT                         ACT   ACT
P45048|HLDD_HAEIN  YCLDREIPFFYAS S AATYG-DTKVFREERE---FEGPLNV Y GYS K FLFDQYVRNILPE-AKSPVCGFRYFNVYGPRE 174
Q9CL97|HLDD_PASMU  YCLDREIPFFYAS S AATYG-DKTEFREERE---FEAPLNV Y GYS K FLFDQYVRAILPE-ANSPVCGFRYFNVYGPRE 174
Q7VKK8|HLDD_HAEDU  FCVDRQIPFLYAS S AATYGGRADNFIEERK---FEGPLNA Y GYS K FLFDEYVRRLLPE-ANSAICGFKYFNVYGPRE 175
Q8ZJN4|HLDD_YERPE  FCLDRSIPFLYAS S AATYGGRTDNFIEDRQ---YEQPLNV Y GYS K FLFDQYVREILPQ-ADSQICGFRYFNVYGPRE 175
P67910|HLDD_ECOLI  YCLEREIPFLYAS S AATYGGRTSDFIESRE---YEKPLNV Y GYS K FLFDEYVRQILPE-ANSQIVGFRYFNVYGPRE 175
Q7NTL6|HLDD_CHRVO  YCQHEEIQFLYAS S AATYG-KGTVFKEERQ---HEGPLNV Y GYS K FLFDQVLRQRIKEGLSAQAVGFRYFNVYGPRE 176
Q51061|HLDD_NEIGO  WCQDERIPFLYAS S AAVYG-KGEIFREERE---LEKPLNV Y GYS K FLFDQVLRRRMKEGLTAQVVGFRYFNVYGQHE 177
Q9WWX6|HLDD_BURPS  ACLAQGTQFLYAS S AAIYG-GSSRFVEARE---FEAPLNV Y GYS K FLFDQVIRRVMPS-AKSQIAGFRYFNVYGPRE 174
Q7WGU9|HLDD_BORBR  YCQAERVPFLYAS S AAVYG-GSSVYVEDPA---NEHPLNV Y GYS K LLFDQVLRTRMSL--TAQVVGLRYFNVYGPHE 172
Q72ET7|HLDD_DESVH  LCMETGARFINAS S AATYGDGSLGFSDDDTTMLRLKPLNM Y GYS K QLFDLWAYREGRL---DGIASLKFFNVYGPNE 176
5.1.3.2                         BIND                       ACT
P09147|GALE_ECOLI  MRAANVKNFIFSS S ATVYGDQPKIPYVESFPTGTPQSP Y GKSKLMVEQILTDLQKAQPDWSIALLRYFNPVGAHPSGDM 188
Q56093|GALE_SALTI  MRAANVKNLIFSS S ATVYGDQPKIPYVESFPTGTPQSP Y GKSKLMVEQILTDLQKAQPEWSIALLRYFNPVGAHPSGDM 188
Q9F7D4|GALE_YERPE  MRAAQVKNLIFSS S ATVYGDQPQIPYVESFPTGSPSSP Y GRSKLMVEQILQDVQLADPQWNMTILRYFNPVGAHPSGLM 188
P35673|GALE_ERWAM  MRSAGVNQFIFSS S ATVYGADAPVPYVETTPIGGTTSP Y GTSKLMVEQILRDYAKANPEFKTIALRYFNPVGAHESGQM 188
P55180|GALE_BACSU  MEKYGVKKIVFSS S ATVYGVPETSPITEDFPLG-ATNP Y GQTKLMLEQILRDLHTADNEWSVALLRYFNPFGAHPSGRI 187
Q42605|GALE1_ARATH MAKYNCKMMVFSS S ATVYGQPEKIPCMEDFELK-AMNP Y GRTKLFLEEIARDIQKAEPEWRIILLRYFNPVGAHESGSI 197
Q43070|GALE1_PEA   MAKHNCKKMVFSS S ATVYGQPEKIPCVEDFKLQ-AMNP Y GRTKLFLEEIARDIQKAEPEWRIVLLRYFNPVGAHESGKL 196
O65780|GALE1_CYATE MSKFNCKKLVISS S ATVYGQPDQIPCVEDSNLH-AMNP Y GRSKLFVEEVARDIQRAEAEWRIILLRYFNPVGAHESGQI 200
Q59083|EXOB_AZOBR  CLRAGIDKVVFSS T AAVYGAPESVPIREDAPTV-PINP Y GASKLMTEQMLRDAGAAH-GLRSVILRYFNVAGADPAGRT 187
O84903|GALE_LACCA  MNQFGIKKIVFSS T AATYGEPKQVPIKETDPQV-PTNP Y GESKLAMEKIMHWADVAY-GLKFVALRYFNVAGAMPDGSI 179
  SP |  peptides
=================================================================
 SP4 | PFLYASSAA LNVYGYSK YGYSKFLFDEYVR RYFNVYGP YFNVYGPRE FSSSATVYG
     | IPYVESFPTG MVEQIL LLRYFNP YFNVAGA
     |
-----------------------------------------------------------------
 SP3 | SSAATYG ASSAAVYG RYFNV
     |
-----------------------------------------------------------------
```

Figure 1. The Occurrence of Specific Peptides within the EC Hierarchy of Enzymes

(A) A sketch of the EC hierarchy and the assignments of SPs to SP classes. SPs can be compared with those appearing in Figure 1B.

(B) Aligned sequences of two groups of enzymes of level 4 that share the same third-level assignment. Alignment is performed according to SPs. The organisms in the upper group, 5.1.3.20, belong to proteobacteria, while those of the lower group, 5.1.3.2, also contain eukaryotes (ARATH, CYATE, and PEA). Boldfaced substrings denote SPs. Amino acids flanked by spaces denote active sites and binding sites, as indicated above. A list of all SPs and their assignments to SPN classes is presented below the sequences.

doi:10.1371/journal.pcbi.0030167.g001

training set. In our case this is problematic, because it effectively calls for eliminating from the test set all enzymes that have four-digit EC numbers appearing in the training set. Alternatively, one could produce for each enzyme in the test set a new training set that does not contain sequences with the same EC number, which is both unconventional and computationally very complex.

