BIOINFORMATICS

Vol. 21 Suppl. 1 2005, pages i328–i337
doi:10.1093/bioinformatics/bti1023

In silico *identification of functional regions in proteins*

Guy Nimrod[1], Fabian Glaser[2], David Steinberg[3], Nir Ben-Tal[1,] and Tal Pupko[4]*

[1]Department of Biochemistry, George S. Wise Faculty of Life Sciences, Tel Aviv University, Ramat Aviv 69978, Israel, [2]European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridge CB10 1SD, UK, [3]Department of Statistics and Operations Research, Raymond and Beverly Sackler Faculty of Exact Sciences and [4]Department of Cell Research and Immunology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Ramat Aviv 69978, Israel

Received on January 15, 2005; accepted on March 27, 2005

ABSTRACT

Motivation: *In silico* prediction of functional regions on protein surfaces, i.e. sites of interaction with DNA, ligands, substrates and other proteins, is of utmost importance in various applications in the emerging fields of proteomics and structural genomics. When a sufficient number of homologs is found, powerful prediction schemes can be based on the observation that evolutionarily conserved regions are often functionally important, typically, only the principal functionally important region of the protein is detected, while secondary functional regions with weaker conservation signals are overlooked. Moreover, it is challenging to unambiguously identify the boundaries of the functional regions.

Methods: We present a new methodology, called PatchFinder, that automatically identifies patches of conserved residues that are located in close proximity to each other on the protein surface. PatchFinder is based on the following steps: (1) Assignment of conservation scores to each amino acid position on the protein surface. (2) Assignment of a score to each putative patch, based on its likelihood to be functionally important. The patch of maximum likelihood is considered to be the main functionally important region, and the search is continued for non-overlapping patches of secondary importance.

Results: We examined the accuracy of the method using the IGPS enzyme, the SH2 domain and a benchmark set of 112 proteins. These examples demonstrated that PatchFinder is capable of identifying both the main and secondary functional patches.

Availability: The PatchFinder program is available at: http://ashtoret.tau.ac.il/∼nimrodg/

Contact: NirB@tauex.tau.ac.il

*To whom correspondence should be addressed.

1 INTRODUCTION

Detection of the amino acid positions that are essential for activities, such as catalysis, protein–protein interactions or protein–ligand interactions, is a critical step in the study of the biological function of proteins (Lichtarge and Sowa, 2002). This task is especially important for proteins with a known structure and unknown function. Detection of key amino acid positions that are functionally important is also essential for drug design studies, for protein classification and annotation and for evolutionary studies (Rost, 2002). This need led to the development of a wide variety of computational methods for the identification of functional regions (Aloy *et al*., 2001; Amitai *et al*., 2004; del Sol Mesa *et al*., 2003; Friedberg and Margalit, 2002; Jones and Thornton, 1997; Innis *et al*., 2004; Ma *et al*., 2003; Madabushi *et al*., 2002; Neuvirth *et al*., 2004; Panchenko *et al*., 2004; Pazos and Sternberg, 2004).

Functionally important positions are often conserved, and many methods exploit the evolutionary history of the protein and its homologs. Dean and Golding (2000) developed a maximum likelihood (ML)-based methodology to discriminate between slowly and rapidly evolving regions of a protein. Unlike other methods, their algorithm does not assign a conservation score for each amino acid position, but rather assigns an estimated rate of evolution to a group of amino acids that are included in a sphere of variable diameter. The strength of their approach is the use of an explicit stochastic process to model protein evolution. However, by considering only spherical shapes, patches of shapes that are far from spherical are likely to be overlooked. Furthermore, the method does not distinguish between exposed and buried amino acid positions. Hence, the predicted functional sphere may include some amino acid positions that are buried in the protein, and are most probably structurally rather than functionally important. Dean and Golding further used a statistical test to check the rate significance for each sphere, comparing it with the rate distribution of a random population.

Aloy *et al*. (2001) used a method based on Evolutionary Trace (ET; Lichtarge *et al*., 1996) to assign conservation scores to each position. The ET method uses a phylogenetic tree of the query protein's family and detects residues that show evolutionary conservation in at least a subgroup (branch) of the tree, i.e. trace residues. Aloy *et al*. consider as candidates for functional patches only sites with amino acids that are polar and totally conserved (or with one 'residue sequence error'). This definition suffers from low sensitivity; an amino acid position is either 'conserved' or not. In practice, there are many levels of conservation, and the exclusion of positions that are not totally conserved or not polar may obscure functional regions. Moreover, the method is very sensitive to the choice of the homologous protein sequences that are used; the addition of more homologs is expected to reduce the number of putative positions, thereby counter-intuitively reducing the power of the method. Aloy *et al*.'s algorithm also assumes that the shape of the patch is spherical, which may not always be suitable.

The algorithm developed by Madabushi *et al*. (2002) for the automated, structure-based prediction of functional sites is also based on ET. A patch is defined as a collection of spatially close trace residues. Sets of patches are detected and a statistical test is used to assign their collective significance, based on the likelihood to obtain at random a cluster of trace residues of the size of the largest patch (Yao *et al*., 2003). Alternatively, the statistical significance is assigned based on the total number of patches of the trace residues that were found. The outcome is a set of key amino acids, some of which are buried and presumably important for the maintenance of the proper protein fold. Others are exposed and comprise the various functional sites.

We have developed the Rate4Site algorithm (Pupko *et al*., 2002), which is a rigorous statistical method for inferring the level of conservation at each amino acid position, taking into account the phylogenetic relations between the sequences and the stochastic process underlying their evolution. When Rate4Site's conservation scores are mapped onto the protein surface, one or more surface patches formed by spatially close and conserved residues often appear. These surface patches are potentially important functional regions. We subsequently launched the ConSurf webserver that implements this algorithm (Glaser *et al*., 2003).

