Roles of Specific Peptides in Enzymes

Yasmine Meroz

A thesis submitted to the

School of Physics and Astronomy

Raymond and Beverly Sackler Faculty of Exact Sciences

Tel-Aviv University

in Partial Fulfillment of the Requirements for the Degree

of

Master of Science

Under the Supervision of

Prof. David Horn

March 2007

Abstract

Specific Peptides (SPs) have been shown [22] to specify the functions of 93% of the enzymes on which they occur. In this work we will focus on the biological importance and possible roles of SPs in the realisation of enzymatic functions.

SPs are shown to provide correct functional classification in problems where conventional methods, based on sequence or structure similarity, fail. These cases include enzymatic functions that diverged or converged with evolution.

Analysing the coverage of functional annotations of enzymes, we demonstrate that SPs contain major fractions of all annotated biological features. One such feature, DNA binding, is further analysed and observed to show interesting coverage patterns. Moreover, its SPs allow for sub-classification of the species which possess this function into phylogenetic classes.

An analysis of sites which have been experimentally altered by mutagenesis leads to the conclusion that SPs contain much more sites that affect the enzyme's function when mutated than a background model, hence are highly important to enzymatic functions.

Events of SPs occurring in three-dimensional pockets of active sites (and other sites of functional importance) are found to be statistically significant. These SPs may play important roles in bringing about the enzymatic function, mostly unknown so far. These SPs are shown to be significantly enriched by glycine, thus leading to the hypothesis that they are responsible for the induced-fit mechanism.

Acknowledgements

I am greatly indebted to my advisor, Prof. David Horn, who showed me the fun in delving into things I do not know and do not understand. As Dante Alighieri put it in *La Divina Commedia* (*Inferno, Canto XI*);

> "O sol che sani ogni vista turbata,
> tu mi contenti sì quando tu solvi,
> che, non men che saver, dubbiar m'aggrata."
>
> "O Sun that healest all dim sight, thou so
> dost charm me in resolving of my doubt,
> to be perplexed is pleasant as to know."

I would also like to thank my lab-mates - Assaf, Liat, Roy, Uri, Vered, Yaron and Zach - who made me smile every day, infallibly.

Last, but certainly not least, I thank my parents, sisters and grandmother for their constant love and support. And then there's Ely, who always believed in me more than I did myself.

Contents

Chapter 1

INTRODUCTION

An outstanding challenge in molecular biology is to predict the spatial structure of proteins and their function from the protein sequence of amino-acids [13, 33]. The conventional approach is to rely on sequence similarity (homology) of the protein in question with other proteins whose structure and function is known: high sequence-similarity ensures similar structures and functions [40], but this is sometimes misleading [16, 32]. Alternatively one may use motif approaches [2, 5, 10, 14, 19, 29], trying to extract from the data sub-sequences 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 uses deterministic sequence-motifs, and concentrates on shedding light on their relationships with protein functionality.

Conventional sequence-motif extraction in enzymes is 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. These motifs are then postulated to being crucial in the enzymes' functional performance. Large-scale studies often make use of multiple sequence alignment, phylogenetic information, and sophisticated mathematical models, thus leading to the plethora of time and resource demanding 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. This is the approach we have taken in [22]. The derivation of motifs does not use pre-processing by multiple sequence alignment, does not search for over-representation in functional categories, and does not rely on any phylogenetic information.

A large-scale search for deterministic sequence-motifs was performed [22], without specifying a-priori their exact functional roles by applying an unsupervised motif extraction algorithm (MEX - described in Appendix A) to 50,698 enzyme sequences. The resulting motifs were then filtered by their specificity relative to the four-level classification hierarchy of the Enzyme Commission (EC), obtaining 52,216 exact motifs, named Specific Peptides (SPs). For a precise account of the SP extraction and filtering procedures see Methods in Chapter 7. By representing some 50,000 enzymes (of average length 380 amino acids) in terms of about the same number of SPs (on average 8.4 $\pm$4.5 amino-acids), a largely compressed functional representation and an EC classification with 93% accuracy is obtained [22]. This may be compared with other methods such as the one based on sequence motifs of [7] or predicting functionality on the basis of sequence similarity using SVM classification [23]. In comparison with the large-scale and popular motif database Pro-Site [5], our approach displays a wide-margin advantage - 93% coverage compared to Pro-Site motifs coverage extending only to 63% of all enzymes in the database. This is further analysed in Appendix B.

This work puts its focus on the question whether SPs are of biological importance. This can be dealt with by asking, first, whether SPs contain regions of the enzyme that are already known to be of importance to the performance of the enzymatic feature. This point can be further strengthened by carrying a statistical analysis of the effects of experimental mutations in the SP on the enzymes' performance. Based upon the results to these questions, one can then ask whether certain SPs may contain regions of previously unsuspected functional importance.

Chapter 2 gives basic biological terms needed for this work. Statistical calculations carried out to obtain results that appear throughout this work are available in Chapter 7, the Methods section. In Chapter 3 we point out the roles of SPs in correctly classifying classes of enzymes that pose a particular problem in conventional methods of classification, such as enzymes with functions that converged or diverged with evolution. The biological relevance of SPs is investigated in Chapter 4. Although only enzyme sequences were used in the analysis, and no further biological constraints served as input to the derivation of these classification markers, we will show that most annotated active and binding sites of enzymes are covered by SPs. We further investigate this point, by calculating SP coverage of annotated features other than active and binding sites and evaluating the sensitivity of enzymes to mutations of amino-acids on SPs. Moreover, other SPs were found to reside in 3D pockets inhabited by active sites, these being candidates for motifs holding novel biological features, previously un-annotated. Chapter 5 discusses this point, analysing characteristics of certain groups of SPs that may be of biological importance. SPs in pockets of certain annotations are found to be significantly glycine-enriched, hinting at their role in the induced fit mechanism.

This work is based on the following papers:

1. V. Kunik, Y. Meroz, B. Sandbank, E. Ruppin and D. Horn (2007) Functional representation of enzymes by Specific Peptides, *submitted for publication.*

2. Y. Meroz and D. Horn (2007) Roles of specific peptides, *submitted to ISMB 2007.*

Chapter 2

BIOLOGICAL BACKGROUND

2.1 Enzymes

In the 19th century, Louis Pasteur's attention was directed to the study of the fermentation of sugar to alcohol by yeast. It was in 1860 that he concluded that the fermentation was catalyzed by something inherent to yeast cells, what he called ferments. In 1878 Wilhelm Kühne described this process using the term enzyme, from Greek $\epsilon\nu\zeta\upsilon\mu\omicron\nu$ "in leaven". The enzymes, or ferments, that Pasteur talked about were thought to act only within living cells, but this was proven wrong in 1897, when Eduard Buchner discovered the ability of yeast extracts to ferment sugar outside living yeast cells.

Many chemical processes take place all the time in a properly functioning cell, most of which naturally occur at a rate too low to fulfill the cells needs; this is where enzymes come into action. Enzymes are proteins that catalyze , i.e accelerate, chemical reactions, converting molecules (substrates) into other molecules (the product). The dramatic acceleration of the rate of the reaction is obtained by effectively lowering its activation energy.

Like all proteins, enzymes are macromolecules made up of long chains of amino acids, 20 kinds of small molecules, that fold into a three-dimensional structure. The names of the 20 amino acids are represented by 20 letters. Different enzymes catalyze different processes, and their functions are greatly determined by their spatial structure. Very few of the enzymes' amino acids have direct contact with

```
  1 MSDQQQPPVY KIALGIEYDG SKYYGWQRQN EVRSVQEKLE KALSQVANEP ITVFCAGRTD   60
 61 AGVHGTGQVV HFETTALRKD AAWTLGVNAN LPGDIAVRWV KTVPDDFHAR FSATARRYRY  120
121 IIYNHRLRPA VLSKGVTHFY EPLDAERMHR AAQCLLGEND FTSFRAVQCQ SRTPWRNVMH  180
181 INVTRHGPYV VVDIKANAFV HHMVRNIVGS LMEVGAHNQP ESWIAELLAA KDRTLAAATA  240
241 KAEGLYLVAV DYPDRYDLPK PPMGPLFLAD                                   270
```

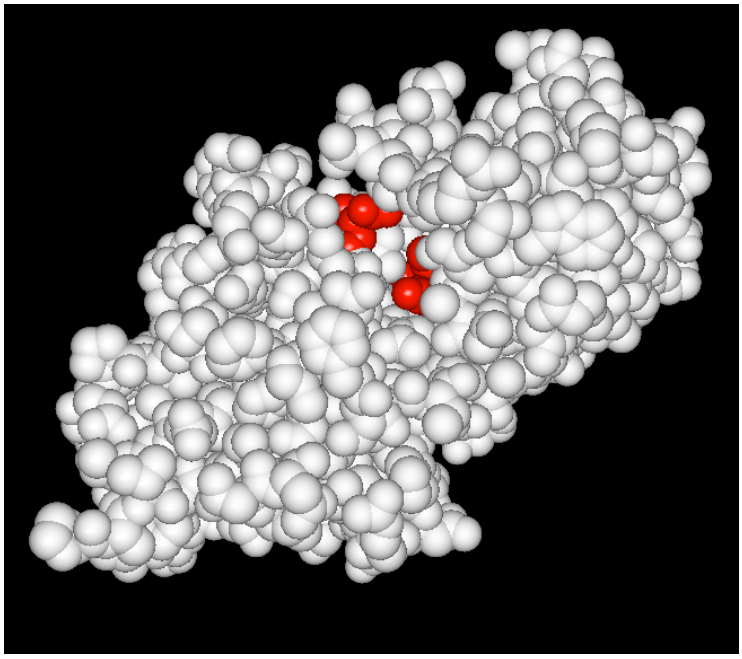

Figure 2-1: The sequence and 3D structure of the enzyme *tRNA pseudouridine synthase A* (PDB 1DJ0A). Active sites are marked in red.

the substrate in question, and these are called active sites. Figure 2-1 shows an example of an enzyme's sequence of amino acids, and its three-dimensional structure. The active sites are highlighted in both.

2.2 Functional Classification

The International Union of Biochemistry and Molecular Biology developed a nomenclature for enzymes; the Enzyme Comission (EC) numbers. Enzymes are classified according to their functionality, the classification being represented by four numbers, four hierarchical levels of classification: N1.N2.N3.N4. Each number represents the classification in the appropriate hierarchy. The first number, N1, represents the first level of hierarchy, broadly classifying the enzyme based on its mechanism. The top-level classification is: EC 1 Oxidoreductases: catalyze oxidation/reduction reactions, EC 2 Transferases: transfer a functional group (e.g.

a methyl or phosphate group), EC 3 Hydrolases: catalyze the hydrolysis of various bonds, EC 4 Lyases: cleave various bonds by means other than hydrolysis and oxidation, EC 5 Isomerases: catalyze isomerization changes within a single molecule and EC 6 Ligases: join two molecules with covalent bonds. These groups, in turn, have more subclassifications, for example transferases have nine subdivisions, meaning that for N1=2 N2 ranges between the values 1 to 9. In the end, the EC number N1.N2.N3.N4 defines the enzymes function, so that all the enzymes bearing the same EC number may belong to different species but have the same function.

2.3 Spatial structure

As was mentioned before, proteins are sequences of amino acids that fold into specific three-dimensional structures, in which they perform their particular biological function. The tertiary structure of a protein is its overall shape, also known as its fold. Proteins can be classified according to these folds, as is done in databases such as SCOP [27] and CATH [30]. The secondary structure is the general three-dimensional form of *local* segments of the protein; the protein fold is made up of 'structural building blocks'. The most common secondary-structures are α-helices and β-sheets. An α-helix looks much like a drill or fusilli pasta, while a β-sheet consists of β strands (stretches of about 5-10 amino acids whose peptide backbones are almost fully extended) connected laterally by three or more hydrogen bonds, forming a generally twisted, pleated sheet. An example of the tertiary and secondary structure of the enzyme from Figure 2-1 is brought in Figure 2-2. The two most common ways of determining experimentally the structure of a protein are X-ray crystallography and NMR spectroscopy. Determining the structure of proteins is imperative since the structure is greatly responsible for the proteins function.

