PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Biological roles of specific peptides in enzymes

Yasmine Meroz and David Horn*

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

ABSTRACT

It has recently been shown (Kunik et al., PLOS Comput Biol 2007;3(8):e167) that the occurrence of specific peptides (SPs) on sequences of enzymes allows for accurate EC classification of enzymes. We inquire whether these SPs play important roles in bringing about the enzymatic function. This is assessed by cross-checking the occurrence of SPs on enzymes with Swiss-Prot annotations and PDB spatial structures of enzymes. Analyzing the coverage of functional annotations of enzymes, we demonstrate that SPs contain major fractions of all annotated features. This result is statistically highly significant and associates over 10% of all SPs with important biological markers. Concentrating on DNA binding regions, relevant to LexA repressor enzymes, we find interesting coverage patterns. Moreover, for the same data, we demonstrate that SPs allow for subclassification of the relevant bacteria into phylogenetic classes. An analysis of mutagen annotations on SPs appearing on all enzymes leads to the conclusion that mutations on SPs tend to damage the enzymatic function much more than expected from a background model, hence SPs are of high importance to enzymatic functions. SPs that lie in 3D pockets that are shared by active and binding sites, are shown to be significantly enriched by glycine, leading to the hypothesis that they are responsible for conformational plasticity. Finally we show that SPs can partially resolve outstanding difficult problems of convergent evolution by representing correctly enzyme functions in spite of remote homologies in sequence and in structure.

Proteins 2008; 72:606–612.

Key words: specific peptides; enzymes; remote homology; conformational plasticity; mutagenesis.

INTRODUCTION

Conventional methods for protein classification utilize sequence similarity to proteins whose functions or structures are well known. Such homology-based annotation transfer may be problematic[1] and alternative methods are needed. A recent study[2] has shown that deterministic linear motifs may turn out to be a very useful tool for this purpose. Focusing on functional classification of all enzymes, they employed a novel unsupervised motif extraction algorithm MEX,[3] looking for sequences of amino-acids that obey context-dependent statistical conditions. Kunik *et al.*[2] have proceeded to screen the MEX motifs by using the enzyme commision (EC) four-level hierarchy, picking out those that are specific to particular branches (sets of indices) of the EC. The latter were named specific peptides (SPs). It turned out that they provide classification of enzymes according to the EC hierarchy with coverage of 93%. The SPs are strings of, on average, 8 ± 5 amino-acids. Although only enzyme sequences were used in the analysis, and no further biological constraints served as input to the derivation of these classification markers, it was shown[2] that most annotated active and binding sites of enzymes are covered by SPs. Moreover, other SPs were found to reside in 3D pockets inhabited by active sites. Thus it seems that SPs play important roles in carrying out the function of the enzyme. Here we wish to further investigate this point, by calculating SP coverage of annotated features other than active and binding sites, looking for the sensitivity of enzymes to mutations of amino-acids on SPs, and pointing out new observations concerning SPs, such as their roles in the realization of conformational plasticity of enzymes, and their success in classification of enzymes whose function converged with evolution.

MATERIALS AND METHODS

SP sets

Kunik *et al.*[2] have investigated 50,698 enzyme sequences of the 48.3 SwissProt release of October 2005. The same dataset is used in our study. They have extracted 42,874 SPs that specify the full EC number, that is, it corresponds to level four of the EC hierarchy and are denoted the SP4 set. Other SPs divide into 2945 in the SP3 set, 1159 in the SP2 set, and 5414 in the SP1 set. We employ these sets of SPs in our analysis.