The automatic nature of the ConSurf server makes it easy to use, but the interpretation of the results is subjective; there is no stringent criterion for determining the amino acid positions that compose a conserved region. Furthermore, while the largest and most conserved surface patch is usually easily detected by the naked eye, it might be difficult to identify secondary functional regions that are smaller in size and/or include residues that are less conserved, and distinguish them from the background noise. Thus, it is essential to discriminate between a random cluster of conserved amino acids and conserved patches that are statistically significant.

In this work, we present a novel methodology for the automatic identification of functionally important regions in proteins with a known 3D structure. This methodology, which is referred to as PatchFinder, is based on three features. First, PatchFinder permits patches of any shape and size on the protein surface rather than limiting the patch to any pre-set shape (e.g. sphere) around a particular amino acid. Second, amino acid positions that are on the protein surface are distinguished from those that are buried in the protein core (Jones and Thornton, 1997). This allows us to approximately distinguish between amino acids that are conserved due to structural constraints (Schueler-Furman and Baker, 2003), and those that are conserved because they are functionally important. Third, PatchFinder assigns a probability to each inferred patch to be functionally important, and searches for the patch with the highest probability. Typically, this patch corresponds to the main conservation signal, and is relatively easy to detect. The following non-overlapping patches are associated with lower likelihood values, and are much more difficult to detect even though they are often biologically important.

2 ALGORITHM

PatchFinder's input includes the 3D structure of a query protein and a multiple-sequence alignment (MSA) of the protein and its sequence homologs.

2.1 Assignment of conservation scores

The first step in the PatchFinder algorithm is the assignment of an evolutionary conservation score to each amino acid position using the Rate4Site algorithm (Pupko *et al*., 2002), as implemented in the ConSurf webserver (Glaser *et al*., 2003). In brief, Rate4Site produces a ML estimate of the score at each position, based on a MSA, a phylogenetic tree and a model of sequence evolution, e.g. JTT (Jones *et al*., 1992).

2.2 Identification of exposed and buried residues

In order to approximately distinguish between positions that are conserved due to structural constraints and those that are conserved due to their functional importance, the patch search procedure is limited to positions that are exposed to the solvent. Discrimination between 'exposed' and 'buried' positions is based on the solvent accessible surface area (ASA), computed using the Surface Racer program (Tsodikov *et al*., 2002). The computation is done on the 3D structure of the query protein, using a probe sphere of radius 1.4 Å, corresponding to a water molecule.

PatchFinder further defines a residue as 'exposed' if its relative accessible surface area (RSA) is greater than a fixed value; the default is 1%. The RSA for each residue is defined as the fraction of its ASA relative to its maximal ASA. The maximal ASA of a residue is calculated in an extended GXG tripeptide, where G is glycine and X is the residue in question (Miller *et al*., 1987).

2.3 Identification of patches of conserved residues

Our algorithm aims to find the most significant primary and secondary patches containing conserved and exposed amino acids. The challenge is to distinguish between a random clustering of a few conserved amino acids and a conserved patch that is functionally important. Based on previous studies (Aloy *et al*., 2001; Dean and Golding, 2000; Landgraf *et al*., 2001; Madabushi *et al*., 2002; Valdar and Thornton, 2001a,b), our intuition is that the functionality of a patch often correlates with two main factors: the number of amino acids that comprise it, and their average conservation. The assumption here is that a significant cluster of conserved residues on the protein surface is usually indicative of an evolutionary pressure presumably to maintain function. In summary, the larger and more conserved a patch is, the more likely it is to be a functional region.

2.3.1 Formal definition of terms A patch is defined as a cluster of spatially close amino acids. Clustering is based on a default cutoff distance of 4 Å between any two heavy atoms (Valdar and Thornton, 2001a). The patch size is defined in terms of the number of amino acids that comprise it. We define the conservation of the patch as the average conservation score of its residues.

2.3.2 The overall search procedure Our objective is to search the space of possible patches and to find the patch with the lowest probability to occur by chance. We use the following search procedure: we set a minimal average conservation (MAC) cutoff for a patch, search for the biggest patches with an average conservation equal to or higher than the cutoff and then compute its probability. Thus, for each possible MAC cutoff, we obtain a candidate patch and its probability. We finally choose the value of MAC cutoff that results with the patch with the highest probability, or in other words, the ML patch.

In practice, the time complexity of a complete search procedure is exponential (in sequence length) and therefore not applicable for proteins. We used a heuristic greedy search procedure because it is much faster and results in a satisfactory performance. The search procedure has two stages: initiation and extension. In the initiation stage, each of the 10 residues with the highest conservation score on the protein surface is selected as a starting point. In the extension stage, the most highly conserved of the neighboring residues is added to the existing patch, and the average conservation of the new patch is calculated. If the average conservation is higher than the cutoff, the new residue is accepted and the extension stage is repeated. The search procedure ends when the average conservation of the patch drops below the cutoff. Subsequently, a likelihood score is assigned to each patch (see below). Thus, for each set of possible MAC cutoffs, a list of patches and their corresponding likelihoods are obtained. The patch with the highest likelihood is defined as the primary (main) patch.

2.3.3 Assigning likelihood to each conserved patch Each of the patches that were found in the previous stage is assigned a probability to be functional. This is achieved by shuffling the conservation scores (Madabushi *et al*., 2002; Valdar and Thornton, 2001a; Yao *et al*., 2003) between the exposed residues, and performing the clustering identification stage again for each MAC cutoff. This procedure is repeatedly carried out N times (the default is 50 000). In this way, a distribution of patch sizes is obtained for each MAC cutoff. The probability of obtaining a cluster of size X and conservation score Y equals the probability of obtaining a patch of size X or more, with average conservation $\geq Y$. This value is approximated by measuring the frequency of patch of size X or bigger for the same MAC cutoff used for the estimated patch. The probability assigned to the patch would be the complementary event.