(A)

```
  1 MSDQQQPPVY KIALGIEYDG SKYYGWQRQN EVRSVQEKLE KALSQVANEP ITVFCAGRTD    60
 61 AGVHGTGQVV HFETTALRKD AAWTLGVNAN LPGDIAVRWV KTVPDDFHAR FSATARRYRY   120
121 IIYNHRLRPA VLSKGVTHFY EPLDAERMHR AAQCLLGEND FTSFRAVQCQ SRTPWRNVMH   180
181 INVTRHGPYV VVDIKANAFV HHMVRNIVGS LMEVGAHNQP ESWIAELLAA KDRTLAAATA   240
241 KAEGLYLVAV DYPDRYDLPK PPMGPLFLAD                                    270
```

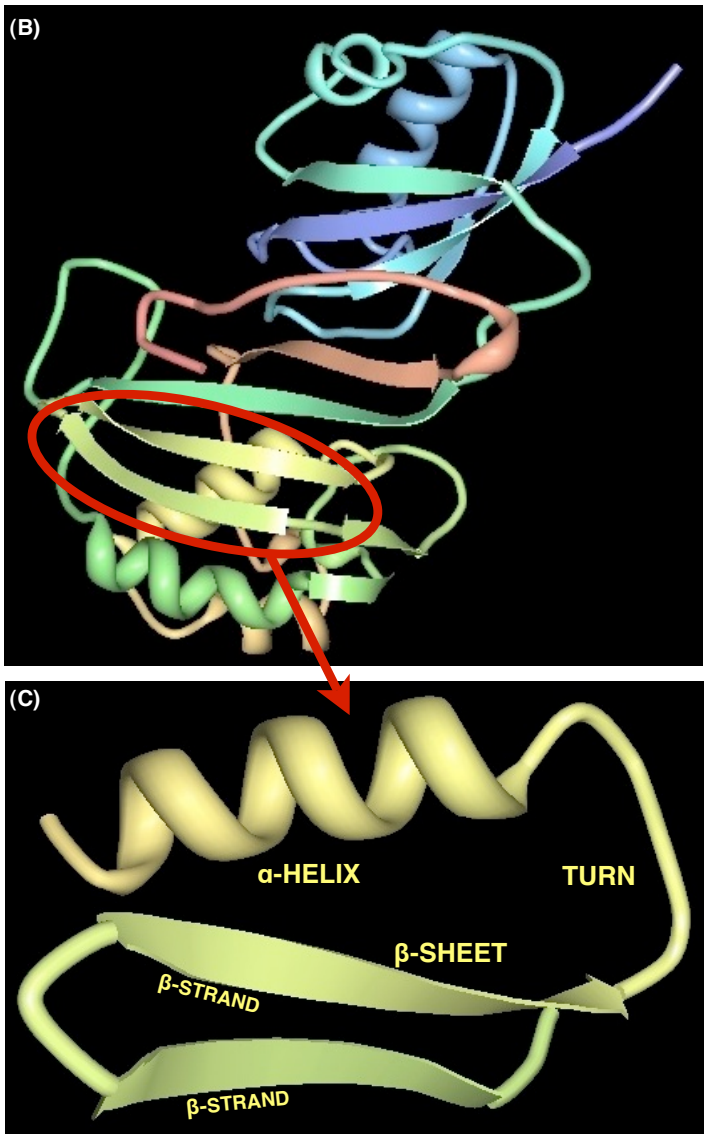

Figure 2-2: The sequence and secondary structures of enzyme 1DJ0 (the same enzyme shown in Figure 2-1). (a) The sequence, with a β-sheet highlighted in green, an α-helix in blue, and a turn that connects the two in red. (b) The structure in *ribbons*, meaning that the secondary structures can be identified sequentially. (c) The structural form of the highlighted sub-sequence of the secondary structures in (a) is magnified, revealing the β-sheet (made of two β-strands represented by flat arrows), the α-helix the turn connecting them.

Chapter 3

FUNCTIONAL CLASSIFICATION

Breakthrough developments in high throughput sequencing in the last years are the main reason for the information overload currently experienced in biological research. This is evident when considering the very high quantities of sequence information available in public databases, and their rapid growth rate. The number of protein sequences is far greater than the number of proteins whose structure and function has been experimentally determined. Therefore, researchers rely on automatic methods to classify new protein sequences into functional and structural hierarchies.

Conventional methods used to determine the functional classification of an enzyme are based on homology. Two proteins are said to be homologous if they have evolved from the same common ancestor, having similar structures and/or sequences. One strategy uses sequence similarity. For example the Smith Waterman algorithm [36] may be used to assess the sequence similarity between two proteins. The resulting scores will determine whether the two are homologous or not. This algorithm will be used throughout this chapter, and a short account of it may be found in Appendix C. Another main method takes advantage of fold similarity. The Structural Classification of Proteins (SCOP) database [27] classifies proteins (as its name hints) by their structures, creating a hierarchy of classes, folds, superfamilies, families, function (protein domains) and lastly the actual structures in different species.

In both cases there are counter examples that show that the similarity as-

sumption may lead to wrong conclusions. Such examples include enzymes with convergent and divergent functions (as shall be seen in sections 4.2 and 4.3).

It is important to note that many of the SPs may have been formed due to sequence similarity, since over-represented motifs are obviously extracted by the MEX algorithm. On the other hand, it has been pointed out that MEX also produces motifs that are not necessarily over-represented, meaning that SPs may be refined enough to classify correctly some extreme examples, where sequence and even structure similarity are not enough. SPs may be regarded as the essence of homology. This point will be examined in this chapter.

3.1 Success rate of classification with SPs.

Applying MEX to the data, and filtering the results by requiring specific peptides within the EC hierarchy, [22] were able to classify most enzymes by SPs occurring on them with coverage between 87% to 93% depending on the EC level that is being looked for (see Table 1 in [22]). Classification success of novel sequences that belong to the same type of data is of order 84-86% (see Table 2 in [22]). With a restriction to low bias (Table 3 in [22]), a precision of 88% is reached on the enzymes covered by SPs.

3.2 Non-homologous enzymes with high sequence similarity

Rost *et.al.* [32] (see Table 1 there) show examples of enzymes where conventional functional classification based on sequence similarity fails. These examples are comprised of pairs of enzymes that have different functional assignments, yet share large sequence identity. Table 3.1 further demonstrates the point made by [6], showing successful classification of the said pairs of enzymes, using SPs. All displayed EC assignments are substantiated by corresponding SPs located on these enzymes, most belonging to SP4. For each pair of enzymes in the table, a Smith-

Waterman alignment [26, 36] has been calculated with a gap opening penalty of 11, gap extension penalty of 1 and the BLOSUM62 similarity matrix (see Appendix C). The results of the alignment include the percentage of sequence identity, the alignment length, the Smith-Waterman alignment score and the appropriate E-value. The E-value is a statistical estimator for the validity of alignment scores. It is defined as the expected number of false positives with a score higher than the observed score. This value depends on the number of random alignments, determined by the size of the aligned sequences. A lower E-value indicates that the score has a higher confidence level. These results demonstrate the high sequence similarity these pairs of enzymes share.

As a more detailed example we note the sixth pair of enzymes in Table 3.1, GTFB_STRMU and AMY3B_ORYSA, having 42% sequence identity along an alignment of 105 amino acids with an alignment score of 106 and and E-value of 7.4e-08. The pair is correctly classified by different SPs: AMY3B_ORYSA contains 24 SPs, none of which have an exact match on GTFB_STRMU, and a single SP4 (GGAFLE) found on the latter determines correctly its EC classification. It should be noted that 7 of the 16 enzymes in Table 3.1 were not in the original data set on which MEX was run, and were correctly classified nonetheless.

Enzyme 1	Enzyme 2	Seq. id.	Al. len.	score	E-value
GUNA_PSEFL 3.2.1.4	MDHP_FLABI 1.1.1.82	69 %	29	64	1.6 e-03
PLB1_YEAST 3.1.1.5	METB_ARATH 2.5.1.48	60 %	30	73	5.9 e-05
RPB1_PLAFD* 2.7.7.6	UBC2_YEAST 6.3.2.19	61%	28	84	1.8 e-05
CHIB_POPTR 3.2.1.14	DGK2_DROME* 2.7.1.107	58%	24	80	6.0 e-06
ODO2_FUGRU 2.3.1.61	PP2BB_HUMAN 3.1.3.16	48%	46	86	1.1 e-06
GTFB_STRMU* 2.4.1.5	AMY3B_ORYSA 3.2.1.1	36%	105	106	7.4 e-08
RPB1_PLAFD* 2.7.7.6	PDE3B_RA* 3.1.4.17	57%	37	107	8.4 e-08
IGF1R_HUMAN* 2.7.10.1	PTPRU_HUMAN* 3.1.3.48	28%	170	123	1.5 e-09

Table 3.1: Enzymes with high sequence similarity and different EC assignments. Alignment length and sequence identity are calculated according to the Smith-Waterman algorithm [26, 36]. The alignment score and its statistical validity appear in the last two columns, demonstrating the high sequence similarity shared by each pair of enzymes. EC assignments agree with SPs occurring on the enzymes. Enzymes denoted by a star were not in our original data set on which MEX was run.

3.3 Functionally divergent enzymes

Another example along the lines of the previous section, is enzymes whose functions diverged with evolution. These are enzymes that originate from the same common ancestor (and therefore have similar structures and sequences), yet evolved to perform different functions (different EC numbers). Conventional functional classification based on sequence similarity fails on such enzymes, since the high sequence similarity leads to misleading conclusions.

A classical example of divergent functionality is TIM barrel enzymes that possess similar folds but different functions [18]. Figure 3-1 shows a schematic representation of the structure of such TIM barrel enzymes. Out of 87 known TIM barrel enzymes we studied 73 which belong to our data set. We found SP hits on 84% of them. SPs specify correctly the full EC number of 73% of the TIM barrels, and classify correctly 6% to the 3rd EC level, 1% to the 2nd EC level, and 4% to the 1st level.

3.4 Functionally convergent enzymes involving different folds

An opposite problem to the previous section is that of functionally convergent enzymes. Such enzymes originate from different ancestors, but evolve to perform exactly the same function. Although these pairs of enzymes share the same four components of the EC number, they have completely unrelated spatial structures (i.e involve different folds) and also very low sequence similarity (an extreme example of *remote homology*). Figure 3-2 shows an example of such a pair.

In this case both methods using sequence similarity and methods based on structure similarity will fail. Another method used to deal with small functional shifts, such as changes in the 4th level of EC classification among homologous enzymes, relies on studies of amino acid evolutionary changes on various positions of multiply-aligned enzymes, such as in [1], although this strategy is very resource-

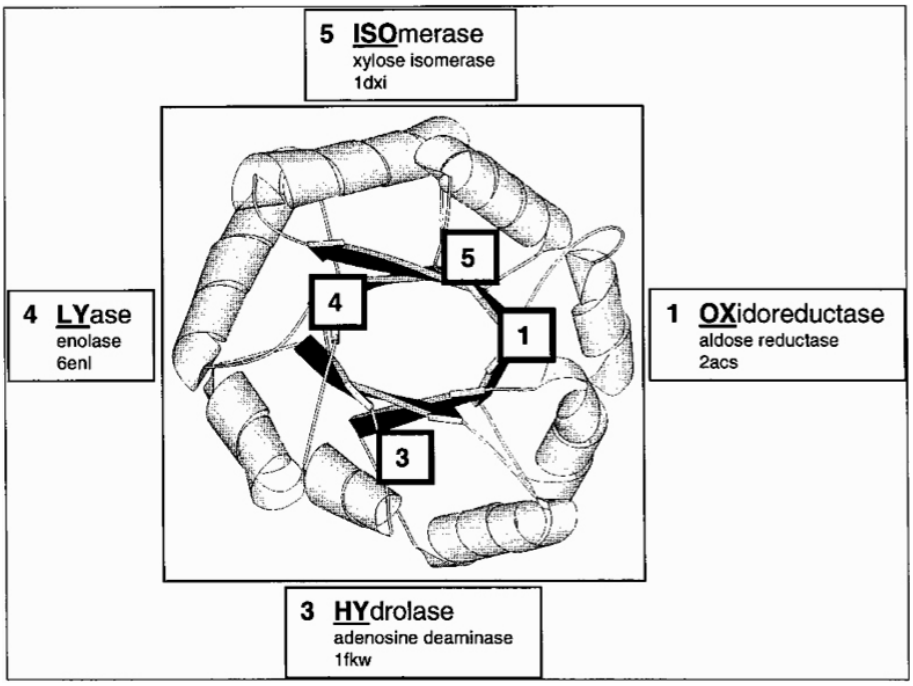

Figure 3-1: An example of divergent evolution, the TIM-barrel. This fold functions as a generic scaffold catalysing 15 different enzymatic functions. A schematic figure of the TIM-barrrel fold is shown with numbers in boxes indicating the different location of the active sites in four proteins that have this fold. These four proteins, xylose isomerase, aldose redductase, enolase and adenosine deaminase, carry our very different enzymatic functions, in four of the main EC classes (1.-.-.-, 3.-.-.-, 4.-.-.- and 5.-.-.-). They have active sites at very different locations (identified by the boxed numbers in the barrel) yet they all share the same fold.

demanding.

Discrete sequence motifs, although often extracted from homology, may serve as measures for functional specification of proteins [6]. Indeed, using the SP methodology we do not rely on sequence similarity or multiple sequence alignments, yet we can attack convergence and divergence problems even at the 1st EC level, as is shown in this chapter.

Hegyi & Gerstein [18] quote 13 sets of enzymes (12 pairs and one triplet) with specific functional convergences involving different folds. These examples are shown in Table 3.2 where, in addition to the pairs of enzymes sharing the same function, we display the levels of correct EC hierarchy as determined by SPs located on these enzymes. For example, the two enzymes in the 8th row in Table 3.2 perform beta-glucanase (EC 3.2.1.73). The first enzyme, gub_nicpl, has a fold

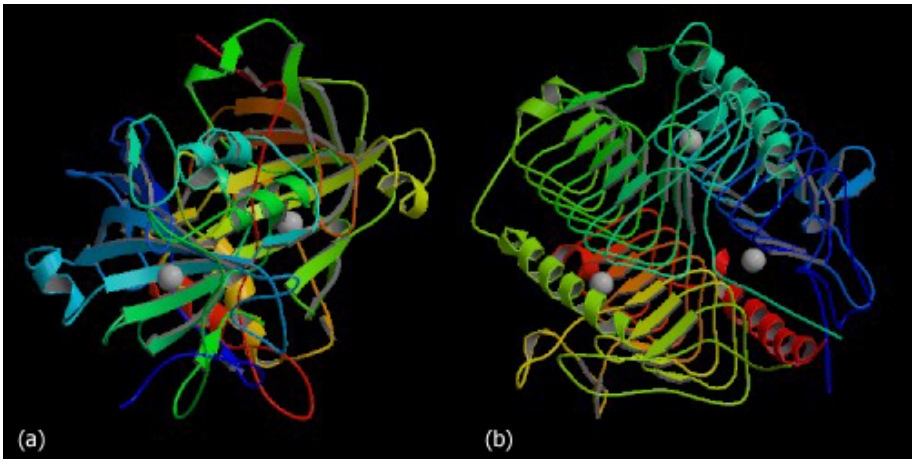

Figure 3-2: An example of a pair of functionally convergent enzymes: (a) 1DMX and (b) 1THJ. These are two carbonic anhydrases with the same enzymatic function (EC number 4.2.1.1) but with different folds; 1DMX is a flat beta sheet, while 1THJ os a left-handed beta helix. The figures were taken from the PDB database([20] and [12]

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), i.e. L2=4. SPs classify correctly, to some level of specificity, 8 pairs out of the 13 sets, 5 of which are classified completely (the full EC number).