To overcome this predicament, we have used the following procedure: a) start with the test set consisting of all sequences of SwissProt release 48.3 that do not appear in release 45; b)

Table 1. Specific Peptides in All Six Classes of Swiss-Prot Release 48.3

EC Class	Number of Enzymes	Number of SP4	Coverage	Number of SP3	Coverage	Number of SP2	Coverage	Number of SP1	Coverage
Oxidoreductases	9,437	8,314	86.1%	681	89%	310	90.8%	1,260	93.9%
Transferases	16,196	12,708	88.4%	726	90.7%	476	91.4%	2,068	93.7%
Hydrolases	10,901	7,535	78.7%	809	83.2%	196	83.9%	1,136	87.4%
Lyases	5,229	4,728	91.4%	186	92.3%	59	92.3%	296	93.4%
Isomerases	2,887	2,588	91.5%	48	92.2%	25	92.3%	154	93.2%
Ligases	6,048	6,974	96.1%	495	97.1%	93	97.3%	500	98.2%
Total	50,698	42,874	87.3%	2,945	89.8%	1,159	90.5%	5,414	92.9%

Displayed are the name of the class, the number of enzymes within each class, the number of SPs, and their cumulative coverage of the data (thus, the sixth column displays the coverage of all SPs belonging to SP3 U SP4, the eighth column displays coverage of SP2 U SP3 U SP4, etc.).
doi:10.1371/journal.pcbi.0030167.t001

blast each one of these (test set) sequences against the sequences of the training set (SwissProt release 45) that do not have the same four-digit EC number; c) include in the non-redundant test set only sequences whose BLAST score [20] with all other training sequences (including those with the same first three EC digits) is larger than 10^{-3}; d) test generalization on this non-redundant set only for peptides in SP1, SP2, and SP3, thus avoiding the SP4 peptides that were extracted from the same fourth-level EC sequences as those of the non-redundant test set. It should be noted that removing the SP4 peptides makes the functional annotation task much more difficult because the coverage of enzymes by SPs is strongly reduced. Only 440 enzymes obey the BLAST $> 10^{-3}$ condition, and less than 40% of them carry SP1, SP2, and SP3 matches.

The results are displayed in Table 3. We obtain correct classification with an accuracy of 88%. The test is that of precision of SP assignments, i.e., to what extent do the EC labels of the SPs, observed to exist on the enzyme sequences, correspond to "48.3" EC classifications.

Whereas even the unbiased tests have high precision, we should emphasize that many successes of the SP approach are due to SP4 peptides, whose existence stems from high homology among different sequences that belong to the same EC number. These successes include the high coverage of enzymes (see Table 1) and the coverage of active and binding sites to be discussed below. The fact that these SPs have been extracted by MEX may be viewed as the essence of homology, as illustrated in Figure 1B, where the existence of SPs is displayed on various enzymes aligned according to their matching SPs.

We provide a Web tool, available at http://adios.tau.ac.il/SPMatch, which displays the hits of SPs for any protein sequence that is presented as a query, together with the EC assignments due to these SPs.

Comparison with Other Methods

We have tested the usefulness of the SP approach by comparing it with conventional functional prediction methods. For this purpose we have used all oxidoreductases in the 48.3 data and divided them into training data and test data with a 75%:25% ratio. MEX was run on all data and SPs were selected from the MEX motifs according to the training data. Only this subset of motifs was then employed to classify the test data. This procedure has been repeated 45 times to gain statistics, and has been subjected to a support vector machine (SVM) analysis. It has been compared with a state-of-the-art method [17] based on an analogous SVM procedure, applied to the same data using the same divisions and relying on classification of (train and test) data according to a matrix of Smith-Waterman distances from all oxidoreductases. The results are displayed in Tables S4 and S5 and show a clear advantage to SP classification. For comparison, we use the Jaccard score defined as $J = TP / (TP + FP + FN)$ where TP, FP, and FN denote true positives, false positives, and false negatives, accordingly. Whereas sequence similarity leads to an average Jaccard score of 0.86 on the second EC level and

Table 2. Performance of SPs Extracted from the Swiss-Prot 45 Dataset on Novel Enzyme Sequences in Swiss-Prot 48.3

EC Class	Number of Sequences	Recall	Precision (SP)	Precision (Enzyme)
Oxidoreductases	1,661	74.4%	98.9%	78.2%
Transferases	3,722	87.3%	98.9%	84.6%
Hydrolases	2,173	74.3%	97.8%	71.8%
Lyases	1,089	85.4%	99.4%	91.2%
Isomerases	541	88.0%	93.8%	79.0%
Ligases	1,399	99.0%	99.3%	87.1%
Total	10,585	84.1%	98.7%	81.7%

Recall refers to the coverage by "45" SPs, and precision is the accuracy measured by "48.3" EC classification. Precision is quoted both at the SP level, where the EC of each SP hit is compared with that of the enzyme, and at the enzyme level when all EC assignments of all SP hits are required to be consistent with that of the enzyme.
doi:10.1371/journal.pcbi.0030167.t002

Table 3. Coverage of a Non-Redundant Test Set by Motifs in SP1, SP2, and SP3

Class	Number of Sequences	SP1	tp_1	fp_1	SP2	tp_2	fp_2	SP3	tp_3	fp_3
Oxidoreductases	36	15 (35)	34	1	0	0	0	0	0	0
Transferases	15	7 (13)	13	0	2 (2)	2	0	2 (2)	2	0
Hydrolases	98	30 (41)	39	2	5 (5)	4	1	4 (4)	2	2
Lyases	134	22 (23)	23	0	10 (12)	11	1	13 (18)	18	0
Isomerases	147	38 (42)	26	16	6 (6)	6	0	9 (14)	8	6
Ligases	10	3 (5)	5	0	4 (10)	10	0	0	0	0
Total	440	115 (159)	140	19	27 (35)	33	2	28 (38)	30	8

Numbers in the three SPN columns indicate the number of sequences that have been covered by SPs. Numbers in parentheses indicate the numbers of SPs observed to occur on the sequences. Columns of tp$_N$ and fp$_N$ display true-positive and false-positive predictions of SPN peptides, where tp corresponds to the SP indicating correctly the EC classification and fp otherwise.