2.3.4 Use of conservation percentiles to set the MAC cutoffs In the search procedure described above, a patch is found for each MAC cutoff and the patch with the highest likelihood is selected. In PatchFinder, the MAC cutoffs are chosen based on the distribution of the conservation scores of all the exposed amino acids. The interval between query MAC cutoff values is 1%. The default search was limited to the 60–100 percentile interval with band size 1 (100, 99, … , 60), because our results showed that the likelihood generally drops drastically for percentiles <60.

In some cases, two patches are assigned to very close probability scores. It is thus also recommended to inspect overlapping patches that were ranked highly by the PatchFinder. These alternative patches can serve as a 'confidence interval' around the patch found. Patches with identical computed likelihood (generally 100%) are ranked by their Z-scores (i.e. the size of the patch minus the average size divided by the standard deviation).

2.4 Identification of secondary patches

The methodology described above yields a primary functional patch, with a likelihood value that is often $\sim$100%. This patch is most likely to be the main functional region of the protein. However, in many cases, we find that there are additional, non-overlapping patches with high likelihood to be functional. These patches may represent secondary functional regions. Once the primary patch with the highest likelihood on the list is found, we begin to search for significant patches, limiting the search to only those residues that are not found in the primary patch. Also at this stage, the likelihood of the patch is determined using a random shuffling procedure, in which for each randomization we discard the most conserved patch with size that is equal to that of the principal patch of the previous stage(s). This procedure is performed by the hill climbing heuristic.

The calculated likelihood values are used only to rank the putative patches according to their statistical significance. In most cases, the actual probability of the most significant patch to be functional is reflected by its average conservation.

2.5 Final output

The final output includes up to three non-overlapping patches. The likelihood and size of each patch, and the identities of the residues that compose it are reported.

3 IMPLEMENTATION

The algorithm was implemented using C++ and PERL and is available at: http://ashtoret.tau.ac.il/∼nimrodg/

3.1 Evaluation of sensitivity and specificity

We evaluated the performance of PatchFinder in terms of sensitivity (the fraction of functional residues identified) and specificity (the fraction of functionally important residues identified out of the total number of residues in the patch). For example, consider a case where the true patch was composed of 9 amino acids and PatchFinder inferred a patch of size 8, of which 7 amino acids were part of the true patch. In such a case, the sensitivity is 7/9 and the specificity 7/8.

3.2 In-depth analysis: IGPS and the SH2 domain

The results were evaluated using previously known data and SURFV (Sridharan *et al*., 1992) analysis. The gold standard active-site residues were considered to be amino acids known in the literature to be catalytically important, as well as those with exposure levels that decreased by >10% when comparing the protein chain alone to the protein–ligand complex. Similar results were obtained where the gold standard was based on the LIGPLOT (Wallace *et al*., 1995) data.

Src was divided into domains as follows: SH2: Trp148 to Pro246; SH3: Met82 to Glu147; Kinase: Thr247 to Glu524. pTyr527 and the three residues following it were considered as the phosphopeptide (Xu *et al*., 1997).

Conservation analysis for IGPS was carried out using an MSA of 153 homologs from the HSSP database (Sander and Schneider, 1991). For the SH2 domain, an MSA of residues Trp148 to Pro246 was used (Pupko *et al*., 2002). It contained 34 Src-like SH2 domains with sequence identity of >60%. The MSAs are available as Supplementary material.

3.3 Benchmarking

3.3.1 Dataset Benchmarking was based on a non-redundant set of 112 single-chain proteins with functional residues that are documented in the PDB (del Sol Mesa *et al*., 2003). The documentation is of varied quality. Nevertheless, since the researchers who determined the structures manually annotated it, we assume that the annotations are reliable, and represent at least portion of the functional positions.

3.3.2 Evaluation of the benchmarking results The active site residues documentation is, in many cases, partial, so the SITE residues cannot be treated as a perfect gold standard. We therefore used a statistical test similar to the one presented by del Sol Mesa *et al*. (2003) to assess the performance of PatchFinder results. The test examines the null hypothesis that PatchFinder locates the patch at random, regardless of functionality, and is based on the distance between the residue that is closest to the patch center and the residue that is closest to the SITE's center. Under the null hypothesis, the patch center is a random choice from among the exposed residues, and the P-value of the test for a single protein is computed from the resulting distribution of distances from the SITE's center. The full collection of P-values is then compared to a uniform distribution.

4 RESULTS

4.1 The IGPS Enzyme

4.1.1 General description Indole-3-Glycerol Phosphate Synthase (IGPS) catalyses one of the reactions in the pathway of biosynthesis of tryptophan. The structure of the IGPS from *Sulfolobus solfataricus* was determined with both CdRP (PDB ID: 1lbl) and IGP (PDB ID: 1a53) (Hennig *et al*., 2002). This allows a comprehensive examination of the active site.

4.1.2 Patch analysis PatchFinder analysis of this structure (PDB ID: 1lbl) yielded a few overlapping patches, each with likelihood of 100%. Of these, the one that is composed of 19 residues and was assigned the best Z-score is presented in Table 1. The experimental data indicate that six residues of IGPS are directly involved in ligand binding (Darimont *et al*., 1998) and the patch includes them all. A total of 23 residues are in contact with either IGP or CdRP, 18 of which are in the patch (Table 1). The other five residues (Leu83, Asp111, Ile213, Ile232 and Ser234) were detected in patches that are assigned lower MAC cutoffs and with high likelihood values. One residue, Gly 91, was in the patch but does not contact any of the ligands examined. Nevertheless, it shows a high level of conservation.