It should be noted that in all the examples of Table 3.2 the sequence similarity of the pairs of enzymes is very small. The Smith-Waterman similarity test was run on the aforementioned pairs with the same parameters as in the previous section, resulting in an average score of 43 ± 25, compared to 90 ± 19 from section 4.2. The two score distributions do not overlap, being $\sim$2 standard deviations apart. 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 [18].

EC	Swiss-Prot1	Fold1	L1	Swiss-Prot2	Fold2	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	-

Table 3.2: 13 sets of functionally convergent enzymes from [18]. Each row contains a pair of enzymes sharing the same function. In rows 8 and 9 is a triplet of enzymes. For each enzyme we quote the Swiss-Prot identification, the fold number from the SCOP database and the EC level to which we were able to classify it using SPs (L1, L2). A - means that no SPs were found on the enzyme.

Chapter 4

SPECIFIC PEPTIDES CONTAIN BIOLOGICALLY IMPORTANT FEATURES

Some areas of the enzyme's sequence are crucial to its performance. For example, it is obvious that the mutation of an active site will completely destroy the enzyme's ability to catalyse its intended chemical reaction, and with it its *raison d'être*. Therefore sequence regions that are crucial to an enzyme will be under *evolutionary pressure*, i.e. more conserved than other regions. Thus over-represented sequence motifs such as SPs may be presumed to be of biological importance. It is of interest to try and establish particular biological roles for the SPs.

4.1 Coverage of active sites

We start by enquiring how many of the known active and binding sites are located on SPs, and what is the percentage of SPs involved in hosting them. Out of all enzymes in Swiss-Prot release 48.3, 42% have annotations of loci of active sites and binding sites. For simplicity we will refer to both annotations, which are always indications of single amino acids, as active sites. An explanation on the Swiss-Prot format and its annotations is available in Appendix D. Given these loci we find that 65% of all active sites are covered by SPs. This can be compared with the coverage of random positions on enzyme sequences which, on average, is only 27%, being 80 standard deviations away.

If an active site annotation is covered by an SP on a given enzyme, it is

probable that the active sites on other enzymes belonging to the same EC class will be covered too, due to the high levels of homology. This is apparent in Figure 4-1. The results for coverage of active sites by SPs may therefore be misleading. In order to estimate the statistical significance of these results more rigourously (see Methods) we construct a non-redundant set by choosing only one enzyme for each 4-component EC class. The results, displayed in Table 4.1, show some differences between the total and the non-redundant set. Note the high significance of these results, and the estimate that about 12% of the relevant SPs (i.e. those that occur on the queried enzymes) hit active sites. In both data sets the score is given in number of standard deviations since the p-value is smaller thanthe smallest positive normalized floating-point in MATLAB.

dataset	#enzymes	sites hit by SPs	random sites hit by SPs	score in STDs	#SPs	SPs hitting sites
all	21,228	65%	27%	80	26,931	8%
non redundant	582	52%	21%	33	6,660	12%

Table 4.1: Occurrence of SPs on active sites. Analysis has been carried out on enzymes that have an active (or binding) site annotation with SPs occurring on them. The first column states the data set used (*all* being the total set of enzymes in Swiss-Prot 48.3 and *non-redundant* standing for a non-redundant set in which a single enzyme was chosen for each EC class (see Methods)). The next column displays the number of enzymes in each data set. In the third column is the percentage of annotated active / binding sites covered by SPs. Next is the average percentage of random sites covered by SPs, followed by the score calculated in standard deviations (STDs). For both results the p-value is smaller than the smallest positive normalized floating-point in MATLAB. The next column displays the number of SPs occurring on the given data set, and the last column displays the percentage of these SPs that cover annotated active / binding sites.