*Correspondence to: David Horn, School of Physics and Astronomy, Tel Aviv University, Tel Aviv 69978, Israel. E-mail: horn@tau.ac.il
Received 15 July 2007; Revised 22 October 2007; Accepted 6 November 2007
Published online 4 February 2008 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/prot.21951

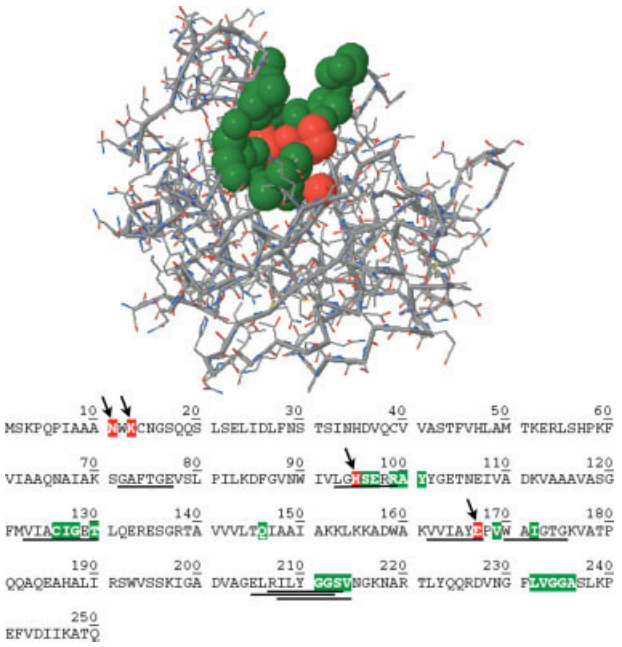

Figure 1

Example of an active pocket, on enzyme 1AG1O (PDB id), drawn using JMOL (www.jmol.org). Amino acids belonging to the active pocket are marked both in the spatial structure (above) and in the sequence (below, highlighted). The active sites are in red, and the SPs appearing on this enzyme are underlined. It can be noted that the first SP, GAFTGE, does not contain any amino acids belonging to the active pocket, while the second SP, LGHSERR, covers the active site, H, and contains three more amino acids belonging to the active pocket. Since it contains at least four amino acids belonging to the active pocket, it is defined as residing in the pocket. Three partially ovelapping SPs exist around location 210. Only one of them, ILYGGSV, is defined as residing in the pocket.

The appearance of an SP on the sequence of an enzyme implies that the enzyme belongs to the particular EC branch to which the SP belongs. On average 15.6 SPs appear on an enzyme. Obviously their EC assignments have to be consistent with one another. Examples of SPs on an enzyme sequence are shown in Figure 1. A web tool that returns the SPs appearing on a given enzyme is available at http://adios.tau.ac.il/SPSearch. When using SPs for functional classification of novel enzymes, Kunik *et al.*[2] have shown that EC classification based on the occurrence of more than one SP on an enzyme's sequence has a high degree of accuracy (see caption of Table S2 there).

Statistical significance of SP coverage of annotated features

To analyze the significance of SP coverage of a given SwissProt annotated feature, as in Table I, we compare this coverage to the expected value of a background model. The latter is defined by labeling randomly picked strings from the enzyme's sequence as pseudo-features, equivalent in size and number to the occurrences of real features. This is carried out on a nonredundant set of enzymes, constructed by finding all enzymes that include the said feature, and choosing one arbitrary enzyme for each EC number. We calculate the score as the number of standard deviations (of the background model) between the coverage of the feature by SPs and the average coverage in the background model. The background model is run 30 times to accumulate statistics.

Statistical significance of the mutagen analysis

We analyze the effect of single site mutations of SPs on enzymatic function. To do so the SwissProt annotation MUTAGEN is used. It denotes single sites which have been experimentally altered by mutagenesis, specifying whether it was successful (damaging the enzymatic function) or not. Using our dataset, a population is constructed of all MUTAGEN annotations that do not coincide with active, binding, or metal binding sites, since the latter are already known to be crucial sites for the enzymes' performance. The size of the population is defined as N. The number of successes in the population is defined as D. A hypergeometric distribution is constructed, describing the number of successes k in a sequence of n draws from a finite population N without replacement. In our application n will be the number of MUTAGEN annotations that are covered by SPs. The said probability for $k = X$ successful events is given by:

$$\Pr(k = X) = f(k; N, D, n) = \frac{\binom{D}{k}\binom{N-D}{n-k}}{\binom{N}{n}}$$

In our analysis we have $X = 867$, $N = 2814$, $D = 2562$ and $n = 919$. The P-value is the cumulant of this distribution, that is, $Pr(k \geq 867)$.