In this example, the PatchFinder analysis yields a single functional patch that corresponds approximately to the ligand-binding site (Fig. 1). The sensitivity in this case is 18 out of 23 amino acids (∼78%) and the specificity is 18 out of 19 residues in the patch (∼95%).

4.2 The SH2 domain

4.2.1 General description Src is a tyrosine kinase that functions in receptor-mediated signal transduction. It is an intracellular protein, composed of three main domains: kinase, SH2 and SH3. Regulation of the kinase activity is achieved through the interactions among these domains (Brown and Cooper, 1996). In the basal state, Src's Tyr527, which is located on the C-tail, is phosphorylated; the tail binds to a phosphopeptide-binding pocket in the SH2 domain and the enzyme is inactive. The exchange of the C tail with an external phosphopeptide may lead to Src activation.

Table 1. PatchFinder results for IGPS

	Residue[a]	IGP[b]	CdRP[c]
False negatives	Leu 83	+	
	Asp 111	+	
	Ile 213	+	+
	Ile 232	+	+
	Ser 234	+	+
True	Glu 51*	+	
	Lys 53*	+	+
	Ser 56		+
	Pro 57		+
	Phe 89	+	+
	Lys 110*	+	+
	Phe 112	+	+
	Leu 131	+	
	Ile 133	+	
	Glu 159*	+	+
	Asn 180*	+	+
	Arg 182*	+	+
	Leu 184	+	+
	Glu 210	+	+
	Ser 211	+	+
	Gly 212	+	+
	Leu 231	+	+
	Gly 233	+	+
False positives	Gly 91		

The residues were partitioned into three groups: True, false negatives and false positives.
[a]The amino acid type and number. Amino acids marked with '*' were confirmed experimentally as catalytically important residues (Darimont *et al*., 1998).
[b]The '+' sign marks amino acids that were found to be in contact with the IGP substrate (PDB ID 1a53).
[c]The '+' sign marks amino acids that were found to be in contact with the CdRP substrate (PDB ID 1lbl).

Here we present an analysis of Src's SH2 domain. This regulatory domain contains ∼100 amino acids and is very common in signaling proteins (Xu *et al*., 1997). For the analysis, we used the crystal structure of Xu *et al*. (1997) (PDB ID: 1fmk), where Src is in its inactive state.

4.2.2 Patch analysis PatchFinder analysis of the domain yielded patches as described in Table 2 and Figure 2. The first patch (patch 1; Table 2) includes 14 highly conserved residues. This is the peptide-binding groove. Using SURFV, 11 residues were inferred to be a part of the binding site. Eight of them are in the patch identified by PatchFinder; one (Thr215) is identified in the next best patch and two (Thr179 and Cys185) are located on the boundary of the binding groove and assigned conservation scores that are significantly lower than the average conservation of the patch.

We next focused on SH2's interactions with the SH3 and kinase domains within the intact Src protein. SH2's interfaces with these domains were identified based on SURFV analysis of the difference in the water-accessible surface area of each SH2 residue with and without the other domains. The results yielded the interface between the domains. However, the interfaces can only be viewed as an approximated gold standard, since the functionally important residues are only part of the interface (DeLano, 2002). Thus, the SURFV analysis can only be used to confirm true positive hits and point to a possible function.

Of the secondary patches, the one that was assigned the ML has three invariable residues (Gly210, Gly211 and Phe220). These three residues constitute the core of a larger patch of 16 residues, which was ranked second in its likelihood (patch 2; Table 2). The latter patch corresponds to the interface between the SH2 and SH3 domains and to a part of the interface between the SH2 and kinase domains.

The third, non-overlapping patch (patch 3; Table 2) is composed of three amino acids: Glu159, Arg160 and Leu163, that are mainly located at the interface between the kinase and SH2 domains.

Overall, 33 amino acids were detected in the 3 patches. As demonstrated in Figure 2, they correspond well to the peptide-binding groove and the interfaces of the SH2 domain with the rest of intact Src. The sensitivity in the peptide-binding groove in this case was 8/11 (∼72%), and the specificity 8/14 (∼57%).

The SH2 domain examined here was also crystallized with a high-affinity phosphotyrosyl peptide (Waksman *et al*., 1993). In this complex, three extra residues were added to those identified as the C-tail contacts, two of which (Gly236 and Leu237) were in the PatchFinder result. The differences between the two experimental results demonstrate some of the problems associated with the definition of a gold standard in this and similar cases.

4.3 Benchmarking

4.3.1 General description Analysis of 112 protein structures, that contain 931 annotated SITE residues, was conducted in order to assess the overall performance of PatchFinder. The data are available on the accompanied website. A total of 3221 residues, including 431 of the annotated SITEs, were found in the 112 main patches that were detected by PatchFinder. In 85% of the cases the first patch is composed of at least one of the SITE residues, and in 63% of the cases, at least half of these residues are in the patch.

A simple approach to the detection of functional sites involves mapping of all the surface residues that are evolutionarily conserved regardless of whether they are located in spatial proximity to each other. Our survey included all the exposed residues for which the conservation scores were at least as high as the minimal score in the ML patch. In comparison with the PatchFinder results, the number of SITE residues that are detected correctly this way is 476, i.e. a ∼10% increase in the detection rate. However, the total number of predicted functional residues is 4548, i.e. a ∼41% increase. Thus, the