As an example of these features in the data we display in Figure 4-1 aligned sub-sequences of enzymes, belonging to the same 3rd level but to two different 4th levels of the EC hierarchy: 6 out of 35 enzymes of 5.1.3.2 and 7 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 red highlighted substrings denoting SPs from our lists. Whereas in 5.1.3.20 most active sites are

```
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 4-1: Aligned sub-sequences of two different groups of enzymes of level 4 that share the same 3rd level assignment. The organisms in the upper group, 5.1.3.20, belong to proteobacteria, while those of the lower group, 5.1.3.2, contain also eukaryotes (ARATH, CYATE and PEA). Red-highlighted 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.

covered by SPs, this is not the case for the active site of 5.1.3.2.

4.2 Coverage of all annotated biological features

We extend the previous analysis to cover most annotated features in Swiss-Prot, and present the results in Table 4.2.

Amongst the most impressive results are the SP coverage of DNA binding annotations (DNA_BIND with 79%), of nucleotide phosphate-binding annotations (NP_BIND with 75%) and of annotations of short sequence motifs of biological interest (MOTIF with 71%). Next come the active sites, binding sites and metal binding sites (ACT_SITE, BINDING, METAL) whose coverage is quite impressive too, when considering the large amounts of features in the data. To evaluate the significance of the coverage of a certain feature, we use again the non-redundant

Feature	#features	coverage	cov NR	cov rand NR	score in STDs (pval)
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	28305	64%	55%	21%	30.8 (p=0)
BINDING	22429	64%	45%	22%	16.0 (p=0)
METAL	38587	59%	39%	17%	23.6 (p=0)
all	113485	59%	43%	22%	34.4 (p=0)

Table 4.2: Coverage of biological function sites. The first column contains the feature annotation as it appears in Swiss-Prot. Their descriptions can be found in Appendix D. Next is the number of annotations found on the data set, next is the percentage of these annotations that were covered by SPs. Beyond the division are the calculations for the non-redundant (NR) data set: the first 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 when compared to 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.

data set as in the previous section (see Methods). Significance is evaluated only for features whose non-redundant set contains more than 100 enzymes. The non-redundant calculation leads to an over-all coverage of 43% of all features, with significance of 34.4 standard deviations with comparison to the background model.

Apart from functional classification of novel enzymes, as seen in the previous chapter, the SPs that cover a certain biological feature may be used to identify the possible location of the said feature on a novel enzyme.

4.3 SP coverage and classification of DNA binding regions

We further investigate the SP coverage of one of the features mentioned above, DNA_BIND (DNA binding regions). This feature consists of a sequence of usually 21 amino-acids, and is mainly observed in hydrolases, 3.4.21.88. We have analyzed all enzymes in this EC number that possess the DNA binding region annotation. Figure 4-2 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 and quite low in the middle. This sort of positional preference has not been found in other features that are annotations of more than one amino acid (i.e. CA_BIND, PEPTIDE and ZN_FING).

Another interesting result is obtained by looking at the SPs occurring on enzymes of the different sub-classes of bacteria. Table 4.3 shows the sets of SPs observed on proteobacteria of the types α, β, γ and others. The interesting point to be made here is that the sets of SPs clearly allow for sub-classification 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.

Type	#Enzymes		Sets of SPs
α	17	16	KSGIHR, PSFDEMK, SKSGIHRLI
		1	KSGIHR, SKSGIHRLI
β	8	4	GFRSPNAAE, PPTRAEI
		3	NAAEEHL, PPTRAEI
		1	PPTRAEI
γ	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

Table 4.3: Sub-classification of DNA-binding bacteria according to the SPs that cover the binding region. 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. The sets of SPs occurring on the three different classes of bacteria are disjoint.

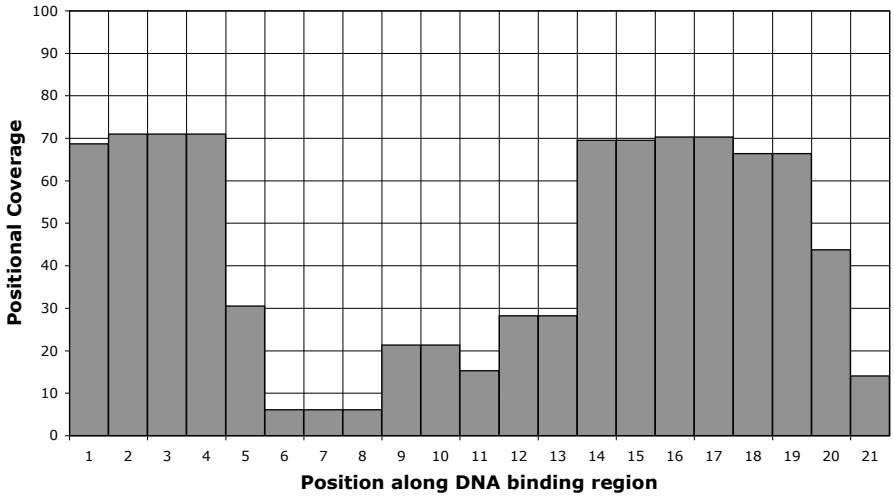

Figure 4-2: Positional coverage of DNA binding regions (as annotated in Swiss-Prot) by SPs. The x axis represents the position from the beginning of the DNA binding region. The y axis denotes the percentage of DNA binding regions covered by SPs per position. A preference of SPs to occur at the starting point and at the end of the feature is apparent.

Chapter 5

NOVEL BIOLOGICAL FEATURES

5.1 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 one of its amino acids by mutagenesis, and looking for changes in the enzymatic function. Amongst the Swiss-Prot annotations is one called MUTAGEN, that annotates a site which has been experimentally altered by mutagenesis. We can exploit this fact and calculate whether mutated SPs damage enzymatic function significantly more than other tested amino acids.

3.509 MUTAGEN annotations exist on our data set. Since active sites, binding sites and metal binding sites are already known to be crucial to the enzymes' performance, we eliminate MUTAGEN annotations that refer to sites that are also annotated as such. We are then left with 2,814 MUTAGEN annotations, 2,562 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 *un-successful* if not.

The size of the population is 2,814, the number of successes in the population is 2,562. From this population we pick a sample of 919 such MUTAGEN anno-

tations that are covered by SPs, 867 of which are *successful*. The hypergeometric distribution describes the probability that in a sample of n distinctive events drawn from the population (without returns) exactly k events are successful. 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.

5.2 SPs in active pockets

In the preceding chapter we saw that SPs cover sequences of amino acids of known biological importance in a statistically significant manner. It was also shown that mutated SPs, in general, tend to damage enzymatic function of enzymes significantly more than at other loci. Here we analyse SPs that do not necessarily cover annotated amino acids with known biological importance.

A certain chain of events led us to the three-dimensional analysis discussed in this chapter: Taking a second look at Figure 4-1, one notices that if the enzymes from both EC classes are aligned by the second active site (Y in both classes) and gaps are disregarded, the SP3 RYFNV occurs 26-27 amino acids away from the said active site. We came across a number of such examples of SPs appearing in what seemed like constant distances from active sites. This led to the conclusion that such SPs might take part in the catalytic activity of active sites. The obvious place to look for such cooperation is in the three-dimensional structure of enzymes.

Figure 5-1 shows the relationship between SPs and spatial structure. The active site and two binding sites appear close in the structure, creating a catalytic area, even though they occur hundreds of amino acids apart on the sequence. This enzyme contains many SPs. Two overlapping SPs cover the active site and lie along the catalytic area. Another one - HMVRNI - lies along the other side of the catalytic area. The active and binding sites are cradled by SPs, some of which are far away on the sequence. These SPs share a pocket with the active site and the two binding sites, and one suspects they may play a role in the catalytic

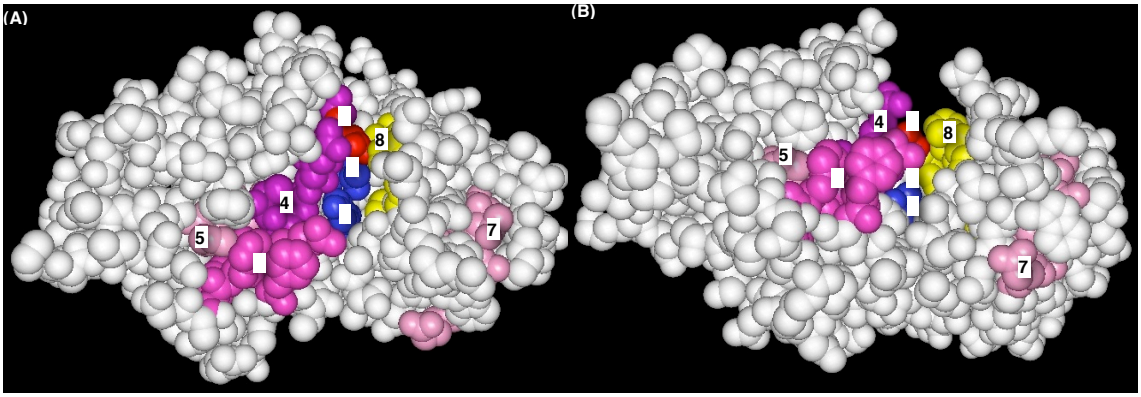

Figure 5-1: (a) Three-dimensional display of enzyme P07649, belonging to 5.4.99.12, showing [1] an active site D at sequence location 60, [2] a binding site Y at location 118, [3] a binding site L at 245. The active site is common to two overlapping SPs [4] (CAGRT(D)AGVH). Other shown SPs are [5] GQVVH 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. GQVVH and ENDFTS belong to SP3, all other motifs to SP4. (b) A different perspective of the same enzyme emphasising the pocket containing the active sites and cradling SPs.

activity.

One is naturally tempted to assign importance to SPs with high solvent-accessibility and in proximity of active sites. High solvent-accessibility means that the SP lies on the surface of the enzyme's structure, and therefore can interact with external substrates. The proximity to an active site reinforces the argument that the SP may play a role in the catalytic activity of the enzyme. SPs that automatically obey these demands are those that reside in the pockets of active sites in the spatial structures of enzymes.

For this study we use the CASTp [9] database, that runs a geometric algorithm on spatial structures listed in the PDB, identifying their pockets. Each structure has a list of pockets and the amino-acids composing them. The interest is directed only, for the moment, towards *active pockets* - pockets that contain an active or binding site amongst their constituting amino-acids. We define an SP as lying within an active pocket if at least four of its amino-acids reside in the pocket (i.e. are amongst the amino-acids constituting the active pocket). Figure 5-2 illustrates an example of an active pocket both in structure and in sequence. SPs

are highlighted on the sequence, demonstrating the definition of SPs residing in active pockets.

We select 1031 enzymes that possess such active pockets. There are 8860 SPs that occur on them, 28% of which lie within these active pockets. Defining a background model (see Methods) of random peptides selected for each event of an SP hitting a particular enzyme, we estimate that 18% of all SPs belong to events that pass an FDR limit [8] of 0.05, i.e. are statistically significant events. Most of them (88%) do not contain an active site, and have not yet been studied experimentally. Table 5.1 presents the these results.

#enzymes	#SP	#SPs in pockets	#Significant SPs FDR=0.05	#Significant SPs without site
1031	8860	2487 (28%)	1622 (18%)	1426

Table 5.1: Occurrence of SPs in spatial proximity to active sites. This is an analysis of 1031 enzymes whose spatial structure is known (in PDB) and possess three-dimensional pockets that include active sites (using CASTp [9]). The table lists the number of enzymes that were analysed 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 in the next column, the number and percentage of SPs whose events passed the FDR test. The last column shows the number of statistically significant SPs that do not contain the active or binding site.

5.3 SPs in pockets of other biological features

Not only active sites tend to reside in pockets; other biological features do so too, such as metal binding sites, nucleotide phosphate-binding regions and more. In light of the interesting results for pockets of active sites, we perform the same analysis on pockets of the other biological features (as seen in Chapter 3). Table 5.2 illustrates results of statistically significant occurrences of SPs in pockets of SwissProt annotated features. Again, these statistically significant occurrences of SPs in pockets of different SwissProt annotated features may have crucial biological roles in the performance of the enzymatic functions.

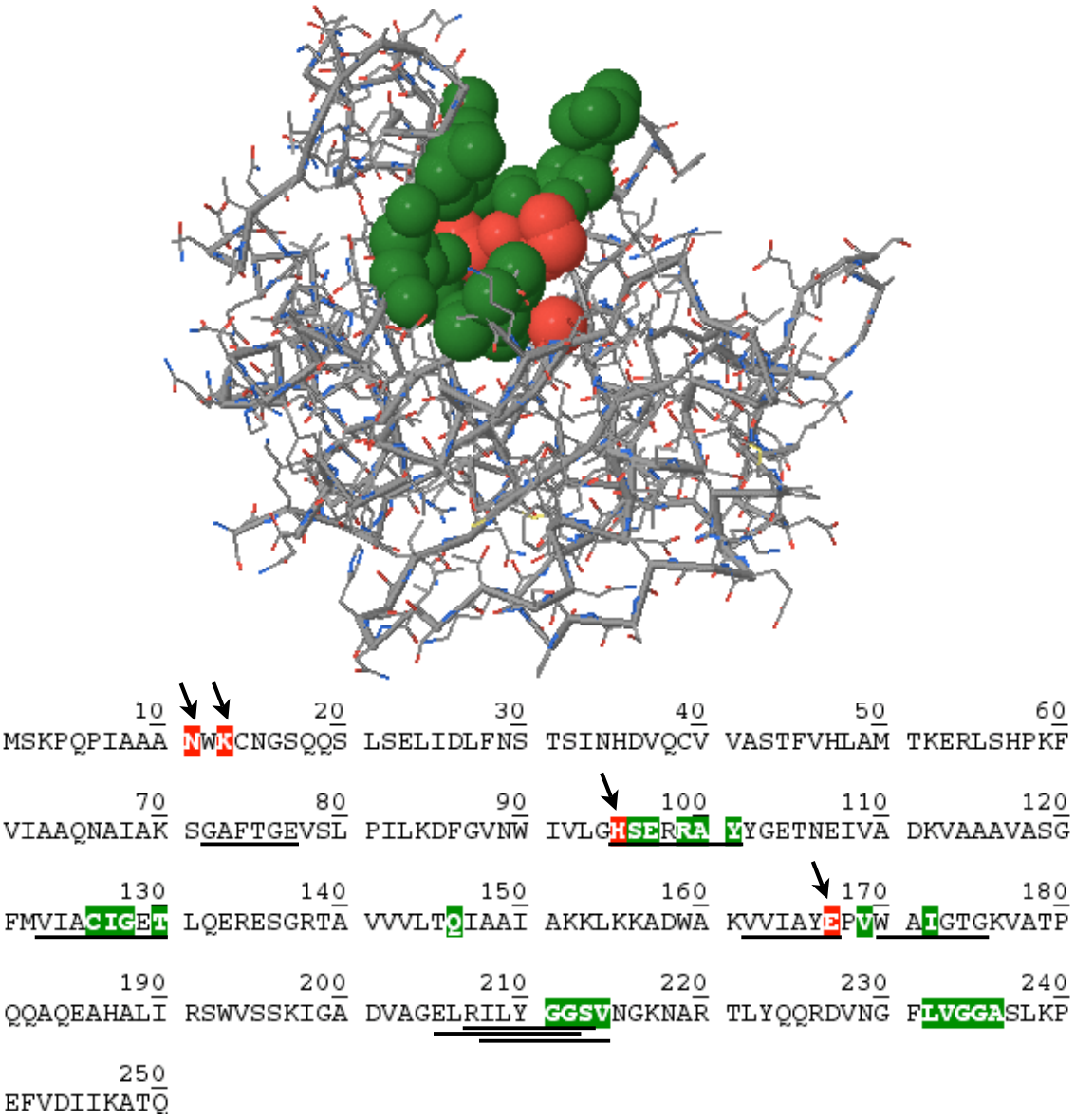

Figure 5-2: Example of an active pocket in enzyme 1AG1O (PDB id), drawn using JMOL (www.jmol.org). The atoms of the amino acids belonging to the active pocket are shown in red and green on the backbone spatial structure (above). The same amino-acids are highlighted in red and green in the sequence (below). The active sites are in red and the rest of the amino acids constituting the pocket are in green. The SPs appearing on this enzyme are underlined in the sequence. It can be noted that none of the amino acids of the first SP, GAFTGE, belong to the active pocket. The second SP, HSERRAY, covers the active site, H, and contains 5 more amino acids belonging to the active pocket (and therefore is said to lie in the active pocket). Three partially ovelapping SPs appear around location 210. Only one of them, ILYGGSV, is said to reside in the pocket.

Feature	#Significant SPs in pockets	#Significant SPs without feature	#Feature pockets with SPs
NP_BIND	1,246	1,103	135
MOTIF	116	92	16
SITE	263	258	28
CARBOHYD	56	44	7
ACT_SITE	1,057	1,024	116
BINDING	945	828	95
METAL	867	751	96
all	2,919	2,300	329

Table 5.2: Occurrence of SPs in spatial proximity to SwissProt annotated features. This is an analysis of enzymes whose spatial structure is known (in PDB) and possess 3D pockets that include any SwissProt feature annotation (using CASTp [9]). For each feature, the table shows the number of SPs lying (with at least four residues) in pockets including the given feature's residues. Requiring high significance of the latter, through a background model described in Methods, and using the FDR limit of 0.05, we obtain the SPs in the first column. The second column shows the number of these SPs that do not contain the feature, and therefore without experimental verification. In the last column is the number of enzymes which have a pocket with the said feature, and that have an SP lying in it.