Statistical significance of SPs residing in active pockets

We define[2] an active pocket to be a 3D pocket in the spatial structure of an enzyme that includes an active or binding site. An SP is defined as residing in an active pocket if at least four of its amino acids belong to it. Figure 1 shows an example of the structure of an active pocket, the amino-acids belonging to it, SPs residing in it, and other SPs. The data regarding the amino acids that constitute pockets was taken from the CASTp database.[4] For every event, defined as the occurrence of a given SP within an active pocket in a given enzyme, a P-value is defined based on the probability that a randomly chosen motif, of the same length as the SP, falls within the active pocket. Significant events are selected according to an FDR limit[5] of 0.05.

Table I
Coverage of Biological Function Sites

Feature	No. of features	Coverage	cov NR	cov rand NR	Score in STDs (*P*-value)
PEPTIDE	33	36%	—	—	—
CA_BIND	141	19%	—	—	—
ZN_FING	349	48%	—	—	—
DNA_BIND	131	79%	—	—	—
NP_BIND	9331	75%	65%	42%	7.7 (P = 6.8e-15)
MOTIF	3346	71%	—	—	—
SITE	3757	52%	—	—	—
CARBOHYD	8895	15%	13%	20%	3.7 (P = 1.1e-04)
ACT_SITE	28,305	64%	55%	21%	30.8 (P = 0)
BINDING	22,429	64%	45%	22%	16.0 (P = 0)
METAL	38,587	59%	39%	17%	23.6 (P = 0)
All	113,485	59%	43%	22%	34.4 (P = 0)

The first column contains the feature annotation as it appears in SwissProt. Their descriptions are listed below. Next is the number of annotations found on the data set, then the percentage of these annotations that were covered by SPs. Next follow the nonredundant (NR) data set: the fourth column shows the coverage of features within the NR set, the next column shows the expected value of the coverage in the background model. Finally is the score in standard deviations and when compared with the background model (see Methods). The *P*-value is given in brackets, $P = 0$ meaning that it is smaller than the smallest positive normalized floating-point in MATLAB. The score is left blank if the NR data set (of enzymes that are annotated with the given feature) is smaller than 100. The last row displays the result when all the annotations are taken into consideration, avoiding double counting.
SwissProt feature identifiers: PEPTIDE - Extent of a released active peptide; CA_BIND - Extent of a calcium-binding region; ZN_FING - Extent of a zinc finger region; DNA_BIND - Extent of a DNA-binding region; NP_BIND - Extent of a nucleotide phosphate-binding region; MOTIF - Short (up to 20 amino acids) sequence motif of biological interest; SITE - Any interesting single amino-acid site on the sequence, that is not defined by another feature key. It can also apply to an amino acid bond which is represented by the positions of the two flanking amino acids; CARBOHYD - Glycosylation site; ACT_SITE - Amino acid(s) involved in the activity of an enzyme; BINDING - Binding site for any chemical group (coenzyme, prosthetic group, etc.); METAL - Binding site for a metal ion.

Statistical significance of differences between amino-acid distributions

Comparing amino-acid frequency distributions, we define the frequency of each amino acid in all enzymes as X_i (i = A, C, D…W, Y) and the corresponding frequency in a certain set of SPs as Y_i. We then calculate the differences between the frequencies for each amino acid, $Z_i = Y_i - X_i$, and evaluate the average and standard deviation. The *P*-value can be calculated for each amino-acid.

RESULTS

SPs and annotated biological features

Coverage of annotated biological features

SPs are known to contain 64% of annotated active and binding sites, compared to 23% ± 1% of randomly selected amino acids.[2] Here we extend the analysis of SPs to cover most annotated features in SwissProt, and present the results in Table I.