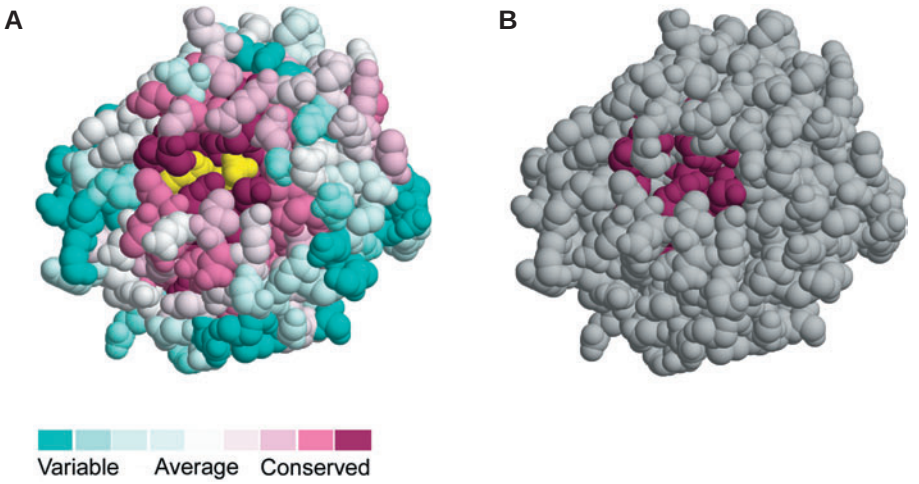

Fig. 1. The active site found by PatchFinder in the IGPS enzyme (PDB ID: 1lbl). (**A**). ConSurf conservation profile of the enzyme mapped onto the space-filling representation of the protein. The conservation color-coding plate is presented below and the ligand is colored yellow. (**B**). The same view as in A, with the residues identified by PatchFinder color-coded by conservation, while the rest of the molecule is gray. Only residues with a RSA >0.1% were considered in the search procedure. The picture was produced using RASMOL (Sayle and Milner-White, 1995).

results suggest that PatchFinder's search for patches of spatially close residues helps to reduce the rate of false positive predictions.

4.3.2 P-value assessment Because of the incomplete nature of PDB annotation of the active sites, we consider the *P*-value measure, based on the 'centers' of the patch and the active site, to be particularly suitable for this statistical survey. In 48 of the proteins (∼43%), the main patch that was detected has a *P*-value <0.05. Sixty nine results (62%) have a *P*-value <0.1. Similar *P*-values were obtained for both small and large proteins.

4.3.3 A closer look at failures We examined about 20 of the cases that were assigned the highest *P*-values, in search for the main reasons for failures. Our analysis revealed four main categories: (1) The MSA included only several proteins that were too closely related to each other. In such cases, it is difficult to distinguish between amino acid positions that are conserved due to the evolutionary pressure and those that appear to be conserved due to insufficient evolutionary time. (2) Most of the SITE residues are less conserved than the patches that were detected (or not conserved at all). For example, the serine protease of PDB ID 1 thm (Teplyakov *et al*., 1990) includes 17 SITE residues and PatchFinder detected an evolutionarily conserved patch of 14 residues. However, the vast majority of the SITE residues were assigned a conservation score that was lower than the average conservation of the patch residues. Thus, only three of the SITE residues were included in the patch. Interestingly, eight of the remaining patch residues are in contact with the protein's inhibitor (Gros *et al*., 1989), which suggests that PatchFinder's patch is functionally important nevertheless. (3) In PDB ID: 1mup (Bocskei *et al*., 1992) and other cases, most of the SITE residues are buried. (4) PatchFinder's patch is much bigger than the documented active site; this may indicate a partial documentation of the SITE residues. Pyruvate dehydrogenase (PDB ID 1iyu; Berg *et al*., 1996), which has only one documented SITE residue, is a very likely example of such a case.

In the first two categories, evolutionary conservation is obviously unsuitable for functional site inference, and complementary approaches should be used.

5 DISCUSSION

In this work, we presented a novel procedure for the identification of functional regions of proteins with known 3D structures. The procedure identifies the boundaries, sizes and statistical significance of evolutionarily conserved patches of residues on the protein surface. In the following, we will relate the current methodology to similar approaches.

Unlike the previously described methods, PatchFinder separates exposed and buried residues as an approximate means to distinguish between functionally and structurally important residues. The importance of this feature was demonstrated in test calculations that were carried out without taking into account the buried/exposed nature of the amino acids. The rates of false-negative and -positive predictions in these calculations (data not shown) were significantly higher than the rates reported above. Moreover, analysis of the benchmark dataset showed that ∼86% of the residues that are documented as SITEs are exposed. In comparison, only ∼64% of the conserved residues are exposed. Our survey (Supplementary

Table 2. PatchFinder results for the SH2 domain

Interface	Residue[a]	PatchFinder[b]
Phosphopeptide	Arg 155	1
	Arg 175	1
	Ser 177	1
	Thr 179	
	Cys 185	
	Asp 190*	1
	Lys 200	1
	His 201	1
	Tyr 202	1
	Lys 203	1
	Arg 205*	2
	Tyr 213*	1
	Ile 214	1
	Thr 215	2
	Leu 237*	1
	Cys 238*	1
SH3	Trp 148	2
	Tyr 149	2
	Tyr 184	
	Leu 223	
	Gln 224	
	Val 227*	2
	Val 244	
Kinase	Phe 150	2
	Arg 155	1
	Arg 156	
	Glu 157	
	Glu 159	3
	Arg 160	3
	Leu 161	
	Leu 163*	3
	Asn 164	
	Glu 178	
	Thr 179	
	Cys 245	2
	Pro 246	
Others	Leu 207	2
	Asp 208	2
	Gly 210	2
	Gly 211	2
	Arg 217	2
	Gln 219	2
	Phe 220	2
	Tyr 229	2
	Asp 235	2
	Gly 236	1
	Leu 241	1

Residues are segregated into four classes: 'Phosphopeptide' those that comprise the interface with the phosphopeptide; 'SH3' the SH3 domain; 'Kinase' the kinase domains, and 'Others' residues that are not in any of these interfaces. It is noteworthy that a residue can belong to more than one category.

[a]The amino acid type and number are partitioned according to their contacts with the relevant domain. Amino acids marked with '*' were not detected as contacts but are in close proximity (<4 Å) to the ones that compose the patch.