5.4 Glycine-enriched SPs in active pockets

Here we discuss a novel finding concerning SPs located in active pockets, whose statistical significance was determined. Comparing the relative frequencies of all amino-acids occurring on these SPs with the frequencies observed on enzymes in general one finds a clear over-representation of glycine. Table 5.3 compares the glycine frequency on enzymes with that on SPs in general and with SPs whose occurrence in active pockets is statistically significant. It turns out that it is highest for those SPs that lie in these pockets, regardless of whether they contain the active site or not. It should be stressed that amongst the amino acids that constitute the active pockets, glycine frequency is normal (as in all the data set), emphasising that the glycine enrichment is a specific characteristic of the statistically significant occurrences of SPs in active pockets. Glycine frequency is normal also for amino acids constituting other pockets.

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

Data set	Glycine frequency
all enzymes	7.5%
all SPs	9.2%
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	8.6%

Table 5.3: Frequencies of the glycine amino acid in various data sets. *SPs in active pockets* refers to SPs whose occurrence in active pockets is statistically significant (see Methods). The p-values in rows 3 and 4 refer to a comparison with frequency distributions of amino-acids in all enzymes (see Methods). The glycine frequency on all enzymes is compared to that on SPs in general and SPs in active pockets. Significance is calculated for SPs that lie in these pockets and do not cover the active site. Glycine frequency is normal (8.6%) amongst the amino acids that constitute the active pockets as a whole.

contributes to packing of nearby residues. It is generally accepted [38] that the location of glycines in the structure of a protein influences its motion.

One can asses whether the glycine enriched SPs tend to appear on turns and hinges by checking on which secondary structures they appear. As explained in the Introduction, the most common secondary structures are α-helices, β-sheets and turns. SwissProt annotates the amino acids that take part in these secondary structures, in enzymes for which the three dimensional structure is known. Table 5.4 shows the analysis of the characteristic secondary structure of our SPs, with glycines and without. Results show that the glycine enriched SPs tend to occur on turns and un-annotated secondary structures, both compared to a random model and compared to the SPs that are not glycine enriched. Turns and unannotated secondary structures obviously need to be flexible, while α-helices and β-sheets are pretty rigid. An example of a glycine-enriched SP appearing in an active pocket, on a sharp bend of an enzyme structure is shown in Figure 5-3.

Following the reasoning according to which glycine influences protein motion and bestows rotational felxibility, [44] have examined 23 enzymes and suggested that glycine residues may provide flexibility to active pockets in enzymes.This is in line with the *induced fit* model that has been proposed by [21]. According to this model active pockets are flexible and go through conformational changes when

			At least 1 AA belongs to a secondary structure			
data	#SPs		none	turn	helix	strand
At least 1 Gly	795	#SPs	583	443	281	345
		#rand	523	353	274	373
		p-val	3.1e-09	0	0.29	0.02

Table 5.4: Glycine enriched SPs' secondary structures (turn, α-helix, β-sheet or no annotation). The data set analysed in this table is SPs whose occurrences in active pockets were statistically significant and that consist of at least one glycine. The number of SPs is presented in the second column. An SP may take part in different secondary structures (one demands that at least one of the amino acids of the SP be annotated as taking part in a certain secondary structure). Thus, for example, the 281 SPs that are shown to take part in α-helices each have at least one amino acid annotated as occurring on an α-helix, but some of these SPs may also occur on β-sheets or turns or have no annotation. One also performs the same count for randomly picked SPs (of the same length and number), presenting here the average count. The p-value is calculated for each test. 0 is marked where the calculated p-value is smaller than the smallest positive normalized floating-point in MATLAB. The results show that glycine enriched SPs tend to appear on turns and un-annotated secondary structures significantly more than expected by random motifs.

Figure 5-3: Example of a glycine enriched SP in an active pocket. A schematic structure of 1B2M, with the SP *ASGNNF* in light blue. The SP lies on a sharp bend in the structure, the glycine right in the middle of it. The figure was made using RasMol [34]

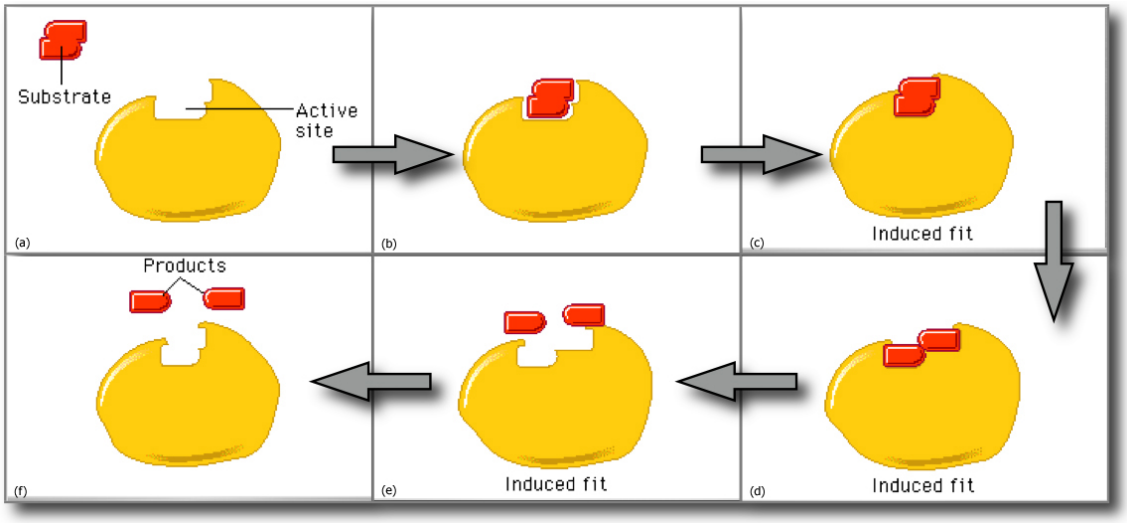

Figure 5-4: The *induced fit* mechanism is shown in its various stages. (a) the active pocket is in an open state, (b) a substrate enters the active pocket, (c) the active pocket changes its conformation to a closed state, *hugging* the substrate, (d) the catalysis takes place, and may change the substrate's conformation, and therefore also the active pocket's conformation, (e) the products leave the active pocket and (f) the active pocket changes it conformation back to the original one. Pictures from www.phschool.com.

binding to a substrate: they have an open form, where the solvent-accessibility is increased revealing the active site, and a closed form when the substrate is bound. This enables both improved catalysis (since excess water molecules may be squeezed out) and permits multiple selectivity of the enzyme. The mechanism is illustrated in Figure 5-4

In addition to the observation of increased glycine content in active pockets [44], we note that the strategic location of glycine has been suggested as a mechanism for achieving an induced fit for some specific enzymes [31], [39], [28]. It has been concluded [44], [31] that fluctuations at or in close proximity to the active site can represent an induced fit mechanism, and that the observed fluctuations are associated with the location of glycines in the protein structure. From Table 5.3 we learn that significant glycine enrichment is observed on SPs residing in active pockets. Thus we may hypothesize that the latter are responsible for the induced fit mechanism.

At this point it might be appropriate to mention that single site mutations in-

Feature	SPs in feature pockets			SPs in feature pockets not containing feature			SPs in feature pockets containing feature		
	#SPs	%Gly	pvalue	#SPs	%Gly	pvalue	#SPs	%Gly	pvalue
NP_BIND	1,246	12.3	1.2e-04	1,103	11.4	5.7e-04	152	19.28	3.1e-05
METAL	867	10.6	2.6e-02	751	10.3	2.9e-02	117	12.4	7.6e-02
MOTIF	116	9.6	3.6e-01	92	9.2	4.9e-01	24	11.36	2.1e-01
SITE	263	10.4	3.7e-02	258	10.6	2.4e-02	5	2.33	-
CARBOHYD	56	8.9	5.7e-01	44	10.0	3.1e-01	12	5.1	-

Table 5.5: Glycine statistics of SPs in pockets of SwissProt annotated features that appear in pockets. Three types of data sets are analysed: SPs residing in the pocket of a given SwissProt annotated feature, SPs residing in the pocket of the feature that do not contain the actual feature, and last of all, SPs in the pocket that do contain the annotated feature. P-values are calculated only for sets of at least 20 SPs. Glycine enrichment is apparent in SPs that occur in pockets of the NP_BIND feature and METAL feature.

volving glycines, i.e mutating glycines or changing non-glycine to glycine residues, can be lethal. Such mutations always affect protein stability [11, 35], cause changes in specificity [41], and are responsible for about 15% of human genetic diseases [42].

It is also interesting to see whether SPs that reside in pockets of other biological features are glycine enriched too. Table 5.5 displays the results for SwissProt annotated features that may appear in pockets: SPs in pockets of NP_BIND annotations are generally very glycine-enriched, especially those containing the feature. SPs in pockets of METAL annotations are less glycine-enriched than in active pockets, but are still interesting. SPs in pockets of the remaining features (MOTIF, SITE and CARBOHYD) do not show significant glycine enrichment.

Chapter 6

DISCUSSION

This study substantiates the importance of Specific Peptides both as classification tools and especially as biologically relevant functional elements.

Conventional classification methods rely on sequence or structure similarity. In Chapter 3 we introduced extreme classification problems, where straightforward sequence or structure similarity analysis may lead to wrong conclusions. While sequence similarity is also at the root of most SPs of level 4 (see some examples in Figure 4-1), we have demonstrated the role of SPs as carriers of information in those extreme situations. SPs were successful in the classification of enzymes with evolutionarily convergent functions (extreme cases of remote homology, the enzymes differing in sequence *and* structure). SPs classified successfully both non-homologous enzymes with high sequence similarity and enzymes with evolutionarily divergent functions, such as the TIM barrel enzymes. This may be attributed to the fact that SPs are deterministic motifs and not forms of general expressions.

The relevance of SPs to biological functions was evaluated in Chapter 4, by finding the coverage of amino acids that are known to be crucial to these functions, such as active-sites, metal binding sites, Ca-binding sites, etc. Many of the functional annotations are well covered by SPs. The statistical significance of the observed coverage was also evaluated on non-redundant data sets. The results are extremely significant.

In the case of DNA binding regions we found that the coverage by SPs is

peaked at the beginning and at the end of the region. We also discovered that these SPs allow for sub-classification of the enzymes of the relevant bacteria into phylogenetic classes. This is quite natural since SPs are highly conserved sections of proteins that belong to different species and share the same EC classification.

Chapter 5 discussed possible biological roles of SPs that do not contain annotated sites known to be crucial to the enzymatic function. In order to verify the biological importance of any individual SP one should perform mutations of the different amino acids of the SP, testing how crucial the SP is to the function of the enzyme. In our large-scale study, we have checked the existing MUTAGEN annotations in the Swiss-Prot database. In doing so we tested a large set of SPs that do not possess known annotations. The number of affecting mutagenesis results occurring on these SPs was compared to the affecting mutageneses in all the data set. Once again the significant p-value confirms the fact that SPs are biologically important.

After assessing that SPs cover sites of known biological importance, and that they affect the the enzyme's function, it is of interest to predict biological roles of previously un-annotated sites. All SPs that reside in active pockets, do not contain annotated sites, and are statistically significant are such novel structures. Their three-dimensional vicinity to a crucial site, and high solvent accessibility, are indications of their expected biological importance. An analysis of such a set of SPs has lead to the following interesting observation: SPs that reside in active pockets and do not contain the actual active or binding site, are significantly enriched with glycine. This holds even when compared to the distribution of all residues in active pockets. We suggested the interpretation that these SPs are responsible for the induced fit mechanism in these enzymes.

Chapter 7

METHODS

7.1 Data set

In this work the data set is the same as that used in [22]. Protein sequences annotated with EC numbers were extracted from the Swiss-Prot database (Release 48.3, 25-Oct-2005). To obtain a high-quality, well-defined training data set, the data was strictly screened and the following sequences were removed: sequences shorter than 100 amino acids or longer than 1200 amino acids, sequences with uncertain annotation, and enzymes that catalyse more than one reaction (e.g. have more than one EC number).

7.2 SP sets

MEX was separately applied [22] to each one of the six enzyme classes, with the parameters $\eta = 0.9$ and $\alpha = 0.01$. The graph vertices of MEX represent the 20 amino acids that comprise the enzyme sequences. The resulting motifs, or peptides, were classified according to their EC classification specificity, resulting in Specific Peptide (SPs). An SP4 is a peptide that occurs only on enzymes belonging to the same four-component EC classification (e.g. only enzymes of EC 5.2.3.20). An SP3 is a peptide that occurs only on enzymes belonging to the the same 3-component EC classification (e.g. enzymes of EC 5.3.2.20, 5.3.2.1, 5.3.2.4 and so on, but only of EC 5.2.3.-). In the same fashion one defines SP2

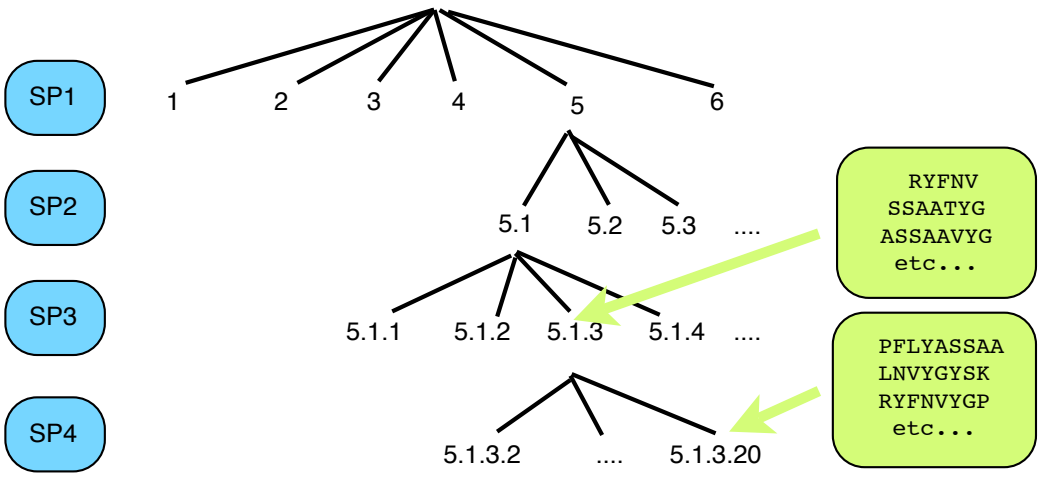

Figure 7-1: The SP (specific peptide) definition: RYFNV is an SP3 since it occurs on sequences of enzymes with the same first three components of the EC number (5.1.3.-) but with a different fourth component. On the other hand RYFNVYGP is specific to a level 4 EC number (5.1.3.20), and therefore is an SP4.