Amongst the most impressive results are the SP coverage of DNA binding annotations (79%), of nucleotide phosphate-binding annotations (75%), and of annotations of short sequence motifs of biological interest (71%). We will return to the DNA binding data in the next section. Next we find in Table I the active and binding site results, mentioned in the introduction. We see that also the coverage of metal binding sites is quite impressive, especially when considering the large amounts of data. To evaluate the significance of the coverage of a certain feature, we assemble a nonredundant (and therefore unbiased) set of enzymes (see Methods). Significance is evaluated only for features whose nonredundant set contains more than 100 enzymes. The nonredundant calculation leads to an over-all coverage of 43% of all features, with significance of 34 standard deviations in comparison with the background model (*P*-value $< 10^{-308}$, the smallest positive normalized floating-point in MATLAB).

Although the results bear high statistical significance it should be pointed out that the number of SPs carrying these biological markers is only a fraction of the total number of SP matches on these enzymes. Considering all events in Table I we find that 3534 out of 38,799 SPs (i.e. 10.8%) carry the relevant annotations.

Coverage of DNA binding regions of LexA repressors

SwissProt annotations of DNA binding regions are continuous sequences of amino-acids and are mostly associated with LexA repressors, selfcatalytic enzymes belonging to EC = 3.4.21.88 that bind to the SOS regulon of genes that are responsible for DNA repair.[6–8] The binding regions tend to be of the same length, around 21 amino-acids, and are part of a spatial H-T-H structure.[8] We have analyzed 93 LexA enzymes that contain an SP match on a DNA binding region annotation. Figure 2 displays the coverage, by SPs, of each location along the DNA binding region. This coverage is high, of the order of 70%, at the beginning and at the end of the domain. These two regions coincide with helix structures in this H-T-H stretch of the enzyme. For comparison we display in Figure 3 the sequence logo[9,10] of information carried by single amino-acids in the DNA binding regions, as determined for the

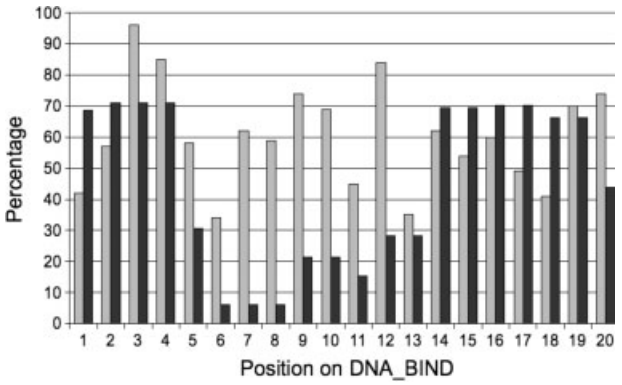

Figure 2

Comparison of SP coverage versus amino acid information in DNA binding regions of LexA enzymes. The x axis represents the position from the beginning of the DNA binding region. Black boxes display the percentage of DNA binding regions covered by SPs per position. White boxes represent the amount of information contained in single amino acids per position (as percentage of maximal information). Positions 1–7 and 13–20 are characteristically associated with helix structures in this H-T-H binding region of LexA.

same set of enzymes (the sum of all information values is also displayed in Fig. 2). Comparing SP coverage with the amino-acid logo plot we find that the conserved positions 3, 4, 19, and 20 in the latter are consistent with the SP coverage pattern, yet the conserved positions 9, 10, and 12 appear in a region (coinciding with the spatial "turn" region) with very low SP coverage. Moreover, regions with large SP coverage contain positions such as 1, 2, and 18 where no single amino-acid carries any large information. We conclude that SP coverage carries information that is complementary to that of single amino-acids. In particular, SPs seem to coincide with the helix regions at the beginning and the end of the DNA binding region.

Phylogenetic classification of bacteria through SPs on LexA

Another interesting result is obtained by looking at the SP content of the binding region of LexA enzymes of the different bacteria belonging to this group: Table II shows the sets of SPs observed in the binding region on proteobacteria of the types α, β, γ, and others. The sets of SPs clearly allow for subclassification of the relevant bacteria in 3.4.21.88 into three classes: (1) α-proteobacteria, (2) β, and γ-proteobacteria, and (3) others. Thus we observe here SPs that are not only EC specific but also specific to phylogenetic classes.