doi:10.1371/journal.pcbi.0030167.t003

0.82 on the third level, SP classification has average Jaccard scores of 0.93 and 0.92, accordingly. Comparing with yet another method, SVM-Prot [18,19], which classifies enzymes on the basis of physical and chemical features of their amino acids, we note that the latter achieves a Jaccard score of only 0.74 on all oxidoreductases data at the second EC level.

The common lore, that large sequence identity between two proteins implies that the two have the same function, has its exceptions. Motifs, although often extracted from homology, may serve as better measures for functional specification of proteins [21] than overall sequence similarity. Table 4 demonstrates this point, by contrasting SP predictions with Smith-Waterman similarity results for pairs of enzymes. These extreme cases have been posed as a problem by Ross [5] (see Table 1 there). All displayed EC assignments correspond to those of SPs located on the enzyme sequences, and match the correct EC numbers. As a more detailed example, we point out that the enzymes of the sixth pair in Table 4, GTFB_STRMU and AMY3B_ORYSA, have 42% sequence identity along an alignment of 105 amino acids. Nonetheless, the sequences are not identical at the SP locations. AMY3B_ORYSA contains 24 SPs, none of which have an exact match on GTFB_STRMU, and a single SP4 (GGAFLE) found on the latter matches correctly its EC number.

It is of interest to compare our SPs with ProSite motifs [8], which are listed in the Swiss-Prot database as standard motif annotations on 63% of the enzymes. ProSite motifs are either regular expressions (of average length 18.3 amino acids) or weight matrices, while SPs are deterministic motifs (with average length of 8.4). We search for all appearances of ProSite regular expression motifs on enzymes. Each such appearance is noted on the enzyme sequence and checked whether it is also (partially) covered by an SP. Figure S6 compares the appearance of SPs and ProSite motifs on the data, and Figure S7 displays the relative coverage of ProSite motifs by SPs as function of the minimal percentage of amino acids belonging to the ProSite motif that are also located on SPs. Thus we find that if at least 40% of the amino acids of the ProSite motif also belong to SPs, which would be appropriate for an average SP to be located within an average ProSite motif, then SPs cover 48% of all ProSite motif occurrences. This may be compared with a random model (see Methods) which covers on average only 24% of ProSite motif occurrences, with a standard deviation of 0.06%. This extremely significant result (400 s.d.) demonstrates that SPs carry information that is highly correlated with that of ProSite motifs.

Biological Roles of Specific Peptides

Coverage of active sites. Next we turn to establishing some particular biological roles for SPs. First we investigate their coverage of active and binding sites. 42% of all enzymes in the Swiss Prot 48.3 database have annotations of loci of active sites and binding sites (single amino acids). For simplicity we will refer to both annotations as active sites. A few examples are shown in Figure 1B. Given these loci, we find that 65% of all active sites are covered by SPs. This can be compared with

Table 4. Enzymes with High Sequence Similarity and Different EC Assignments

Enzyme 1	Enzyme 2	Sequence Identity	Alignment Length	e-Value
GUNA_PSEFL EC 3.2.1.4	MDHP_FLABI EC 1.1.1.82	71%	28	1.6 e-03
PLB1_YEAST EC 3.1.1.5	METB_ARATH EC 2.5.1.48	60%	30	5.9 e-05
RPB1_PLAFD EC 2.7.7.6	UBC2_YEAST EC 6.3.2.19	63%	27	1.8 e-05
CHIB_POPTR EC 3.2.1.14	DGK2_DROME EC 2.7.1.107	58%	24	6.0 e-06
ODO2_FUGRU EC2.3.1.61	PP2BB_HUMAN EC 3.1.3.16	53%	39	1.1 e-06
GTFB_STRMU EC 2.4.1.5	AMY3B_ORYSA EC 3.2.1.1	42%	105	7.4 e-08
RPB1_PLAFD EC 2.7.7.6	PDE3B_RAT EC 3.1.4.17	58%	36	8.4 e-08
IGF1R_HUMAN EC2.7.10.1	PTPRU_HUMAN EC 3.1.3.48	34%	157	1.5 e-09

Alignment and identity are calculated according to the Smith-Waterman method [33]. EC assignments are determined by SPs and are correct.

doi:10.1371/journal.pcbi.0030167.t004

Table 5. Occurrence of Specific Peptides on Active Sites

Dataset	Number of Enzymes	Active Sites Hit by SPs	Random Sites Hit by SPs	Number of Standard Deviations	Number of SPs	SPs Hitting Sites
All	21,228	65%	27%	80	26,931	8%
Non-redundant	582	52%	21%	33	6,660	12%

Analysis has been carried out on enzymes that have an active (or binding) site annotation and are being hit by SPs. Results are given for the total set of such enzymes in Swiss-Prot 48.3 and for a non-redundant set in which a single enzyme was chosen for each EC number. Statistical significance of these results is given in terms of standard deviations (see Methods). The *p*-values are well below 10^{-308}.

doi:10.1371/journal.pcbi.0030167.t005

the coverage of random positions on the same enzyme sequences which, on average, is only 27% (off by 80 standard deviations, see Methods). We also construct a non-redundant set by choosing only one enzyme for each EC number (i.e., EC class of level 4). The results, displayed in Table 5, show some differences between the total and the non-redundant sets. Since the latter is unbiased, it should generalize better, and allow us to get a better estimate of active-site coverage had the annotations existed for all enzymes. This estimate is 12% and has very high statistical significance (zero *p*-value, see Methods).