[b]The three patches found by PatchFinder with numbers corresponding to the iteration where the patch was located.

material) showed a factor of 2.9-enrichment in SITE residues in the conserved-and-exposed amino acids compared with the conserved-and-buried.

Similar to Madabushi *et al*. (2002), PatchFinder does not impose any initial constraints on the shape of the functional patches. Like Dean and Golding (2000), the patch clustering procedure is flexible and allows inclusion of moderately conserved residues by the use of the average conservation of the patch. PatchFinder is also capable of identifying secondary functional regions with a weaker conservation signal than the main one.

PatchFinder depends on the quality of the input MSA provided. When the alignment is unreliable or not diverse enough the inferred conservation scores are incorrect, and the search for functional regions may become meaningless.

The accuracy of the calculations also depends on the input 3D structure of the protein. Some proteins are known to fluctuate between two or more biologically relevant conformations; for example, active and inactive states of an enzyme. This variability might change the level of exposure of amino acids and their mutual distances; both quantities are exploited by PatchFinder. Using only a single structure might increase the rate of false-negative predictions.

In the two examples that were presented in detail (Figs 1 and 2, Tables 1 and 2), we show the potency and advantages of this new methodology. As other algorithms (Madabushi *et al*., 2002; Aloy *et al*., 2001), PatchFinder successfully identified many of the reported/inferred functional residues of the primary site on the protein surface, while the predicted functional patches included a few potentially non-functional residues. Studies performed on the benchmark set yielded similar results (see Supplementary material). Thus, PatchFinder can be used for automatic, high-throughput detection of the approximate location of the main functional regions of proteins.

The detection of secondary patches is more challenging because of their weaker conservation signal and/or smaller size. PatchFinder detected such patches in the SH2 domain and assigned them with reasonably high average conservation scores (Fig. 2, Table 2). However, they were assigned with low likelihood values, thus complicating their automatic identification.

There are several features and improvements that could be introduced into PatchFinder: (1) PatchFinder currently searches for patches of highly conserved positions, while in some cases the functionally important region may even be hyper-variable, as the peptide-binding groove of the MHC class I heavy chain (see the GALLERY of the ConSurf server at http://consurf.tau.ac.il). (2) The inclusion of information about the sequential location of the conserved amino acids may improve PatchFinder's sensitivity (Mihalek *et al*., 2003). (3) Currently, PatchFinder does not take into account specific attributes of different kinds of functional sites; for example, the residues that compose catalytic sites are generally partially