and SP1. This is represented in Figure 7-1. These SPs were filtered to obtain a non-redundant set (i.e. if within an SPN group one motif contained another, the shorter one was kept).

42,874 SPs have been extracted that specify the full EC number, i.e. correspond to the fourth level of the EC hierarchy and denoted as the SP4 set. Other SPs divide into 2,945 in the SP3 set, 1,159 in the SP2 set and 5,414 in the SP1 set, the latter specifying the enzyme class (one out of six classes). We employ these sets of SPs in our analysis. 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 9.5 SPs appear on an enzyme. Obviously their EC assignments have to be consistent with one another.

7.3 Statistical significance of SP coverage of annotated features

In Chapter 4 we calculate the SP coverage of a given Swiss-Prot annotated feature. Let us clarify this by considering an example of the DNA_BIND feature in Figure

D-1 in Appendix D. The 20-amino acid-long annotation starts on the 28th amino acid in the sequence and ends on the 48th amino acid (highlighted in red). Two SPs are found to overlap with the feature, therefore this annotation is said to by covered. In the same manner the coverage of all DNA_BIND appearances is assessed.

To analyse the statistical significance of the results we compare them to the expected value of a background model. The latter is defined by labelling randomly picked strings from the enzyme's sequence as pseudo-features, equivalent in size and number to the occurrences of the real features.

According to the central limit theorem, the distribution of a sum of a large number of independent variables is approximately normal. The rule of thumb is that a sample size of at least 30 will suffice. Therefore the background model is run 30 times, enabling the calculation of an expected value and standard deviation. In our example from Figure D-1, 30 different sub-sequences (each 20 amino acids long) are randomly chosen from the enzyme sequence, and for each the coverage is checked. The procedure is carried out on all the enzymes on which the DNA_BIND feature occurs.

The average number of covered annotations and standard deviation define the normal distribution of the background model. A p-value is calculated as the probability of observing at least as many covered annotations as observed in reality, in the background model.

Since enzymes belonging to the same EC class share high homology, if an active site is covered by an SP on one enzyme it is probably covered also on the other enzymes of the class (as can be seen in Figure 4-1). Coverage results may be biased and therefore the statistical significance is calculated also on a non-redundant set of enzymes (enzymes without high sequence similarity). The set is constructed by finding all enzymes that include the said feature, and choosing one arbitrary enzyme for each EC number.

7.4 Calculation of the p-value for the mutagen analysis

Chapter 5 includes an analysis of the effect on enzymatic function of SPs whose sites have been experimentally altered by mutagenesis.

A hypergeometric distribution is encountered. This is a discrete probability distribution that describes the number of successes in a sequence of n draws form a finite population without replacement.

The size of the population (i.e. the number of MUTAGEN annotations that do not coincide with active, binding or metal binding sites) is defined as N. The number of successes in the population (i.e. the number of MUTAGENs that represent mutations that damaged the enzymatic function) is defined as D. The hypergeometric distribution describes the probability that in a sample of n distinctive events drawn from the population (the MUTAGEN annotations that are covered by an SP) exactly k events are successful. The said probability for k=X successful events is given by:

$$Pr(k = X) = f(k; N, D, n) = \frac{\binom{D}{k}\binom{N-D}{n-k}}{\binom{N}{n}} \tag{7.1}$$

In this case we have X = 1,197, N = 3,509, D = 3,254 and n = 1,251. The p-value is the cumulative of this distribution, i.e $Pr(k \geq 1,197)$.

7.5 Statistical significance of SPs residing in active pockets

In Chapter 5 we define an *active pocket* as a pocket in the three-dimensional 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 also constitute the active pocket. Figure 5-2 shows an example of the structure of an enzyme, demonstrating an active pocket, the amino acids belonging to it and SPs residing

in it. The data regarding the amino acids that constitute pockets was taken from the CASTp database [9].

Since different numbers of SPs of different lengths appear on different enzymes, one has to define the random variable, the statistical event, as the occurrence of a given SP within an active pocket in a given enzyme.

Let us define N as the number of different SPs occurring on an enzyme's sequence. We assign to each SP a random variable representing a Bernoulli trial: the possible outcomes of a randomly picked motif of the same length as the SP:

$$X_i = \{1, 0\}; \quad \forall i \in \{1, 2, ..., N\} \tag{7.2}$$

If the randomly picked motif occurs in an active pocket, it is called a *success* and the appropriate random variable is assigned 1. If the randomly picked motif does not occur in an active pocket, it is assigned 0. The probability of success of each variable is defined as the number of possible motifs on the sequence that occur on an active pocket divided by the number of all possible motifs on the sequence. Let us define the probabilities of *success* for each random variable as:

$$p_i \in [0, 1]; \quad \forall i \in \{1, 2..., N\} \tag{7.3}$$

It should be noted that it is possible for one such probability to be equal to 0 (if the motif isn't long enough to cover four amino acids belonging to an active pocket). Let us now define a new random variable, describing the number of *successes* or the number of motifs occurring in an active pockets our of N randomly picked motifs (equal in number and size to the SPs appearing on the enzyme sequence):

$$Z = \sum_{i=1,2,...,N} X_i \tag{7.4}$$

For a given event of an SP occurring in an active pocket, the p-value is calculated as the probability that at least one of N randomly picked motifs occurs in

an active pocket, i.e.

$$\text{p-value} = Prob(Z \geq 1) = 1 - Prob(Z = 0) = \prod_{i=1,2,\ldots,N} (1 - p_i) \qquad (7.5)$$

P-values are calculated for all events of SPs in active pockets, creating a multiple-hypothesis problem. This occurs when a number of independent observations are filtered using the same acceptance criterion (i.e the same p-value criterion) that one would use when considering a single event. A single event's acceptance criterion requires that the observed data be highly unlikely under a background model. A large number of independent observations, subject to the same acceptance criteria, outweighs the original high unlikelihood associated with each individual test. It becomes increasingly likely that one will observe data that satisfies the acceptance criterion by chance. These errors are called false positives because they positively identify a set of observations as satisfying the acceptance criterion while that data in fact represents the null hypothesis. False discovery rate (FDR) control [8] is a statistical method used in multiple hypothesis testing to correct this phenomenon. Significant events are selected here according to an FDR limit of 0.05.

7.6 Statistical significance of differences between distributions

Tables 5.3 and 5.5 in Chapter 5 compare two amino acid frequency distributions. Let us define the frequency of each amino acid in all enzymes as X_i (i=A, C, D...W, Y). The corresponding frequency in a certain set of SPs is defined 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 of these differences.

The statistical significance of the difference between the frequency of a certain amino acid in a given SP set and its frequency in all the data is evaluated by calculating a p-value. This is done by comparing the difference in frequencies for

a certain amino acid to the average difference.

Appendix A

THE MEX ALGORITHM

MEX (Motif Extraction Algorithm, [37]) is an unsupervised algorithm that extracts statistically significant motifs from a given set of data. The algorithm is data driven, meaning that the motifs that it finds are not necessarily over-represented in the data. This fact gives it an advantage on other motif extraction algorithms, usually based on sequence similarity (and therefore missing motifs that are not over-represented). Originally, MEX emerged from the a linguistic context, and since spoken language is intuitive, it might be a good idea to describe MEX in this context.

Let us define the problem our algorithm confronts: given a corpus with all word delimiters removed (such as spaces, and all forms of punctuation), it must fathom the original words that assemble the corpus. The algorithm uses a directed graph: the vertices are composed of the letters of the alphabet and begin and end vertices. V = {a, b, c, ..., y, z, *begin*, *end*}. A set of ordered pairs of vertices, or edges, represent the order in which the letters appear in the corpus, i.e. the edge e(t, h), connecting between the vertices t and h, represents that fact that h appeared after t at some point in the corpus. The first sentence in the corpus is loaded onto the directed graph, connecting the vertices with directed edges, starting at the *begin* vertex and ending at the *end* vertex. This procedure is illustrated in Figure A-1, and is repeated for all sentences of the corpus.

As can be seen in Figure A-1 (d), the edges connecting the vertices a, l, i, c and e, consecutively, seem to form a bundle; edges seem to converge towards the

Vertex	Conditional probability expression	Conditional probability
a	P(a) = 8,770 / 109,625	0.08
l	P(al|a) = 1,046 / 8,770	0.12
i	P(ali|al) = 486 / 1,046	0.45
c	P(alic|ali) = 397 / 486	0.85
e	P(alice|alic) = 397 / 397	1
w	P(alicew|alice) = 48 / 397	0.12
a	P(alicewa|alicew) = 21 / 48	0.44
a	P(alicewas|alicewa) = 17 / 21	0.81
b	P(alicewasb|alicewas) = 1 / 17	0.12
e	P(alicewasbe|alicewasb) = 1 / 1	1
g	P(alicewasbeg|alicewasbe) = 1 / 1	1

Table A.1: The first column shows which vertex were looking at, the second column shows what conditional probability were interested in, and its calculation. The last column shows the final probability after calculation.

vertex *a*, walk together through *l*, *i*, *c* and *e*, and seem to diverge at the vertex *e*, as is illustrated in refFig:03. This happens since the sequence *alice* appeared in all four phrases, but in different contexts. So if one were able to recognize all such bundles, these would hypothetically lead to the words that constitute the corpus. Let us rephrase this in probabilistic terms by constructing a conditional probabilities matrix. Table A.1 shows the calculation of the matrix for a toy example, the first sentence form *Alice in Wonderland*: *Alice was beginning to get very tired of sitting by her sister on the bank and having nothing to do.* The input of MEX is: *alicewasbeginningtogetverytired.....*

In the first row of Table A.1, P(a) the probability of appearing in the corpus is calculated resulting in 0.08. In the next row we see the calculation for P(l—a) the probability of *l* appearing after *a*, giving 0.12. Continuing along the sentence the probability slowly rises, until it reaches 1 on the fifth vertex, meaning that *e* always appears after the sequence *alic*. But in the next row, there is a sharp drop in the probability: *w* appears after only 12% of the occurrences of *alice*. This is the quantitative expression of the divergence of edges we saw before, meaning that after the word *alice* any new word can begin. At a certain point in the sentence the probability will be 1 and remain so since there is only one such sentence in the corpus. These probabilities are called the right moving probabilities (P_R), since

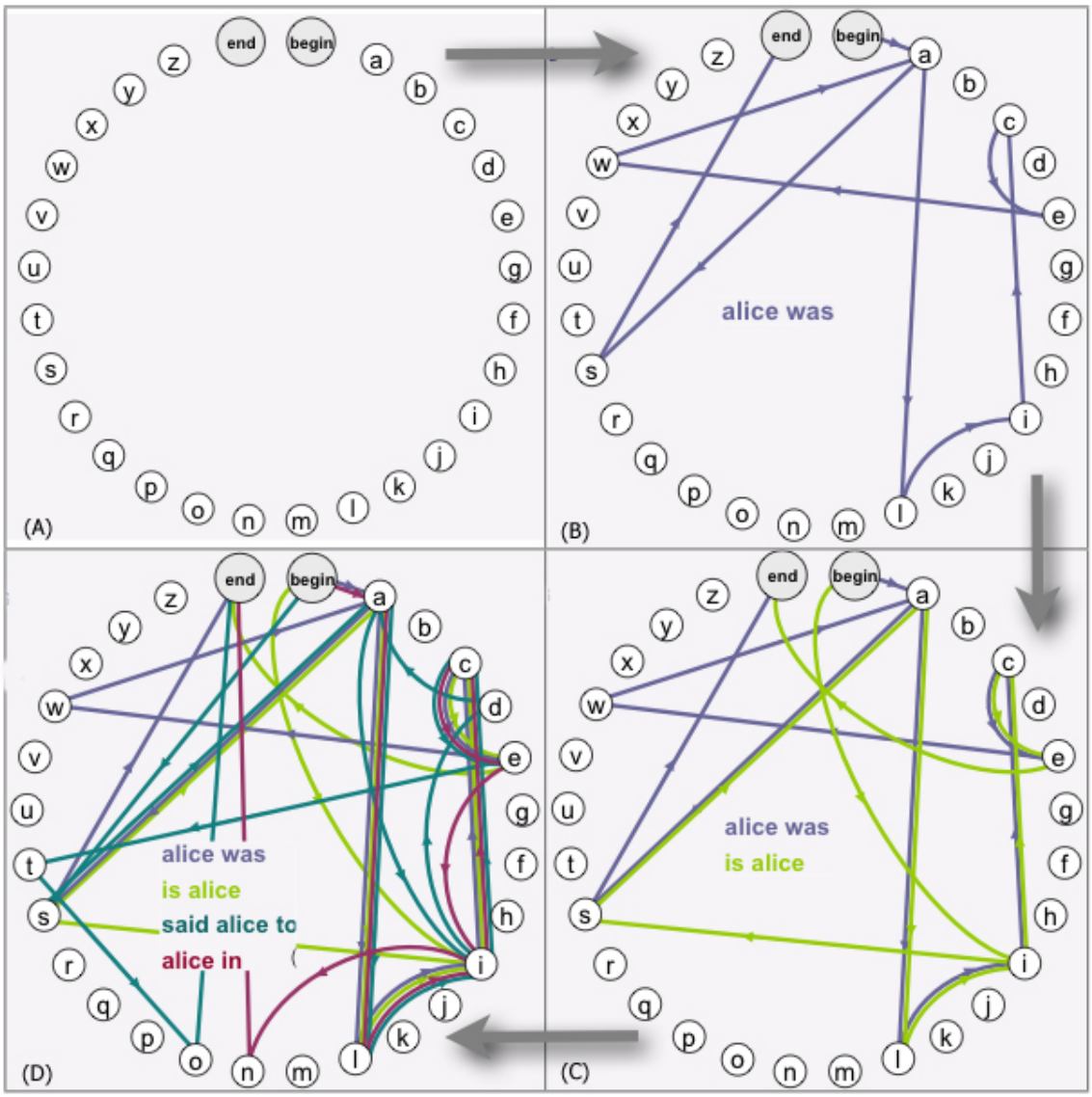

Figure A-1: Loading the directed graph: (a) the empty graph with the vertices V = {a, b, c, ..., y, z, *begin*, *end*} (b) the sentence *alice was* is loaded: the first edge connects vertex *begin* and vertex *a*, the second edge connects *a* and *l* and so on, creating a path on the graph that ends on the vertex *end* (c) the sentence *is alice* is loaded (d) all 4 sentences are loaded on the graph. The edges seem to form a bundle along the vertices *a*, *l*, *i*, *c* and *e*.

they are calculated by moving to the right along the corpus. We also calculate the left moving probabilities (P_L), the conditional probabilities in the opposite direction, from the end to the beginning. These probabilities are calculated for all positions, formally:

$$P_R(e_i; e_j) = p(e_j|e_i e_{i+1} e_{i+2} ... e_{j-1}) = \frac{l(e_i; e_j)}{l(e_i; e_{j-1})}$$

$$P_L(e_j; e_i) = p(e_j|e_{i+1} e_{i+2} ... e_{j-1} e_j) = \frac{l(e_j; e_i)}{l(e_j; e_{i+1})}$$

where e_i is vertex i, and $l(e_i; e_j)$ is the number of sub-paths connecting vertices e_i and e_j. The end of a motif is defined as the vertex in which a dramatic drop in the right moving probabilities is apparent (expressing the divergence of edges from the vertex), and the beginning of a motif as a dramatic drop in the left moving probabilities (expressing the convergence of edges to that vertex), formally; we define the drops at a given point as:

$$D_R(e_i; e_j) = P_R(e_i; e_j)/P_R(e_i; e_{j-1})$$

$$D_L(e_j; e_i) = P_L(e_j; e_i)/P_L(e_j; e_{i+1})$$

demanding $D_R(e_i; e_j) < \eta$ for the ending of the motif at vertex e_{j-1}, and $D_L(e_i; e_j) < \eta$ for the beginning of the motif at vertex e_{i+1}. η is the threshold parameter. This is illustrated in Figure A-2.

The probabilities have been calculated from finite numbers, creating a low statistics problem which may give misleading results. Another parameter is introduced, $\alpha < 1$, to take care of this: we require that the average of D_L and D_R be smaller than α.

For more details see [37] and http://adios.tau.ac.il.

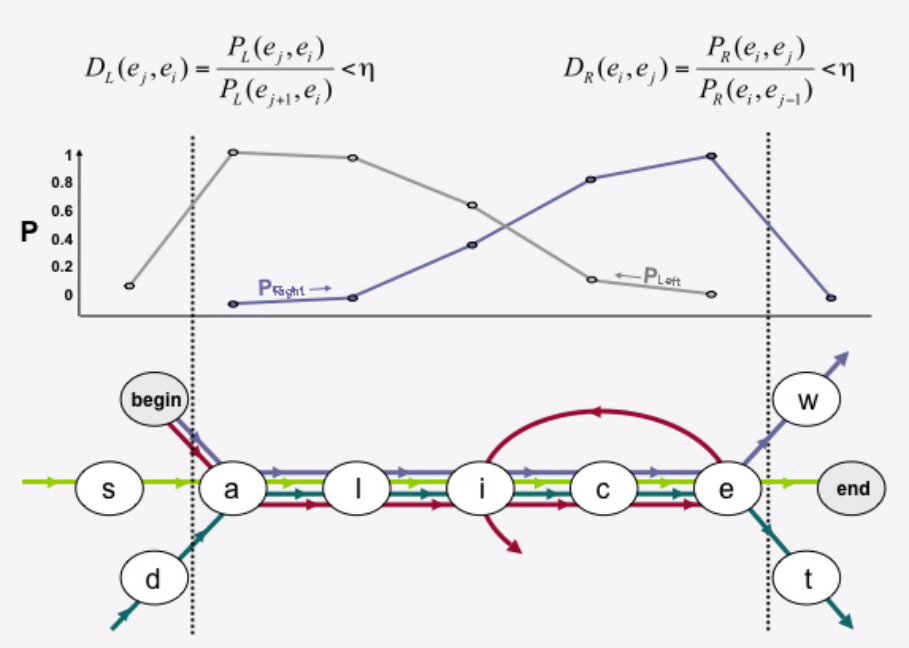

Figure A-2: An *alice* example: a drop in the right moving probability denotes the end of the motif at the vertex e, and a drop in the left moving probability denotes the beginning of the motif at the vertex a, thus detecting the motif alice.

Appendix B

MOTIF EXTRACTION: COMPARISON OF SPS TO PROSITE MOTIFS

Here we compare the SPs performance to that of ProSite motifs. SPs cover 93% of enzymes, while ProSite motifs cover only 63%. Figure B-1 demonstrates this point.

It is of interest to assess how SPs cover ProSite motifs. Since ProSite motifs are expressed as regular expressions or weight matrices, and SPs are deterministic motifs, we search for the appearance of the ProSite regular expression on a given enzyme and check how well SPs cover tit. The average length of ProSite motifs is 18 amino acids (double the average length of SPs). We define a ProSite motif to be covered by SPs if at least 40% of its amino acids also belong to an SP on that enzyme. We can then calculate what percentage of unique ProSite peptides are covered by SPs, with the given definition. To calculate the statistical significance

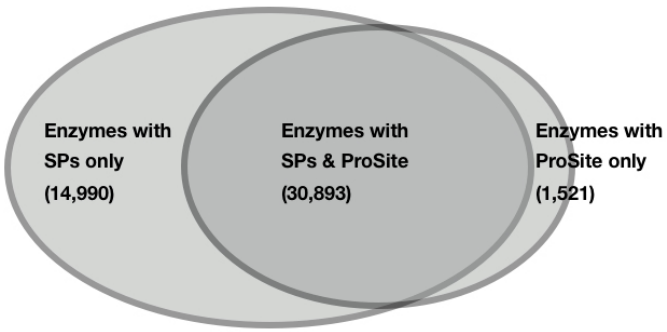

Figure B-1: Venn Diagram of the coverage of enzymes by SPs and/or ProSite motifs. SPs cover 93% of our enzyme data set, while ProSite motifs cover only 63%.

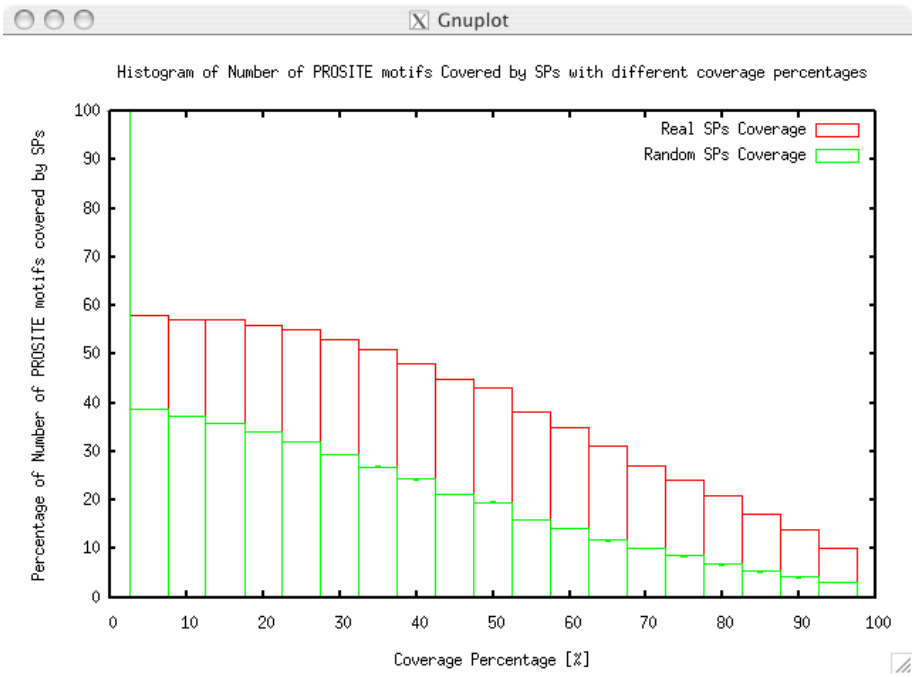

Figure B-2: Comparison of the coverage of ProSite peptides by SPs and by randomly chosen motifs, for different definitions of a covered ProSite motif (i.e. different minimal percentages of ProSite amino-acids that belong also to the motifs in question.

of these results, we calculate the coverage of ProSite motifs by randomly chosen motifs from the enzyme sequences (equal in number and length to the SPs on them). Figure B-1 compares the SP coverage and the average random coverage of ProSite peptides for various minimal percentages.

For example, if we require that at least 40% of a ProSite peptides amino acids be covered by SPs to consider it entirely as covered, then SPs cover 48% of ProSite peptides, and random motifs cover on average only 24%, with a standard deviation of 0.06%. These statistically extremely significant results demonstrate that SPs not only cover more enzymes than ProSite motifs do, but they also cover existing ProSite motifs favourably.

Appendix C

THE SMITH-WATERMAN ALGORITHM

The Smith-Waterman Algorithm performs local alignment between two sequences; it finds a pair of segments, one from each of the two long sequences, such that there is no other pair of segments with greater similarity (homology). The similarity measure used here allows for arbitrary length deletions and insertions.

Let us define the two sequences as $A = a_1a_2...a_n$ and $B = b_1b_2...b_m$. A similarity $s(a, b)$ is given between sequence elements a and b, and a weight W_k is assigned to deletions of length k. So as to find pairs of segments with greatest similarity, a matrix H is constructed. All cell values start at zero and are not allowed to fall below zero (so a new alignment path can begin at any point). So we set: $H_{k0} = H_{0l} = 0$ for $0 \geq k \geq n$ and $0 \geq l \geq m$. Preliminary values of H have the interpretation that H_{ij} is the maximum similarity of two segments *ending* in a_i and b_j, respectively. When calculating the value for a cell in matrix H, there are four possibilities for ending the segments at any a_i and b_j:

1. If a_i and b_j are associated, the similarity is $H_{i-1,j-1} + s(a_i, b_j)$.

2. If a_i is at the end of a deletion of length k, the similarity is $H_{i-k,j} - W_k$.

3. If b_j is at the end of a deletion of length l, the similarity is $H_{i-k,j} - W_l$.

4. Finally, a zero is included to prevent calculated negative similarity, indicating no similarity up to a_i and b_j. (Zero need not be included unless there are negative values for s(a, b)).

The equation expressing this:

$$H_{ij} = max\{H_{i-1,j-1} + s(a_i, b_j), \max_{k\geq 1}\{H_{i-k,j} - W_k\}, \max_{l\geq 1}\{H_{i,j-l} - W_l\}, 0\} \quad \text{(C.1)}$$

for $1 \geq i \geq n$ and $1 \geq j \geq m$.

The pair of segments with maximum similarity is found by starting with the cell of greatest value in H. The other matrix elements leading to this maximum value are then sequentially determined with a traceback procedure ending with an element of H equal to zero. Thus both the segments and the corresponding alignment are produced. The pair of segments with the next best similarity is found by applying the traceback procedure to the second largest element of H not associated with the first traceback.

We bring here a simple example, on two DNA sequences: S_1 =CAGCCUCGCUUAG and S_2 = AAUGCCAUUGACGG. In this example, if a_i and b_j are a match, then the similarity is set to $s(a_i, b_j) = 1$. If a_i and b_j are a mismatch, then the similarity is set to $s(a_i, b_j) = -1/3$. The deletion weight was chosen to be $W_k = 1 + k/3$ (in general it must be at least the difference between a match and a mismatch). Figure C-1 shows the construction of the H matrix, calculating the value for $H_{6,7}$.

$$\begin{aligned}
\max_{k\geq 1}\{H_{6-k,7} - W_k\} &= \max_{k\geq 1}\{H_{6-k,7} - (1 + k/3)\} \\
&= \max\{H_{5,7} - (1 + 1/3), H_{4,7} - (1 + 2/3), H_{3,7} - (1 + 3/3), \\
&H_{2,7} - (1 + 4/3), H_{1,7} - (1 + 5/3)\} \\
&= \max\{1.0 - (1 + 1/3), 0.7 - (1 + 2/3), 0.0 - (1 + 3/3), \\
&0.0 - (1 + 4/3), 0.0 - (1 + 5/3)\} \\
&= \max\{-1/3, -29/30, -2.0, -7/3, -8/3\} \\
&= -1/3
\end{aligned} \quad \text{(C.2)}$$

	C	A	G	C	C	U	**C**	G	C	U	U	A	G
A	0.0	1.0	0.0	0.0	0.0	0.0	**0.0**	0.0	0.0	0.0	0.0	1.0	0.0
A	0.0	1.0	0.7	0.0	0.0	0.0	**0.0**	0.0	0.0	0.0	0.0	1.0	0.7
U	0.0	0.0	0.7	0.3	0.0	1.0	**0.0**	0.0	0.0	1.0	1.0	0.0	0.7
G	0.0	0.0	1.0	0.3	0.0	0.0	**0.7**	1.0	0.0	0.0	0.7	0.7	1.0
C	1.0	0.0	0.0	2.0	1.3	**0.3**	**1.0**	0.3	2.0	0.7	0.3	0.3	0.3
C	**1.0**	**0.7**	**0.0**	**1.0**	**3.0**	**1.7**	?						
A													
U													
U													
G													
A													
C													
G													
G													

Figure C-1: Partially filled H matrix. The the value for the element $H_{6,7}$ is calculated by considering the maximal element in its sub-row, the maximal element in its sub-column, and the element in the its diagonal, while taking into consideration mismatch penalties. Figures are taken from http://www.maths.tcd.ie/ lily/pres2/sld001.htm.

$$\begin{aligned}\max_{l\geq 1}\{H_{6,7-l} - W_l\} &= \max_{l\geq 1}\{H_{6,7-l} - (1 + l/3)\} \\ &= \max\{H_{6,6} - (1 + 1/3), H_{6,5} - (1 + 2/3), H_{6,4} - (1 + 3/3), \\ &H_{6,3} - (1 + 4/3), H_{6,2} - (1 + 5/3), H_{6,1} - (1 + 6/3)\} \\ &= \max\{1.7 - (1 + 1/3), 3.0 - (1 + 2/3), 1.0 - (1 + 3/3), \\ &0.0 - (1 + 4/3), 0.7 - (1 + 5/3), 1.0 - (1 + 6/3)\} \\ &= \max\{11/30, 4/3, -1, -7/3, -59/30, -2\} \\ &= 4/3\end{aligned} \tag{C.3}$$

$$H_{6-1,7-1} + s(a_6, b_7) = H_{5,6} + s(C, C) = 0.3 + 1(match) = 1.3 \tag{C.4}$$

So if we plug in (C.1) the calculated values (C.4), (C.2), (C.3), we will get the value for $H_{6,7}$:

$$\begin{aligned}H_{6,7} &= \max\{H_{6-1,7-1} + s(a_6, b_7), \max_{k\geq 1}\{H_{6-k,7} - W_k\}, \max_{l\geq 1}\{H_{6,7-l} - W_l\}, 0\} \\ &= \max\{1.3, -1/3, 4.3, 0\} = 4/3\end{aligned} \tag{C.5}$$

In this way all the cells of matrix H are calculated. The complete matrix is presented in Figure C-2. The matrix element with the greatest value is then located (in this example it is $H_{10,8} = 3.3$. From this maximal element, one traces back which other matrix element lead to it (must be wither in the maximal element's diagonal or in its sub-column or sub-row). In this example, $H_{9,7} = 2.3$ is the element that lead to the maximal element. This trace-back procedure is continued until a zero element is hit (in this case$H_{3,4}$ is the last element).

We can now retrieve the two segments, and the corresponding alignment, as

	C	A	G	C	C	U	C	G	C	U	U	A	G
A	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
A	0.0	1.0	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.7
U	0.0	0.0	0.7	0.3	0.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	0.7
G	0.0	0.0	1.0	0.3	0.0	0.0	0.7	1.0	0.0	0.0	0.7	0.7	1.0
C	1.0	0.0	0.0	2.0	1.3	0.3	1.0	0.3	2.0	0.7	0.3	0.3	0.3
C	1.0	0.7	0.0	1.0	3.0	1.7	1.3	1.0	1.3	1.7	0.3	0.0	0.0
A	0.0	2.0	0.7	0.3	1.7	2.7	1.3	1.0	0.7	1.0	1.3	1.3	0.0
U	0.0	0.7	1.7	0.3	1.3	2.7	2.3	1.0	0.7	1.7	2.0	1.0	1.0
U	0.0	0.3	0.3	1.3	1.0	2.3	2.3	2.0	0.7	1.7	2.7	1.7	1.0
G	0.0	0.0	1.3	0.0	1.0	1.0	2.0	3.3	2.0	1.7	1.3	2.3	2.7
A	0.0	1.0	0.0	1.0	0.3	0.7	0.7	2.0	3.0	1.7	1.3	2.3	2.0
C	1.0	0.0	0.7	1.0	2.0	0.7	1.7	1.7	3.0	2.7	1.3	1.0	2.0
G	0.0	0.7	1.0	0.3	0.7	1.7	0.3	2.7	1.7	2.7	2.3	1.0	2.0
G	0.0	0.0	1.7	0.7	0.3	0.3	1.3	1.3	2.3	1.3	2.3	2.0	2.0

Figure C-2: Complete H matrix: the underlined elements indicate the trace-back path from the maximal element 3.3.

shown in Figure C-3.

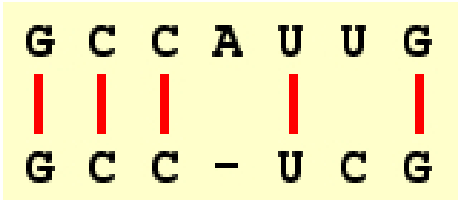

Figure C-3: The resulting segments and the corresponding alignment.

Appendix D

THE SWISS-PROT FORMAT

The Swiss-Prot Protein Knowledgebase is an annotated protein sequence database. It was established in 1986 and maintained collaboratively by the group of Amos Bairoch and the EMBL Outstation - The European Bioinformatics Institute (EBI). This appendix is based on the UniProt Knowledgebase User Manual available at http://www.expasy.org/sprot/userman.html.

The Swiss-Prot Protein Knowledgebase consists of sequence entries. Sequence entries are composed of different line types, each with their own format. For standardisation purposes the format of Swiss-Prot follows as closely as possible that of the EMBL Nucleotide Sequence Database. In Swiss-Prot, as in many sequence databases, two classes of data can be distinguished: the core data and the annotation. For each sequence entry the core data consists of: sequence data; citation information (bibliographical references); taxonomic data (description of the biological source of the protein).

The annotation consists of the description of the following items: Function(s) of the protein; post-translational modification(s) such as carbohydrates, phosphorylation, acetylation and GPI-anchor; domains and sites, for example, calcium-binding regions, ATP-binding sites, zinc fingers, homeoboxes, SH2 and SH3 domains and kringle; secondary structure, e.g. alpha helix, beta sheet; quaternary structure, i.e. homodimer, heterotrimer, etc.; similarities to other proteins; disease(s) associated with any number of deficiencies in the protein; sequence conflicts, variants, etc.

The entries in the database are structured so as to be usable by human readers as well as by computer programs. Each sequence entry is composed of lines. Different types of lines, each with their own format, are used to record the various data that make up the entry. A sample sequence entry is shown in Figure D-1. Each line begins with a two-character line code, which indicates the type of data contained in the line. Some of the important line types and line codes, as shown in Figure D-1, are explained here.

The **ID** (IDentification) line is always the first line of an entry. The first item on the ID line is the entry name of the sequence. The Swiss-Prot entry name consists of up to 11 uppercase alphanumeric characters. Swiss-Prot uses a general purpose naming convention that can be symbolised as X_Y, where: X is a mnemonic code of at most 5 alphanumeric characters representing the protein name. Examples: B2MG is for Beta-2-microglobulin, HBA is for Hemoglobin alpha chain and INS is for Insulin, CAD17 for Cadherin-17; The '_' sign serves as a separator; Y is a mnemonic species identification code of at most 5 alphanumeric characters representing the biological source of the protein. This code is generally made of the first three letters of the genus and the first two letters of the species.

The **AC** (ACcession number) line lists the accession number(s) associated with an entry. The purpose of accession numbers is to provide a stable way of identifying entries from release to release. It is sometimes necessary for reasons of consistency to change the names of the entries, for example, to ensure that related entries have similar names. However, an accession number is always conserved, and therefore allows unambiguous citation of entries.

The **DT** (DaTe) lines show the date of creation and last modification of the database entry.

The **DE** (DEscription) lines contain general descriptive information about the sequence stored. In the case of enzymes the EC number is given. This information is generally sufficient to identify the protein precisely.

The **GN** (Gene Name) line indicates the name(s) of the gene(s) that code for