Mutated SPs damage enzyme function

Having assessed that SPs cover annotations of biological importance in a statistically significant manner, it is of interest to obtain a result on a more global scale. What is the relevance of SPs in general, especially those that do not cover sites of known biological importance? The ultimate test for biological relevance of a certain motif is experimentally altering its amino acids by mutagenesis, and looking for changes in the enzymatic function.

There are 3509 MUTAGEN annotations in our dataset. Since active sites, binding sites, and metal binding sites are already known to be crucial to the enzymes' performance, we eliminate them from the set that we consider. We are then left with 2814 MUTAGEN annotations, 2562 of which affect the enzymatic function. An event of a MUTAGEN annotation is defined as successful if the mutation in question has damaged the enzyme's performance and unsuccessful if not. 919 MUTAGEN annotations are covered by SPs, 867 of which are successful. Using the hypergeometric distribution to compare this result with the total set of MUTAGEN successes we conclude that the *P*-value of the observed results (see Methods) is 3.5e−06, making them highly significant. This supports the statement that mutated SPs, as a whole, tend to damage the enzymatic function.

Glycine-enriched SPs in active pockets

SPs were found to be located in 3D pockets containing active or binding sites,[2] to which we refer as active

Figure 3

Sequence Logo of LexA DNA binding regions.

Table II
Subclassification of Proteobacteria According to SPs that Cover the Binding Regions of Their LexA Enzymes

Type	No. of Enzymes		Sets of SPs
Alpha	17	16	KSGIHR, PSFDEMK, SKSGIHRLI
		1	KSGIHR, SKSGIHRLI
Beta	8	4	GFRSPNAAE, PPTRAEI
		3	NAAEEHL, PPTRAEI
		1	PPTRAEI
Gamma	37	14	NAAEEHL, PPTRAEI
		11	AEEHLKALARKGVIEI, GFRSPNAAE, NAAEEHL, PPTRAEI
		5	RAAQYHLEALE
		4	GFRSPNAAE, NAAEEHL, PPTRAEI
		1	NAAEEHL
		1	AEEHLKALARKGVIEI, PPTRAEI
		1	GFRSPNAAE, PPTRAEI
Other	42	16	SVREIG, GYPPSVREI, STVHGH
		8	RGYPPSIREI
		5	SVREIG, GYPPSVREI, REIGQAVGL, STVHGH
		4	GYPPSVREI, STVHGH
		4	STVHGH
		3	REIGQAVGL
		2	GYPPSVREI

The first column contains the class of proteobacteria, followed by the numbers of relevant enzymes belonging to them. Next come sets of SPs whose common appearance is observed on these enzymes, preceded by the number of their occurrences. SP sets appearing on different classes (1: α, 2: β and γ, 3: other) are disjoint.

pockets. An example is shown in Figure 1. Their statistical significance was determined by comparison to a background model (see Methods). Here we stress a novel finding concerning this particular set of SPs. Comparing the relative frequencies of all amino-acids occurring on these SPs with the frequencies observed on enzymes in general one finds a clear overrepresentation of glycine. Table III compares the glycine frequency on enzymes with that on SPs in general and with SPs in active pockets. It turns out that it is highest for SPs that lie in these active pockets.

Table III
Frequencies of the Glycine Amino Acid in Various Data Sets

Dataset	Glycine frequency
All enzymes	7.50
All SPs	9.20
SPs in active pockets	11.1% (P = 4.0e-04)
SPs in active pockets, not on site	11.0% (P = 2.9e-04)
Amino acids in active pockets	8.60%
Amino acids in active pockets not on SPs	7.30%

SPs in active pockets refers to SPs whose occurrence in active pockets is statistically significant (see Methods). The P-value refers to a comparison with frequency distributions of amino-acids in all enzymes (see Methods). Glycine frequency is normal when considering all SPs and amongst the amino acids that constitute the active pockets and that are not contained in SPs.