As an example of these features in the data, we display in Figure 1B aligned subsequences of enzymes, belonging to the same third level but to two different fourth levels of the EC hierarchy: six out of 35 enzymes of 5.1.3.2 and seven out of 29 enzymes of 5.1.3.20. Shown are strings belonging to the sequences that include active sites and binding sites as indicated in Swiss-Prot annotations, and boldfaced substrings denoting SPs from our lists. Whereas in 5.1.3.20, most active sites are covered by SPs, this is not the case for the active site of 5.1.3.2. Nonetheless, it turns out from investigating spatial structures of these enzymes that RYFNV, an SP that appears in both groups, is located within the same pocket in which the active site resides. This may be regarded as an indication that RYFNV plays an important role in fostering the biological function of this enzyme.

An example stressing the relationships among SPs and spatial structures is presented in Figure 2. This enzyme contains many SPs. Two SPs cover the active site, one—HMVRNI—shares a pocket with the active site and the two binding sites, and another one—FHARFV—plays the role of RNA binding in this tRNA pseudouridine synthase I.

FHARF is one example of previously discovered motifs [22]. Some other examples are: a) GFGRIG (SP of 1.1.1.26) [23], a conserved region of GAPDH that is active in the glycolytic pathway; b) HRDLKP (SP of 2.7.1.37) [24], appearing in protein kinases; c) IFIDEID (SP of 3.6.4.3), the Walker B motif of ATPase [25]; to name a few. However, most of the SPs have not been studied before.

These results raise the question how many SPs can be found in the neighborhood of active sites, as defined by the pockets in the spatial structures of enzymes. One is naturally tempted to assign importance to all SPs of this kind, not just those that carry the active site annotation (single amino acid). For this study we use the CASTp [26] database, which lists all amino acids belonging to pockets appearing in spatial structures of proteins. We select 1,031 enzymes that possess pockets including active (or binding) site annotations. There are 8,860 SPs that occur on these enzymes, 31% of which lie within these "active pockets," i.e., have at least four amino acids that reside in the pocket. Defining a background model (see Methods) of random peptides selected for each event of an SP hitting an active pocket in a particular enzyme, we estimate that 11% of all SPs belong to events that pass an FDR limit [27] of 0.05. Most of them (70%) do not contain an active site; hence, they are of potential interest for experimental verification of their importance in defining and maintaining the enzymatic function. Table 6 summarizes the results of this analysis. Further details of all significant events are presented in Table S6. All 1,910 listed SP occurrences on enzymes should be of high relevance to the biological functions of these enzymes, and the elimination of any one of them from the enzyme sequence on which it occurs should be deleterious to the function of that enzyme or to its stability.

SPs may also have biological roles that are not connected to active or binding sites. Examples are DNA and RNA binding, metal binding, protein–protein interactions, etc. Given the large number of SPs, we may look forward to a plethora of predictions.

Minimal SP Sets with Maximal Coverage

We started our study with 50,698 enzymes from which 52,365 SPs were extracted. These SPs provided coverage of about 93% of all enzymes. By introducing further screening of SPs according to biological findings, a much reduced number of SPs may suffice for the purpose of classification. 21,228 enzymes carry active or binding site annotations in the 48.3 data. The number of SPs hitting these enzymes is 26,931; however, only 2,337 cover the active or binding sites. These 2,337 are found to occur on 79% of the 21,228 enzymes. Thus, instead of the approximately 1:1 ratio between the number of SPs and the number of enzymes they cover as found previously, we now obtain an order of magnitude parsimonious ratio, of about 1:8, while maintaining a similar level of classification accuracy.

The same SPs cover 36% of all original enzymes of our dataset. Performing a similar analysis on the 45 data, one finds that the 2,014 SPs that cover the annotated enzymes in it hit 75% of the relevant set of enzymes. Moreover, using the same SPs to classify the 10,585 novel enzymes contained in the 48.3 release and absent from the 45 release, one obtains coverage of 28% of them. This last fact demonstrates that the relatively large coverage reached by the small fraction of SPs that hit active sites is not limited to the dataset (training set) used to define the SPs. All these results are summarized in Table 7. It seems therefore quite reasonable to conclude that,