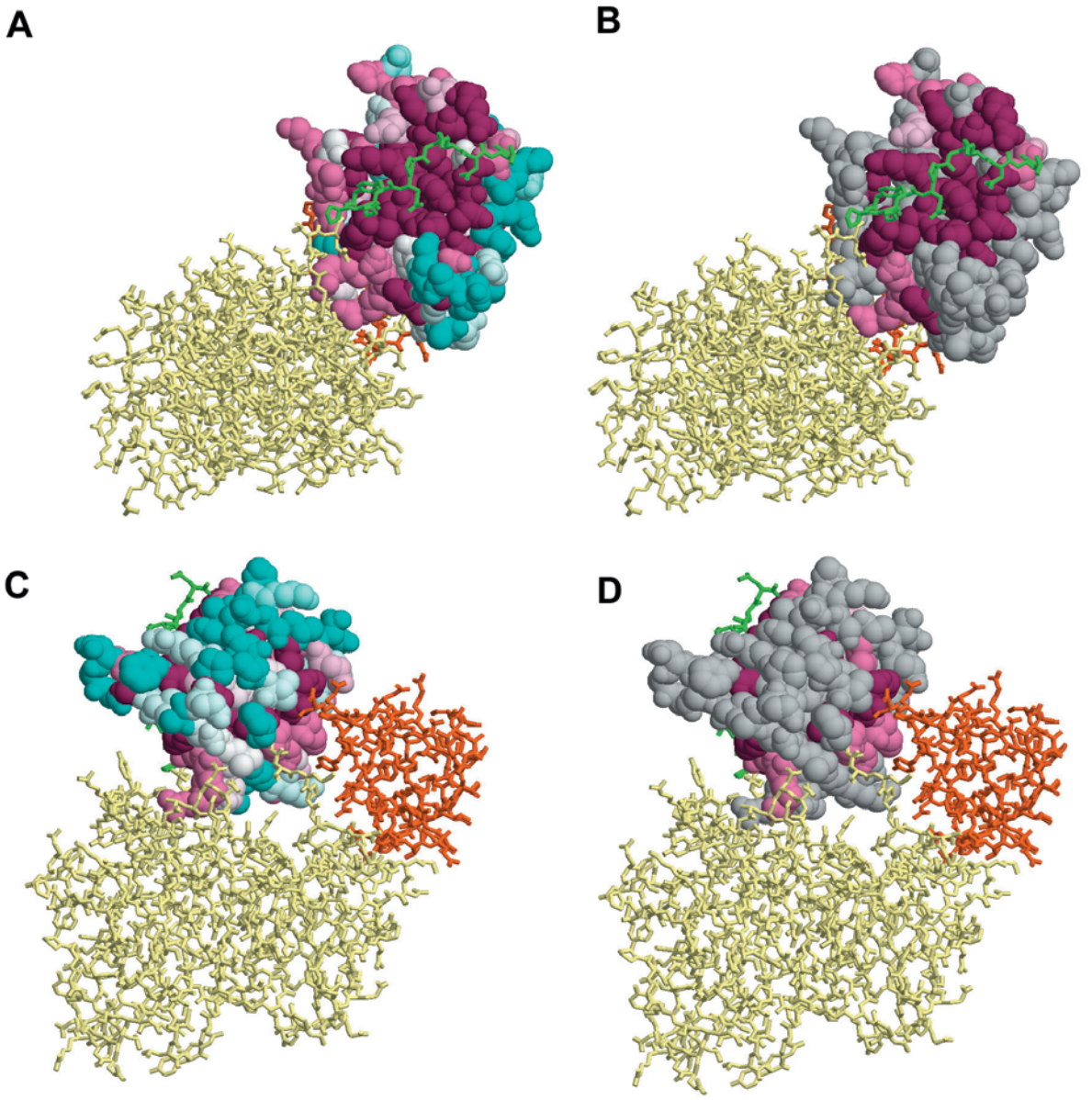

Fig. 2. The functional regions found by PatchFinder in the SH2 domain of human Src (PDB ID: 1fmk, Xu *et al*., 1997). The kinase domain is colored yellow, the SH3 domain is colored orange and the C-tail is in bright green. The SH2 domain is shown in a space filled representation, colored as follows: (**A**) and (**C**) colored by the conservation scale of Figure 1, (**B**) and (**D**) patch residues are colored by conservation, and the rest of the amino acids are gray. Only residues with a RSA >1% were considered in the search procedure. The picture was produced using RASMOL (Sayle and Milner-White, 1995).

buried and highly conserved, while protein–protein interaction regions are typically found in larger cavities with lower average conservation. This suggests the adjustment of parameters when searching for a patch of a specific type. In fact, we took a first step in this direction by using different criteria for discrimination between buried and exposed residues in active sites (a cutoff of 0.1%) and in inter-protein interfaces (a cutoff of 1%). Similarly, the accuracy of detection of enzyme active sites can be enhanced by the addition of a requirement that at least one of the residues be polar (Aloy *et al*., 2001). (4) We could also consider other properties of the patch, such as the curvature, the identity and nature of the amino acids comprising it and their class specificity (Chakrabarti and Jannin, 2002; Shanahan *et al*., 2004). This is likely to improve PatchFinder's accuracy in the detection of functional regions. Moreover, the improved procedure is likely to provide information on the function of the region (protein–protein interface, ligand- or DNA-binding, etc.),

which is completely missing now. Finally, the combination of some of these measures in the clustering procedure may also improve the accuracy and sensitivity of the calculations (Mihalek *et al*., 2004; Oliveira *et al*., 2003).

In conclusion, we believe that PatchFinder's capacity for high-throughput screening of functional regions in proteins of known 3D structure will be useful in the context of the proteomics and structural genomics initiatives, and may be used to further characterize known functional regions, as well as to reveal conserved regions that are as yet unknown.

ACKNOWLEDGEMENTS

We thank Itay Mayrose and Sarel Fleishman for helpful discussions and comments on the manuscript. This work was supported by the Israel Cancer Association grant to N.B.-T. The research was initiated when T.P. was doing his Postdoctorate with Dr Dave Swofford. He thanks Dave Swofford for his support and for helpful discussion. T.P. was supported by a grant in Complexity Science from the Yeshaia Horvitz Association.

REFERENCES

Aloy,P., Querol,E., Aviles,F.X. and Sternberg,M.J. (2001) Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. *J. Mol. Biol.*, **311**, 395–408.

Amitai,G., Shemesh,A., Sitbon,E., Shklar,M., Netanely,D., Venger,I. and Pietrokovski,S. (2004) Network analysis of protein structures identifies functional residues. *J. Mol. Biol.*, **344**, 1135–1146.

Berg,A., Vervoort,J. and de Kok,A. (1996) Solution structure of the lipoyl domain of the 2-oxoglutarate dehydrogenase complex from *Azotobacter vinelandii*. *J. Mol. Biol.*, **261**, 432–442.

Bocskei,Z., Groom,C.R., Flower,D.R., Wright,C.E., Phillips,S.E., Cavaggioni,A., Findlay,J.B. and North,A.C. (1992) Pheromone binding to two rodent urinary proteins revealed by X-ray crystallography. *Nature*, **360**, 186–188.

Brown,M.T. and Cooper,J.A. (1996) Regulation, substrates and functions of src. *Biochim. Biophys. Acta*, **1287**, 121–149.

Chakrabarti, P. and Janin, J. (2002) Dissecting protein–protein recognition sites. *Proteins*, **47**, 334–343.

Darimont,B., Stehlin,C., Szadkowski,H. and Kirschner,K. (1998) Mutational analysis of the active site of indoleglycerol phosphate synthase from *Escherichia coli*. *Protein Sci.*, **7**, 1221–1232.

Dean,A.M. and Golding,G.B. (2000) Enzyme evolution explained (sort of). *Pac. Symp. Biocomput.*, 6–17.

del Sol Mesa,A., Pazos,F. and Valencia,A. (2003) Automatic methods for predicting functionally important residues. *J. Mol. Biol.*, **326**, 1289–1302.

DeLano,W.L. (2002) Unraveling hot spots in binding interfaces: progress and challenges. *Curr. Opin. Struct. Biol.*, **12**, 14–20.

Friedberg,I. and Margalit,H. (2002) Persistently conserved positions in structurally similar, sequence dissimilar proteins: roles in preserving protein fold and function. *Protein Sci.*, **11**, 350–360.

Glaser,F., Pupko,T., Paz,I., Bell,R.E., Bechor-Shental,D., Martz,E. and Ben-Tal,N. (2003) ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics*, **19**, 163–164.

Gros,P., Fujinaga,M., Dijkstra,B.W., Kalk,K.H. and Hol,W.G. (1989) Crystallographic refinement by incorporation of molecular dynamics: thermostable serine protease thermitase complexed with eglin c. *Acta. Crystallogr. B*, **45**, 488–499.