Table IV
Thirteen Sets of Functionally Convergent Enzymes from Ref. 19

EC	SwissProt 1	Fold 1	L1	SwissProt 2	Fold 2	L2
1.11.1.10	prxc_psepy	3.048.001	1*	prxc_curin	1.068.001	1*
1.15.1.1	sodc1_orysa	2.001.007	4	sodm_bacca	4.023.001	4
3.1.3.48	ptpa_strco	3.028.001	—	pyp3_schpo	3.029.001	4
3.1.26.4	rnh_ecoli	3.038.003	4	rnh_bpt4	3.039.001	—
3.2.1.4	gun_bacsz	1.061.001	—	gun_paepo	3.001.001	4*
3.2.1.8	xyn_triha	2.018.001	4	xynb_thene	3.001.001	4
3.2.1.14	chia_tobac	3.001.001	4	chix_pea	4.002.001	4
3.2.1.73	gub_nicpl	3.001.001	3	gub_bacsu	2.018.001	4
3.2.1.73	gub_bacci	1.061.001	—			
3.2.1.91	gux1_trivi	2.018.001	4	gux3_agabi	3.002.001	1
3.5.2.6	blp4_pseae	5.003.001	4	blab_bacce	4.083.001	4
4.2.1.1	cah_mette	2.053.001	—	cahz_brare	2.047.001	4
5.2.1.8	mip_trycr	4.018.001	4	cypr_drome	2.041.001	4
5.4.99.5	chmu_yeast	1.079.001	—	chmu_bacsu	4.037.001	—

Each row contains a pair of enzymes sharing the same function, but for row 9 which contains one enzyme that shares the EC number with the pair of row 8. For each enzyme we quote the SwissProt identification, the fold number and the EC level to which we were able to classify it using SPs (L1, L2). Cases marked by * correspond to doubtful classifications based on just one SP match on the enzyme's sequence.

Glycine is the smallest amino acid, having effectively no side chain, and therefore bestows rotational flexibility to the site (i.e appears in turns and hinges) and contributes to packing of nearby residues. It is generally accepted[11] that the location of glycines in the structure of a protein influences its motion. Yan and Sun[12] have demonstrated in a study of X-ray crystallographic data of 23 enzymes that active site regions are rich in GXY and YXG oligopeptides where X and Y are polar and nonpolar residues. Hence they concluded that glycine residues may provide conformational flexibility[13–15] to active pockets in enzymes. The strategic location of glycine at or in close proximity to the active site has also been noted in the same context for some other enzymes.[16–18] Our large-scale analysis in Table IV shows that glycine enrichment is mostly apparent in SPs that reside in active pockets. Hence we hypothesize that these enriched SPs are of particular importance for the conformational flexibility of the enzyme.

At this point we wish to recall that single site mutations involving glycines, that is, mutating glycines or changing nonglycine to glycine residues, can be lethal. Such mutations always affect protein stability,[20,21] cause changes in specificity,[22] and are responsible for about 15% of human genetic diseases.[23]

SPs may resolve difficult classification problems

Given the results quoted so far, that point to the biological importance of SPs, we turn to a classification problem that is deemed to be very difficult, and could be resolved by SPs provided they capture the essence of the

function that is being assessed. Kunik *et al.*[2] have shown that EC classification based on the occurrence of more than one SP on an enzyme's sequence has a high degree of accuracy (see caption of Table S2 there). Conventional classification methods, using sequence or structural similarity, may fail to classify correctly enzymes whose functions converged with evolution — enzymes with different ancestors that perform the same function. Such enzymes may perform exactly the same function (i.e may have the same four components of the EC number), and yet have completely unrelated spatial structures (i.e involve different folds) as well as different sequences.