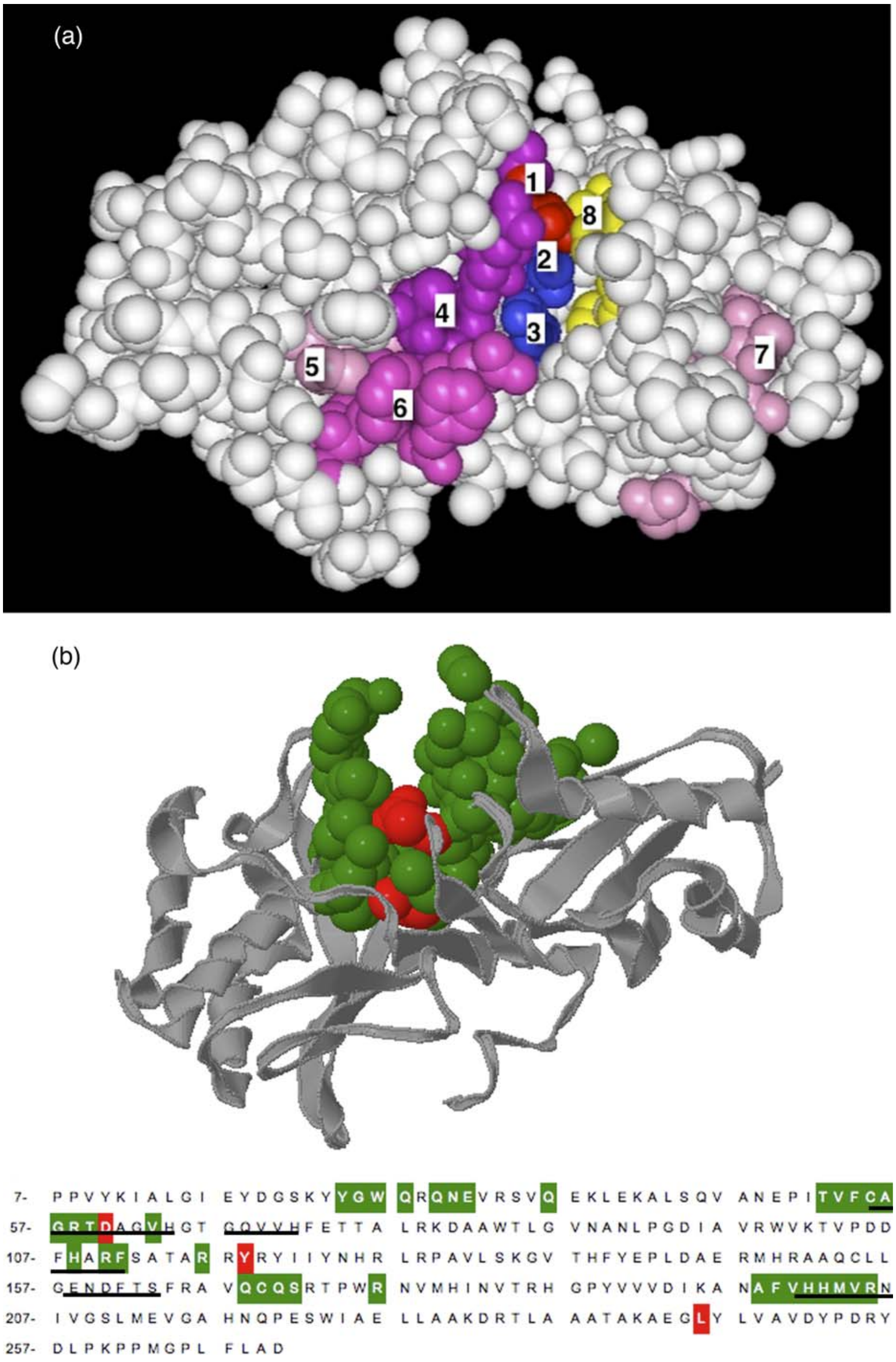

(a)
1
8
2
4
3
5
7
6
(b)
7- PPVYKIALGI EYDGSKYYGW QRQNEVRSVQ EKLEKALSQV ANEPITVFCA
57- GRTDAGVHGT GQVVHFETTA LRKDAAWTLG VNANLPGDIA VRWVKTVPDD
107- FHARFSATAR RYRYIIYNHR LRPAVLSKGV THFYEPLDAE RMHRAAQCLL
157- GENDFTSFRA VQCQSRTPWR NVMHINVTRH GPYYVVVDIKA NAFVHHMVRN
207- IVGSLMEVGA HNQPESWIAE LLAAKDRTLA AATAKAEGLY LVAVDYPDRY
257- DLPKPPMGPL FLAD

Figure 2. SPs Occurrence on a Spatial Structure of an Enzyme

(A) 3-D display of enzyme P07649 (PDB code 1DJ0), belonging to 5.4.99.12, showing (1) an active site D at sequence location 60; (2) a binding site Y at location 118; (3) a binding site L at location 245. The active site is common to two SPs (4) containing (CAGRT(D)AGVH). Other shown SPs are (5) GQVVH at locations 67–71; (6) FHARF at 107–111, known to be a tentative RNA-binding peptide; (7) ENDFTS at 157–163; and (8) HMVRNI at 201–207, sharing a pocket with the active and binding sites. QVVH and ENDFTS belong to SP3, all other peptides belong to SP4.

(B) A different display of the same enzyme focuses on the pocket containing the active site. The relevant section of the sequence is shown, with red residues signifying active and binding sites, green residues corresponding to other amino acids residing in the pocket, and underlined residues corresponding to SPs.

doi:10.1371/journal.pcbi.0030167.g002

adding information of biological markers, one can reduce the ratio of the number of SPs deduced from a certain number of enzymes and needed to label their EC classification from 1:1 to about 1:8.

This, however, does not mean that all other SPs should be disregarded. First, there exist good chances that they are of biological importance for various structural and functional reasons that may warrant further investigation. Second, when extreme classification issues come up, as in the cases displayed in Table 4, every single SP may count.

Table 6. Occurrence of SPs in Spatial Proximity to Active Sites

Number of Enzymes	Number of SPs	Number of SPs in Pockets	Significant SPs FDR = 0.05	Significant SPs without Sites
1,031	8,860	2,487 (28%)	1,622 (18%)	1,422 (16%)

This is an analysis of 1,031 enzymes whose spatial structure is known (in PDB) and possesses 3-D pockets that include active site (using CASTp [26]). This table lists the number of enzymes that were analyzed and the number of SPs that are located on these enzymes. This is followed by numbers of SPs lying (with at least four residues) in pockets including active sites. Requiring high significance of the latter, through a background model described in Methods, and using the FDR limit of 0.05, we obtain the results displayed in the following column. The last column displays the number of significant SPs that lie in the pocket but do not contain the amino acid with active site annotation.
doi:10.1371/journal.pcbi.0030167.t006

Discussion

Conventional wisdom attributes protein functions to large domains, as well as to specific amino acids at strategic structural points on the protein. Large-scale studies often make use of multiple sequence alignment (MSA), phylogenetic information, and sophisticated mathematical models, thus leading to the plethora of algorithms and Web tools that permeate bioinformatics. While all that may be necessary to obtain a thorough understanding of the way proteins develop and perform, much can be gained by shifting attention to deterministic linear motifs on proteins. In doing so, we return to a way that has been often tried in the past. Thus, in the 1990s, many investigations looked for k-mers that are over-represented in sequences of proteins that have common functional properties. Some examples are ProSite [8,12], with which we have compared our results, and papers such as [28–30], where major emphasis has been put on finding a complete dictionary of motifs that cover all strings of amino acids that are of any importance. In the case of [30], the search has been an unsupervised one leading eventually to a coverage of 98% of all amino acids on the protein strings. Some reviews of the motif approaches of the 1990s are [7,9]. More recently, interests have shifted to automated prediction tools that may make use of motifs but are not limited to them. Examples are the GOtcha method [31] that uses sequence-identity searches of various genomes to predict functional annotation, and [32] who pursue the same goal using PSI-BLAST searches with varying resolution.