Hennig,M., Darimont,B.D., Jansonius,J.N. and Kirschner,K. (2002) The catalytic mechanism of indole-3-glycerol phosphate synthase: crystal structures of complexes of the enzyme from *Sulfolobus solfataricus* with substrate analogue, substrate, and product. *J. Mol. Biol.*, **319**, 757–766.

Innis,C.A., Anand,A.P. and Sowdhamini,R. (2004) Prediction of functional sites in proteins using conserved functional group analysis. *J. Mol. Biol.*, **337**, 1053–1068.

Jones,D.T., Taylor,W.R. and Thornton,J.M. (1992) The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.*, **8**, 275–282.

Jones,S. and Thornton,J.M. (1997) Prediction of protein–protein interaction sites using patch analysis. *J. Mol. Biol.*, **272**, 133–143.

Landgraf,R., Xenarios,I. and Eisenberg,D. (2001) Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. *J. Mol. Biol.*, **307**, 1487–1502.

Lichtarge,O., Bourne,H.R. and Cohen,F.E. (1996) An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.*, **257**, 342–358.

Lichtarge,O. and Sowa,M.E. (2002) Evolutionary predictions of binding surfaces and interactions. *Curr. Opin. Struct. Biol.*, **12**, 21–27.

Ma,B., Elkayam,T., Wolfson,H. and Nussinov,R. (2003) Protein–protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. *Proc. Natl Acad. Sci. USA*, **100**, 5772–5777.

Madabushi,S., Yao,H., Marsh,M., Kristensen,D.M., Philippi,A., Sowa,M.E. and Lichtarge,O. (2002) Structural clusters of evolutionary trace residues are statistically significant and common in proteins. *J. Mol. Biol.*, **316**, 139–154.

Mihalek,I., Res,I. and Lichtarge,O. (2004) A family of evolution-entropy hybrid methods for ranking protein residues by importance. *J. Mol. Biol.*, **336**, 1265–1282.

Mihalek,I., Res,I., Yao,H. and Lichtarge,O. (2003) Combining inference from evolution and geometric probability in protein structure evaluation. *J. Mol. Biol.*, **331**, 263–279.

Miller,S., Janin,J., Lesk,A.M. and Chothia,C. (1987) Interior and surface of monomeric proteins. *J. Mol. Biol.*, **196**, 641–656.

Neuvirth,H., Raz,R. and Schreiber,G. (2004) ProMate: a structure based prediction program to identify the location of protein–protein binding sites. *J. Mol. Biol.*, **338**, 181–199.

Oliveira,L., Paiva,P.B., Paiva,A.C. and Vriend,G. (2003) Identification of functionally conserved residues with the use of entropy-variability plots. *Proteins*, **52**, 544–552.

Panchenko,A.R., Kondrashov,F. and Bryant,S. (2004) Prediction of functional sites by analysis of sequence and structure conservation. *Protein Sci.*, **13**, 884–892.

Pazos,F. and Sternberg,M.J. (2004) Automated prediction of protein function and detection of functional sites from structure. *Proc. Natl Acad. Sci. USA*, **101**, 14754–14759.

Pupko,T., Bell,R.E., Mayrose,I., Glaser,F. and Ben-Tal,N. (2002) Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics*, **18** (suppl), S71–S77.

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

Sander,C. and Schneider,R. (1991) Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins*, **9**, 56–68.

Sayle,R.A. and Milner-White,E.J. (1995) RASMOL: biomolecular graphics for all. *Trends Biochem. Sci.*, **20**, 374–376.

Schueler-Furman,O. and Baker,D. (2003) Conserved residue clustering and protein structure prediction. *Proteins*, **52**, 225–235.

Shanahan,H.P., Garcia,M.A., Jones,S. and Thornton,J.M. (2004) Identifying DNA-binding proteins using structural motifs and the electrostatic potential. *Nucleic Acids Res.*, **32**, 4732–4741.

Sridharan,S., Nicholls,A. and Honig,B. (1992) A new vertex algorithm to calculate solvent accessible surface area. *Biophys. J.*, **61**, A174.

Teplyakov,A.V., Kuranova,I.P., Harutyunyan,E.H., Vainshtein,B.K., Frommel,C., Hohne,W.E. and Wilson,K.S. (1990) Crystal structure of thermitase at 1.4 Å resolution. *J. Mol. Biol.*, **214**, 261–279.

Tsodikov,O.V., Record,M.T., Jr and Sergeev,Y.V. (2002) Novel computer program for fast exact calculation of accessible and molecular surface areas and average surface curvature. *J. Compt. Chem.*, **23**, 600–609.

Valdar,W.S. and Thornton,J.M. (2001a) Conservation helps to identify biologically relevant crystal contacts. *J. Mol. Biol.*, **313**, 399–416.

Valdar,W.S. and Thornton,J.M. (2001b) Protein–protein interfaces: analysis of amino acid conservation in homodimers. *Proteins*, **42**, 108–124.

Waksman,G., Shoelson,S.E., Pant,N., Cowburn,D. and Kuriyan,J. (1993) Binding of a high affinity phosphotyrosyl peptide to the Src SH2 domain: crystal structures of the complexed and peptide-free forms. *Cell*, **72**, 779–790.

Wallace,A.C., Laskowski,R.A. and Thornton,J.M. (1995) LIGPLOT: a program to generate schematic diagrams of protein–ligand interactions. *Protein Eng.*, **8**, 127–134.

Xu,W., Harrison,S.C. and Eck,M.J. (1997) Three-dimensional structure of the tyrosine kinase c-Src. *Nature*, **385**, 595–602.

Yao,H., Kristensen,D.M., Mihalek,I., Sowa,M.E., Shaw,C., Kimmel,M., Kavraki,L. and Lichtarge,O. (2003) An accurate, sensitive, and scalable method to identify functional sites in protein structures. *J. Mol. Biol.*, **326**, 255–261.