Hegyi *et al.*[19] quote 13 sets of enzymes (12 pairs and one triplet) with specific functional convergences involving different folds. These examples are shown in Table IV where, in addition to the pairs of enzymes sharing the same function, we display the fold numbers from SCOP 1.35 and the levels of correct EC hierarchy as determined by SPs located on these enzymes (cases where only one SP match occurs on the sequence, should be regarded as doubtful classifications, and are designated by an * in Table IV). For example, the two enzymes in the 8th row in Table IV perform beta-glucanase (EC 3.2.1.73). The first enzyme, gub_nicpl, has a fold number 3.001.001 and SPs classify it with the specificity of the third component of the EC number 3.2.1 (L1 = 3). The second enzyme, gub_bacsu, has a fold number 2.018.001 and SPs classifiy its complete EC number (3.2.1.73), that is, L2 = 4. SPs classify correctly 8 pairs out of the 13 sets, 5 of which are classified completely (the full EC number).

The sequence similarity of the pairs of enzymes in Table IV is very small: Blastp[24] run on these pairs (using the BLOSUM64 similarity matrix, a gap penalty of 11 and a gap extension penalty of 1) finds no significant similarity. Smith-Waterman[25] similarity test with the same parameters results in scores ranging between 24 and 131 (an average of 42.6 $\pm$ 25.2). Hence we conclude that SPs are able to compare correctly pairs of enzymes with remote homology both in sequence and in structure. Thus we have partially resolved a difficult problem in functional classification.[19]

DISCUSSION

The relevance of SPs to biological functions is evaluated by finding the coverage of residues that are known to be crucial to these functions, such as active sites, metal binding sites, Ca binding sites, and so forth. Most of the functional annotations are well covered by SPs. The statistical significance of the observed coverage was evaluated on nonredundant datasets. The results are extremely significant, establishing the existence of biological markers on more than 10% of all SPs.

DNA binding region annotations for enzymes are relevant mainly to LexA repressors. We have found that their coverage by SPs is peaked at the beginning and at the end of the region, coinciding with the helices in an H-T-H spatial structure. Moreover we demonstrated that the information carried by the SPs is complementary to that derived from evolutionary conservation of single amino-acids.

We have also demonstrated on the LexA enzymes, that the SPs on the DNA binding region allow for phylogenetic subclassification of the species carrying these enzymes. This could have been expected since SPs are highly conserved fractions of proteins that belong to different species and share the same EC classification.

To verify the biological importance of any individual SP one should perform some kind of alanine-experiment, or any other single-site mutation on the different residues of the SP, to test how crucial the SP is to the function of the enzyme. In our large-scale study, this can be replaced by checking the MUTAGEN annotations occurring on SP matches in enzyme sequences. Successes of such mutagen experiments on SPs turn out to be highly significant.

Analysing SPs that reside in active pockets we find that they are significantly enriched with glycine. This holds even when compared with the distribution of all residues in active pockets. We suggest the interpretation that these SPs are responsible for conformational plasticity in these enzymes.

Finally we have studied the difficult classification problem of convergent evolution. SPs were found to be important carriers of information in remote homology situations as exemplified in Table IV. This is a particularly difficult problem since the pairs of enzymes studied here are dissimilar in both sequence and 3D structure.

CONCLUSION

Our study substantiates the importance of specific peptides[2] as biologically relevant functional elements. We have also found further proof for their importance as classification tools when straightforward sequence similarity does not provide a clear functional prediction.

In particular, we have presented new evidence for the relevance of SPs to metal binding sites, to DNA binding regions and to other features annotated in SwissProt data. SPs were shown to be very sensitive to mutation experiments, thus demonstrating their crucial importance to maintaining the stability and functionality of enzymes.

Studying SPs that occur within spatial pockets containing active or binding sites, we have discovered that they are glycine-enriched. This striking phenomenon suggests that SPs are responsible for conformational plasticity.

ACKNOWLEDGMENTS

We thank Nir Ben-Tal, Eli Eisenberg, Uri Gophna, Guy Nimrod, Eytan Ruppin, Alexandra Shulman, and Haim Wolfson for helpful discussions.