Our goal is more moderate, restricting ourselves to the functional classification of enzymes. By doing so, and by applying the MEX algorithm together with limiting ourselves to SPs within the EC hierarchy, we are able to classify all enzymes by SPs occurring on them with coverage between 87% to 93%, depending on the EC level that is being looked for (Table 1). Classification success of novel sequences that belong to the same type of data has coverage of 84% and precision of 99% at the SP level and 82% at the enzyme level (Table 2). Restricting ourselves to low bias (Table 3), we still have a large precision of 88% at the SP level. We have demonstrated that our results surpass the classification accuracy of sequence similarity (using Smith-Waterman [33]), and our SPs have a higher coverage than ProSite motifs. As such, they become a powerful tool that may be added to existing automated searches.

It should be noted that the SPs were extracted by an unsupervised motif search algorithm, applied to each one of the six EC classes. This is quite different from conventional supervised approaches. Our method may disregard motifs that obey some over-representation criterion, and choose others that do not satisfy such a global statistics measure. Another major difference from other approaches is that we do not make use of MSA. MEX finds significant motifs without requiring alignment as a preprocessing stage. In fact, MEX can serve as a source for MSA by employing its motifs for alignment (see Figure 1B).

SPs were selected from all MEX motifs by imposing the condition that they should be specific to particular levels of the EC hierarchy. This has led to a large number of SPs, as numerous as the set of all enzymes (but, obviously, providing a much more concise description). Imposing further biological conditions, one may find much smaller sets that suffice for classification. In an analysis of enzymes for which the active sites are known, we have shown that the set of SPs bearing these active sites, which comprises just 8.6% of all relevant SPs (i.e., those occurring anywhere on these enzymes), suffices to cover (and therefore label) all enzymes.

Conventional classification methods rely on homology. While large homology is also at the root of our success for most SPs of level 4 (see some examples in Figure 1B), we have

Table 7. Small Sets of SPs that Contain Active Sites Suffice To Specify Functionality of Many Enzymes

Dataset	Number of Enzymes	Number of SPs	Number of SPs Hitting Active Sites	Percent Enzyme Coverage	Percent SPs
Annotated enzymes	21,228	26,931	2,337	79%	8.6%
All “48.3” enzymes	50,698	52,365	2,337	36%	4.5%
Annotated “45” enzymes	17,005	21,676	2,014	75%	9.3%
Enzymes in “48.3–45”	10,585			28%	

The first two rows refer to the Swiss-Prot 48.3 data, the third to release 45 data, and the fourth to the novel enzymes, as defined in Table S5. The columns specify the numbers of enzymes, the numbers of SPs that appear on them, the number of SPs containing annotated active (or binding) sites, the coverage that this limited set of SPs provides, and the fraction it consists of the total number of SPs. Some of these entries are irrelevant to the last row, which is used to test generalization, i.e., to see if the coverage by the 2,014 SPs of the third row (28%) is similar to that displayed in the second row (36%).
doi:10.1371/journal.pcbi.0030167.t007

demonstrated (in Table 4) that SPs can also be of importance in extreme cases, where straightforward comparison of an enzyme to another one with large sequence similarity may be misleading.

In conclusion, we have established a comprehensive and accurate classification scheme for enzymes based on the occurrence of short peptides on their sequences. The SPs contain, on average, just 8.4 amino acids, yet they suffice to correctly classify an overwhelming majority of known enzymes. Moreover, we have found indications for some of the biological roles of SPs, e.g., covering a majority of active sites. This study has laid the foundations for the further experimental investigation of these intriguing sets of SPs.

Methods

Motif extraction. MEX is a motif extraction algorithm that serves as the basic unit of ADIOS [15], an unsupervised method for extraction of syntax from linguistic corpora. We apply it to the problem of finding sequence motifs in enzymes.

Each enzyme sequence is represented as a path over a graph containing 20 vertices, each vertex representing one amino acid. After uploading all enzyme sequences onto the graph, one counts the number of paths connecting vertices in order to define probabilities such as

$p(e_j|e_i) =$ (number of paths proceeding from e_i to e_j) / (total number of paths leaving e_i)

$p(e_k|e_j,e_i) =$ (number of paths proceeding from e_i to e_j to e_k) / (number of paths proceeding from e_i to e_j)

for all vertices e_i of the graph. These data-driven probabilities allow for the definition of a position-dependent variable-order Markov model describing the data.

A motif that is extracted by MEX is a subpath along the graph defined by probability-based criteria that account for convergence of many paths into the beginning point of a motif, and divergence of many paths from the endpoint of the motif. Motifs are not constrained by length, and may overlap with one another (see, e.g., the two SPs that overlap at the active site D in Figure 2B). The only two parameters of MEX are η, specifying a decrease in probability measures that determine convergence and divergence, and α specifying their statistical significance. For more details, see [15] and http://adios.tau.ac.il. Throughout this paper, we use $\eta = 0.9$ and $\alpha = 0.01$.

Data. Protein sequences annotated with EC numbers were extracted from the Swiss-Prot database (Release 48.3, 25 October 2005). To obtain a high-quality, well-defined training dataset, the data were strictly screened and the following sequences were removed: sequences shorter than 100 amino acids or longer than 1,200 amino acids, sequences with uncertain annotation, and enzymes that catalyze more than one reaction (e.g., have more than one EC number).

Random model for SP hits on ProSite motifs. Enzyme sequences are searched for matches with regular expressions of ProSite motifs. The resulting strings of amino acids are checked for matches with SPs. The latter are compared with matches of a random model where, for each given enzyme, random peptides are selected with the same lengths as those of the SPs that hit this enzyme. The random model provides a probability distribution which serves as a zero model for calculating the significance of the SP hit on the ProSite motif. This comparison is being made for each enzyme and for varying fractions of amino acids that are shared by the SP with the ProSite motif.

Significance of SP hits on active sites. In analyzing the significance of SP coverage of active (and binding) sites, we compare this coverage with that of randomly chosen residues on enzyme sequences. This is carried out on all data (i.e., annotated enzymes with SP hits) and on a non-redundant set composed of only one enzyme from each EC number (i.e., EC classification at level 4). The deviations of the measurements from random distributions are very high, and are quoted in numbers of standard deviations. The corresponding p-values are zero according to Matlab accuracy, i.e., are well bellow 10^{-308}.