```
ID   LEXA1_PSESM             Reviewed;         202 AA.
AC   Q87ZB9;
DT   24-OCT-2003, integrated into UniProtKB/Swiss-Prot.
DT   01-JUN-2003, sequence version 1.
DT   23-JAN-2007, entry version 31.
DE   LexA repressor 1 (EC 3.4.21.88).
GN   Name=lexA1; Synonyms=lexA-2; OrderedLocusNames=PSPTO_3510;
.
.
.
KW   Autocatalytic cleavage; Complete proteome; DNA damage; DNA repair;
KW   DNA replication; DNA-binding; Hydrolase; Repressor; SOS response;
KW   Transcription; Transcription regulation.
FT   CHAIN         1    202       LexA repressor 1.
FT                                /FTId=PRO_0000170072.
FT   DNA_BIND     28     48       H-T-H motif (By similarity).
FT   ACT_SITE    123    123       For autocatalytic cleavage activity (By
FT                                similarity).
FT   ACT_SITE    160    160       For autocatalytic cleavage activity (By
FT                                similarity).
FT   SITE         88     89       Cleavage; by autolysis (By similarity).
SQ   SEQUENCE   202 AA;  22150 MW;  48257AB73A3BB9B6 CRC64;
     MIKLTPROAE ILGFIKRCLE DNGFPPTRAE IAQELGFKSP NAAEEHLKAL ARKGAIEMTP
     GASRGIRIPG FEARPDESSL PVIGRVAAGA PILAQQHIEE SCNINPSFFH PSANYLLRVH
     GMSMKDVGIL DGDLLAVHTT REARNGQIVV ARIGDEVTVK RFKREGSKVW LLAENPDFAP
     IEVDLKDQEL VIEGLSVGVI RR
//
```

Figure D-1: A sample sequence entry of *LexA repressor 1* enzyme. Only the beginning and the end of the entry are shown here, bringing the important entries relevant to this work. Circled in blue are: the ID (IDentification) line with the entry name of the sequence (LEXA1_PSESM); the AC (ACcession) line that lists the accession number associated with the entry (Q872B9); the DE (DEscription) line containing the name of the enzyme (LexA repressor 1) and the EC number. The annotations of biological features are shown in the FT (FeaTure) line. The annotations for DNA_BIND, ACT_SITE and SITE are highlighted in matching colours on the sequence. The SPs that appear on this enzyme are underlined here, for future reference. SPs PPTRAEI and NAAEEHL overlap with the DNA_BIND feature, GMSME covers the ACT_SITE and the SPs DEVTVK and EVTVKR cover the other ACT_SITE. Descriptions of the lines and the feature identifiers are brought in this chapter.

the stored protein sequence. The GN line contains different types of information, such as gene names (a.k.a gene symbols) and ordered locus names (a name used to represent an open reading frame in a completely sequenced genome or chromosome).

The **KW** (KeyWord) lines provide information that can be used to generate indices of sequence entries based on functional, structural, or other categories.

The **FT** (Feature Table) lines provide a precise but simple means for the annotation of sequence data. The table describes regions or sites of interest in the sequence. In general the feature table lists post-translational modifications, binding sites, enzyme active sites, local secondary structure or other characteristics reported in the cited references. Sequence conflicts between references are also included in the feature table. The first item on each FT line is the key name, which is a fixed abbreviation (of up to 8 characters) with a defined meaning. Following the key name are the 'FROM' and 'TO' endpoint specifications. These fields designate (inclusively) the endpoints of the feature named in the key field. In general, these fields simply contain residue numbers which indicate positions in the sequence as listed. The remaining portion of the FT line is a description that contains additional information about the feature.

SwissProt feature identifiers used in this work: 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 (co-enzyme, prosthetic group, etc.); METAL - Binding site for a metal ion.

The **SQ** (SeQuence header) line marks the beginning of the sequence data and gives a quick summary of its content.

Bibliography

[1] S. Abhiman and E.L.L. Sonnhammer (2005) Large-scale prediction of function shift in protein families aith a focus on enzymatic function. *Proteins*, **60**, 758-768.

[2] A. Aitken (1999) Protein consensus sequence motifs, *Mol. Biotechnol.*, **12**, 241-253.

[3] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. and D.J. Lipman (1997) Gapped blast and psi-blst: a new generation of protein database search programs, *Nucl. Acids Res.*, **25**, 3389-3402.

[4] P.Anda, J.A. Gebbia, P.B. Backenson, J.L. Coleman and J.L. Benach (1996) A glyceraldehyde-3-phosphate dehydrogenase homolog in borrelia burgdorferi and borrelia hermsii, *Infect Immun.*, **64**, 262-268.

[5] A. Bairoch, P. Bucher and K. Hofmann (1997) Prosite, *Nucleic Acids Res.*, **25**, 217-22.

[6] A. Ben-Hur and D. Brutlag (2003) Remote homology detection: a motif based approach, *Bioinformatics*, **19**, Suppl 1, 26-33.

[7] A. Ben-Hur and D. Brutlag (2004) Protein sequence motifs: Highly predictive features of protein function. In: *Feature extraction, foundations and applications*, I. Guyon, S. Gunn, M. Nikravesh and L. Zadeh (eds.) Springer Verlag, to be published.

[8] Y. Benjamini, Y. Hochberg (1995) Controlling the flase discovery rate: A practical and powerful approach to multiple testing, *J. Roy. Stat. Soc.*, **57**, 289-300.

[9] T.A. Binkowski, S. Naghibzadeg, and J. Liang (2003) CASTp: computed atlas of surface topography of proteins, *Nucleic Acid Research*, **31**, 3352-3355.

[10] P. Bork and E.V. Koonin (1996) Protein sequence motifs, *Curr. Op. Structural Biology*, **6**, 366-376.

[11] Bordner, A. J., Abagyan, R. A. (2004) Large-Scale Prediction of Protein Geometry and Stability Changes for Arbitrary Single Point Mutations, *PROTEINS: Structure, Function and Bioinformatics*, **57**, 400-413.

[12] P.A. Boriack-Sjodin, R.W. Heck, P.J. Laipis, D.N. Silverman and D.W. Christianson (1995) Structure determination of murine mitochondrial carbonic anhydrase V at 2.45-A resolution: implications for catalytic proton transfer and inhibitor design, *Proc.Natl.Acad.Sci.USA*, **92**, 10949-10953.

[13] F.S. Domingues and T. Lengauer (2003) Protein function from sequence and structure data, *Appl Bioinformatics*, **2**, 3-12.

[14] L. Falquet, M. Pagni, P. Bucher, N. Hulo, C.J. Sigrist, K. Hofmann and A. Bairoch (2002) The prosite database, its status in 2002, *Nucleic Acids Res*, **30**, 235-238.

[15] P.G. Foster, L. Huang, D.V. Santi and R.M. Stroud (2000) The structural basis for tRNA recognition and pseudouridine formation by pseudouridine synthase i, *Nature Struc. Biol.*, **7**, 23-27.

[16] M. von Grotthuss, D. Plewczynski, K. Ginalsky, L. Rychlewski and E. I. Shakhnovich (2006) PDB-UF: database of predicted enzymatic functions for unannotated protein structures from structural genomics, *BMC Bioinformatics*, *7*, 53-62.

[17] S.K. Hanks, A.M. Quinn and T. Hunter (1988) The protein kinase family: Conserved features and deduced phylogeny of the catalytic domains, *Science*, **241**, 42-52.

[18] H. Hegyi, M. Gerstein (1999) The relationship between protein structure and function: a comprehensive survey with application to the yeast genome, *J. Mol. Biol.*, **288**, 147-164.

[19] J.Y. Huang and D.L. Brutlag (2001) The emotif database, *Nuclear Acids research*, **29**, 202–204.

[20] C. Kisker, H. Schindelin, B.E. Alber, J.G. Ferry and D.C. Rees (1996) A left-hand beta-helix revealed by the crystal structure of a carbonic anhydrase from the archaeon Methanosarcina thermophila, *EMBO J.*, **15**, 2323-2330.

[21] D.E. Koshland (1958) Application of a theory of enzyme specificity to protein synthesis, *Proc. Natl. Acad. Sci. U.S.A.*, **44**, 98-104.

[22] V. Kunik, Y. Meroz, B. Sandbank, E. Ruppin and D. Horn (2007) Functional representation of enzymes by Specific Peptides, *submitted for publication.*

[23] L. Liao and W.S. Noble (2003) Combining pairwise sequence analysis and support vector machines for detecting remote protein evolutionary and structural relationships, *J. of Comp. Biology*, **10**, 857-868, 2003.

[24] Y. Meroz and D. Horn (2007) Roles of specific peptides, *submitted to ISMB 2007.*

[25] J.L. Moreland, A. Gramada, O.V. Buzko, Qing Zhang, and P.E. Bourne (2005) The molecular biology toolkit (mbt): A modular platform for developing molecular visualization applications, *BMC Bioinformatics.*, **6**, 21.

[26] R. Mott (2000) Accurate formula for p-values of gapped local sequence and profile alignments, *J. Mol Biol.*, **300**, 649-659.

[27] A. G. Murzin, S. E. Brenner, T. Hubbard, C. Chothia (1995 SCOP: a structural classification of proteins database for the investigation of sequences and structures, *J. Mol. Biol.*, **247**, 536-540.

[28] N. Narayana, S. Cox, X. Nguyen-huu, L.F. Ten Eyck and S.S. Taylor (1997) A binary complex of the catalytic subunit of cAMP-dependent protein kinase and adenosine further defines conformational flexibility, *Structure*, **5**, 921-935.

[29] C.G. Neville-Manning, T.D. Wu and D.L. Brutlag (1998) Highly specific protein sequence motifs for genome analysis, *Proc. Natl. Acad. Sci. USA*, **95**, 5865–5871.

[30] F. M. G. Pearl, C. F. Bennett, J. E. Bray, A. P. Harrison, N. Martin, A. Shepherd, I. Sillitoe, J. Thornton and C. A. Orengo (2003) The CATH database: an extended protein family resource for structural and functional genomics, *Nucleic Acids Research*, **31**, No. 1, 452-455.

[31] G. H. Peters and R.P. Bywater (1999) Computational analysis of chain flexibility and fluctuations in Rhizomucor miehei lipase, *Protein Engeneering*, **12**, 747-754.

[32] B. Rost (2002) Enzyme function less conserved than anticipated, *J. Mol. Biol.*, **318**, 595-608.

[33] B. Rost, G. Yachdav and J. Liu (2004) The predictprotein server, *Nucleic Acids Res.*, **32**, 321-326.

[34] R. Sayle and E.J. Milner-White (1995) RasMol: Biomolecular graphics for all, *Trends in Biochemical Sciences*, **20**, No. 9, 374.

[35] L. Serrano, J. Sancho, M. Hirshbergand A.R. Fersht (1992) 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.*, **227**, 544-559.

[36] T. Smith, M. Waterman (1981) Identification of common molecular subsequences, *J. of Mol. Biology*, **147**, 195-197.

[37] Z. Solan, D. Horn, E. Ruppin, S. Edelman (2005) Unsupervised learning of natural languages, *Proc. Natl. Acad. Sci. USA*, **102**, 11629-11634.

[38] S. Subbiah (1996) Protein Motions, *Molecular Biology Intelligence Unit, Springer-Verlag, Heidelberg, Germany.*

[39] A. Teplyakov, P. Sebastiao, G. Obmolova, A. Perrakis, G.S. Brush, M.J. Bessman and K.S. Wilson (1996) Crystal structure of bacteriophage T4 deoxynucleotide kinase with its substrates dGMP and ATP, *EMBO J.*, **15**, 3487-34997.

[40] W. Tian and J. Skolnick (2003) How well is enzyme function conserved as a function of pairwise sequence identity?, *J. Mol. Biol.*, **333**, 863-882.

[41] P.S. Vermersch, J.G.G. Tesmer, D. Lemon and F.A. Quiocho (1990) 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.*, **265**, 16592-16603.

[42] D. Vitkup, C. Sander and G.M. Church (2003) The amino-acid mutational spectrum of human genetic disease, *Genome Biology*, **4**, R72.

[43] J.E. Walker, M. Saraste, M.J. Runswick and N.J. Gay (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.

[44] B.X. Yan and Y.Q. Sun (1997) glycine residues provide flexibility for enzyme active sites, *J. Bio. Chem.*, **272**, 3190-3194.