REFERENCES

1. Rost B, Liu J, Nair R, Wrzeszczynski KO, Ofran Y. Prediction of protein functional specificity without an alignment. OMICS J Integr Biol 2006;10:56–65.
2. Kunik V, Meroz Y, Sandbank B, Ruppin E, Horn D. Functional representation of enzymes by specific enzymes. PLOS Comput Biol 2007;3(8):e167.
3. Solan Z, Horn D, Ruppin E, Edelman S. Unsupervised learning of natural languages. Proc Natl Acad Sci USA 2005;102:11629–11634.
4. Binkowski TA, Naghibzadeg S, Liang J. Castp: computed atlas of surface topography of proteins. Nucleic Acid Res 2003;31:3352–3355.
5. Benjamini Y, Hochberg Y. Controlling the flase discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc 1995;57:289–300.
6. Horii T, Ogawa T, Ogawa H. Nucleotide sequence of the lexA gene of *E. coli*. Cell 1981;23:689–697.
7. Brent R, Ptashne M. Mechanism of action of the lexA gene product. Proc Natl Acad Sci USA 1981;78:4204–4208.
8. Fogh RH, Ottleben G, Rueterjans H, Schnarr M, Boelens R, Kaptein R. Solution structure of the LexA repressor DNA binding domain determined by 1H NMR spectroscopy. EMBO J 1994;13:3936–3944.
9. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. Genome Res 2004;14:1188–1190.
10. Schneider TD, Stephens RM. Sequence logos: a new way to display consensus sequences. Nucleic Acids Res. 1990;18:6097–6100.
11. Subbiah S. Protein motions, molecular biology intelligence unit. Heidelberg, Germany: Springer-Verlag; 1996.
12. Yan BX, Sun YQ. Glycine residues provide flexibility for enzyme active sites. J Bio Chem 1997;272:3190–3194.
13. Koshland DE. Application of a theory of enzyme specificity to protein synthesis. Proc Natl Acad Sci USA 1958;44:98–104.
14. Kraut J. How do enzymes work? Science 1988;242:533–540.
15. Bone R, Sile JL, Aghard DA. Structural plasticity broadens the specificity of an engineered protease. Nature 1989;339:191–195.
16. Peters GH, Bywater RP. Computational analysis of chain flexibility and fluctuations in Rhizomucor miehei lipase. Protein Eng 1999;12:747–754.
17. Teplyakov A, Sebastiao P, Obmolova G, Perrakis A, Brush GS, Bessman MJ, Wilson KS. Crystal structure of bacteriophage T4 deoxynucleotide kinase with its substrates dGMP and ATP. EMBO J 1996;15:3487–3497.
18. Narayana N, Cox S, Xuong N, Ten Eyck LF, Taylor SS. A binary complex of the catalytic subunit of cAMP-dependent protein kinase and adenosine further defines conformational flexibility. Structure 1997;5:921–935.
19. Hegyi H, Gerstein M. The Relationship between protein structure and function: a comprehensive survey with application to the yeast genome. J Mol Biol 1999;288:147–164.
20. Bordner AJ, Abagyan RA. Large-scale prediction of protein geometry and stability changes for arbitrary single point mutations. Proteins, 2004;57:400–413.
21. Serrano L, Sancho J, Hirshberg M, Fersht AR. Alpha-helix stability in proteins 1. Empirical correlations concerning substitution of side-chains at the N and C-caps and the replacement of alanine by glycine or serine at solvent exposed surfaces. J Mol Biol 1992;227:544–559.
22. Vermersch PS, Tesmer JGG, Lemon DD, Quiocho FA. A pro to gly mutation in the hinge of the arabinose-binding protein enhances binding and alters specificity-sugar-binding and crystallographic studies. J Biol Chem 1999;265:16592–16603.
23. Vitkup D, Sander C, Church GM. The amino-acid mutational spectrum of human genetic disease. Genome Biol 2003;4:R72.
24. Tatusova TA, Madden TL. Blast 2 sequences—a new tool for comparing protein and nucleotide sequences. FEMS Microbiol Lett 1999;174:247–250.
25. Smith T, Waterman M. Identification of common molecular subsequences. J Mol Biol 1981;147:195–197.