Significance of SP residing in active pockets. Let us define an event as the occurrence of a given SP within an active pocket in a given enzyme. For each such event, we evaluate the probability that at least one of randomly selected sequences from this enzyme, which coincide in length with the various SPs that occur on this enzyme, lies (with at least four amino acids) within the active pocket. This defines the p-value that we assign to the event. We then select the significant events according to an FDR limit [33] of 0.05.

Supporting Information

Figure S1. SP Length Distribution

Found at doi:10.1371/journal.pcbi.0030167.sg001 (377 KB JPG).

Figure S2. Distribution of the Numbers of SPs Occurring on Enzymes

Found at doi:10.1371/journal.pcbi.0030167.sg002 (500 KB JPG).

Figure S3. Distribution of Percentages of Sequence Identity for Pairs of Enzymes Sharing the Same SP3 or SP4 of Length Less Than Nine Amino Acids

Found at doi:10.1371/journal.pcbi.0030167.sg003 (172 KB JPG).

Figure S4. Distribution of Percentages of Sequence Identity for Sets of Enzymes That Share the Same SP3 or SP4 of Length between 9 and 12

Found at doi:10.1371/journal.pcbi.0030167.sg004 (156 KB JPG).

Figure S5. Relation of Enzymes in Two Swiss-Prot Releases, 45 (October 2004) and 48.3 (October 2005)

Found at doi:10.1371/journal.pcbi.0030167.sg005 (118 KB JPG).

Figure S6. Data Coverage by ProSite Regular Expression Motifs and by SPs in the Swiss-Prot Database

Found at doi:10.1371/journal.pcbi.0030167.sg006 (183 KB JPG).

Figure S7. Coverage of ProSite Motifs by SPs versus the Required Minimal Amount of Amino Acids Shared by the Two Motifs

For each ProSite motif (of average length 18 amino acids) occurrence on an annotated enzyme, SP matches were searched. The cumulative percentage of ProSite motifs that are covered by SPs is plotted as a function of the relative amount of coverage, i.e., the percent of the number of amino acids belonging to the ProSite motif that is shared by the SP. This is compared with the coverage of ProSite motifs by random motifs that have the same length and number as the SPs appearing on the enzymes

Found at doi:10.1371/journal.pcbi.0030167.sg007 (360 KB JPG).

Table S1. Coverage by SPs of Enzymes in Swiss-Prot Release 48.3

Found at doi:10.1371/journal.pcbi.0030167.st001 (30 KB DOC).

Table S2. Comparison between Swiss-Prot Annotations and SP Predictions for Doubly Annotated Enzymes

Columns indicate the protein ID according to Swiss-Prot, its two EC assignments, the EC assignments according to SP predictions, and the number of SP matches that have the same EC prediction (separated into correct and false predictions). An analysis of the data shows that predictions that are based on a single SP match in the enzyme sequence are often wrong (122 false predictions versus 80 true predictions). The appearance of two SPs whose EC assignments are consistent with each other leads to 19 true predictions and five false predictions. All predictions based on more than two consistent SPs are true. When counting enzymes (rather than SPs), we find that 92 of 157 had one false prediction and no true prediction. 48 enzymes have one false prediction; 31 of them have also one true prediction, and 17 have two true predictions. 65 enzymes have no false prediction; 43 of them have one true prediction and 22 have two true predictions. It should be noted that this list of enzymes contains many related enzymes (i.e., it has high bias), hence successes and failures in different enzymes are correlated. It seems safe, however, to conclude that predictions based on several SPs whose EC assignments are consistent with each other may be trusted.

Found at doi:10.1371/journal.pcbi.0030167.st002 (808 KB DOC).

Table S3. Numbers of Enzymes in Swiss-Prot Release 48.3 and Swiss-Prot Release 45

Found at doi:10.1371/journal.pcbi.0030167.st003 (34 KB DOC).

Table S4. Comparison of SP with Smith-Waterman Performance on Classification at the Subclass Level

Classification based on SPs has been compared with classification based on sequence similarity using the Smith-Waterman (SW)

method. This has been performed on the oxidoreductases data of the 48.3 release, using all subclasses and sub-subclasses that contain more than 20 enzyme sequences. The data were randomly partitioned into 75% training and 25% test sets. Features for the SP classification were determined by running MEX on all oxidoreductases and checking for their specificity using the training data only. These SPs were then used for defining, through training, the SP-SVM. Smith-Waterman analysis was carried out by defining a log(p-value) (with cutoff at p = e–06) distance matrix whose columns (features) were all oxidoreductases. The rows (instances) of the training-set enzymes were used to determine the SW-SVM classifications. 45 different partitions were performed to accumulate statistics. Same partitions were applied to both classification methods. Classification was performed using a soft-margin linear SVM, available online at http://svmlight.joachims.org.
Performance was measured by the Jaccard score J = TP / (TP + FP + FN).

Found at doi:10.1371/journal.pcbi.0030167.st004 (58 KB DOC).

Table S5. Comparison of SP with Smith-Waterman Performance on Classification at the Sub-Subclass Level

Found at doi:10.1371/journal.pcbi.0030167.st005 (99 KB DOC).

Table S6. List of SPs That Lie in Active Pockets

A list of all events of SPs lying in active pockets that have passed the FDR = 0.05 limit, ordered according to their p-values. Entries include the enzyme PDB ID and the details of the SP.

Found at doi:10.1371/journal.pcbi.0030167.st006 (2.1 MB DOC).

Acknowledgments

We thank Nir Ben-Tal, Assaf Gottlieb, Rachel Kolodny, Martin Kupiec, Ruth Nussinov, Yanay Ofran, Tal Pupko, Burkhard Rost, Roded Sharan, and Roy Varshavsky for comments and helpful discussions. We thank Joe Dundas and the CASTp team for making some of their data available to us. BS is supported by the Yeshaya Horowitz Association through the Center of Complexity Science.

Author contributions. ER and DH conceived and designed the computational work. VK and ZS performed the computational work. VK, YM, and DH analyzed the data. ZS, BS, and UW contributed analysis tools. VK, YM, ER, and DH wrote the paper.

Funding. This research was partially supported by the US–Israel Binational Science Foundation.

Competing interests. The authors have declared that no competing interests exist.

References

1. Domingues FS, Lengauer T (2003) Protein function from sequence and structure data. Appl Bioinformatics 2: 3–12.
2. Rost B, Yachdav G, Liu J (2004) The predictprotein server. Nucleic Acids Res 32: W321–W326.
3. Tian W, Skolnick J (2003) How well is enzyme function conserved as a function of pairwise sequence identity? J Mol Biol 333: 863–882.
4. Hegyi H, Gerstein M (1999) The relationship between protein structure and function: A comprehensive survey with application to the yeast genome. J Mol Biol 288: 147–164.
5. Rost B (2002) Enzyme function less conserved than anticipated. J Mol Biol 318: 595–608.
6. von Grotthuss M, Plewczynski D, Ginalsky K, Rychlewski L, Shakhnovich EI (2006) PDB-UF: Database of predicted enzymatic functions for unannotated protein structures from structural genomics. BMC Bioinformatics 7: 53–62.
7. Bork P, Koonin EV (1996) Protein sequence motifs. Curr Op Struct Biol 6: 366–376.
8. Bairoch A, Bucher P, Hofmann K (1997). Prosite. Nucleic Acids Res 25: 217–221.
9. Aitken A (1999) Protein consensus sequence motifs. Mol Biotechnol 12: 241–253.
10. Neville-Manning CG, Wu TD, Brutlag DL (1998) Highly specific protein sequence motifs for genome analysis. Proc Natl Acad Sci U S A 95: 5865–5871.
11. Huang JY, Brutlag DL (2001) The emotif database. Nucleic Acids Res 29: 202–204.
12. Falquet L, Pagni M, Bucher P, Hulo N, Sigrist CJ, et al. (2002) The ProSite database, its status in 2002. Nucleic Acids Res 30: 235–238.
13. Tong AH, Drees B, Nardelli G, Bader GD, Brannetti B, et al. (2002) A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. Science 295: 321–324.
14. Obenauer JC, Yaffe MB (2004) Computational prediction of protein–protein interactions. Methods Mol Biol 261: 445–468.
15. Solan Z, Horn D, Ruppin E, Edelman S (2005) Unsupervised learning of natural languages. Proc Natl Acad Sci U S A 102: 11629–11634.
16. Ben-Hur A, Brutlag D (2006) Protein sequence motifs: Highly predictive features of protein function. In: Guyon I, Gunn S, Nikravesh M, Zadeh L, editors. Feature extraction, foundations and applications. Berlin: Springer Verlag.
17. Liao L, Noble WS (2003) Combining pairwise sequence analysis and support vector machines for detecting remote protein evolutionary and structural relationships. J Comp Biol 10: 857–868.
18. Cai CZ, Han LY, Ji ZL, Chen YZ (2003) SVM-PROT: Web-based support vector machine software for functional classification of a protein from its primary sequence. Nucleic Acids Res 31: 3692–3697.
19. Cai CZ, Han LY, Ji ZL, Chen YZ (2004) Enzyme family classification by support vector machines. Proteins 55: 66–76.
20. Altschul SF, Madden TL, Schaffer AA, Zhan JZ, Lipman DJ (1997) Gapped blast and psi-blst: A new generation of protein database search programs. Nucleic Acids Res 25: 3389–3402.
21. Ben-Hur A, Brutlag D (2003) Remote homology detection: A motif based approach. Bioinformatics 19 (Supplement 1): i26–33.
22. Foster PG, Huang L, Santi DV, Stroud RM (2000) The structural basis for trna recognition and pseudouridine formation by pseudouridine synthase I. Nat Struct Biol 7: 23–27.
23. Anda P, Gebbia JA, Backenson PB, Coleman JL, Benach JL (1996) A glyceraldehyde-3-phosphate dehydrogenase homolog in *Borrelia burgdorferi* and *Borrelia hermsii.* Infect Immun 64: 262–268.
24. Hanks SK, Quinn AM, Hunter T (1988) The protein kinase family: Conserved features and deduced phylogeny of the catalytic domains. Science 241: 42–52.
25. Walker JE, Saraste M, Runswick MJ, Gay NJ (1982) Distantly related sequences in the alpha- and beta-subunits of atp synthase, myosin, kinases and other atp-requiring enzymes and a common nucleotide binding fold. EMBO J 1: 945–951.
26. Binkowski TA, Naghibzadeg S, Liang J (2003) Castp: Computed atlas of surface topography of proteins. Nucleic Acid Res 31: 3352–3355.
27. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. J Roy Stat Soc 57: 289–300.
28. Ogiwara A, Uchiyama I, Seto Y, Kanehisa M (1992) Construction of a dictionary of sequence motifs that characterize groups of related proteins. Protein Eng 5: 479–488.
29. Wang JTL, Marr TG, Shasha D, Shapiro BA, Chirn GW (1994) Discovering active motifs in sets of related protein sequences and using them for classification. Nucleic Acids Res 14: 2769–2775.
30. Rigoutsos I, Floratos A, Ouzounis C, Gao Y, Parida L (1999) Dictionary building via unsupervised hierarchical motif discovery in the sequence space of natural proteins. Proteins 37: 264–277.
31. Martin DM, Berriman M, Barton GJ (2004) GOtcha: A new method for prediction of protein function assessed by the annotation of seven genomes. BMC Bioinformatics 5: 178.
32. Hawkins T, Luban S, Kihara D (2006) Enhanced automated function prediction using distantly related sequences and contextual association by PFP. Protein Sci 15: 1550–1556.
33. Smith T, Waterman M (1981) Identification of common molecular subsequences. J Mol Biol 147: 195–197.