doi:10.1006/jmbi.2001.4474 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (2001) **307**, 447–463

JMB

ConSurf: An Algorithmic Tool for the Identification of Functional Regions in Proteins by Surface Mapping of Phylogenetic Information

Aharon Armon[1], Dan Graur[2] and Nir Ben-Tal[1]*

[1]*Department of Biochemistry*

[2]*Department of Zoology*
George S. Wise Faculty of Life Sciences, Tel Aviv University
Ramat Aviv 69978, Israel

**Corresponding author*

Experimental approaches for the identification of functionally important regions on the surface of a protein involve mutagenesis, in which exposed residues are replaced one after another while the change in binding to other proteins or changes in activity are recorded. However, practical considerations limit the use of these methods to small-scale studies, precluding a full mapping of all the functionally important residues on the surface of a protein. We present here an alternative approach involving the use of evolutionary data in the form of multiple-sequence alignment for a protein family to identify hot spots and surface patches that are likely to be in contact with other proteins, domains, peptides, DNA, RNA or ligands. The underlying assumption in this approach is that key residues that are important for binding should be conserved throughout evolution, just like residues that are crucial for maintaining the protein fold, i.e. buried residues. A main limitation in the implementation of this approach is that the sequence space of a protein family may be unevenly sampled, e.g. mammals may be overly represented. Thus, a seemingly conserved position in the alignment may reflect a taxonomically uneven sampling, rather than being indicative of structural or functional importance. To avoid this problem, we present here a novel methodology based on evolutionary relations among proteins as revealed by inferred phylogenetic trees, and demonstrate its capabilities for mapping binding sites in SH2 and PTB signaling domains. A computer program that implements these ideas is available freely at: http://ashtoret.tau.ac.il/ ~ rony

Keywords: molecular recognition; protein-protein interactions; protein modeling; phylogenetic trees

Introduction

Mutual interactions between proteins and between proteins and peptides, nucleic acids or ligands play a vital role in every biological process. Thus, detailed understanding of the mechanism of these processes requires the identification of functionally important amino acids at the protein surface that mediate these interactions. Studies to determine the three-dimensional (3D) structure of protein complexes are useful to single out residues at protein-protein interfaces that are functionally important. However, it is often difficult to determine the 3D structure of protein complexes, and often only the structures of the unbound proteins (or domains) are available. In such cases, it is common to carry out tedious mutagenesis studies to determine functionally important residues. However, because of the amount of work required for such an approach, a number of entries in the RCSB Protein Data Bank[1] exist, for which we have only partial information about the function; for example, we may know that a certain protein is a kinase without being able to map the exact location of its active site. The fraction of such entries is expected to increase rapidly due to the different structural genomics initiatives.[2,3]

An alternative method to identify functionally important residues in proteins of known 3D

Abbreviations used: MSA, multiple sequence alignment; ConSurf, consevation surface mapping; PTB, phosphotyrosine binding; rmsd, root-mean-square deviation.

E-mail address of the corresponding author: bental@ashtoret.tau.ac.il

structure is to use evolutionary information, that is, to deduce the importance of residues from their level of conservation in families of homologous proteins. It is well established that residues buried in the protein core are conserved throughout evolution.[4] The reason for buried residues to be evolutionarily conserved is known; the packed structure of proteins tolerates only conservative amino acid replacements, whereas radical replacements, such as exchanges between residues of different sizes, often destabilize the structure of the protein and results in malfunctioning proteins.

Likewise, protein complexes are very sensitive to replacements at the inter-protein interface.[5] Thus, it is reasonable to assume that functionally important residues, which are involved in molecular recognition between proteins (or between proteins and DNA) or in enzymatic activity, should be evolutionarily conserved.[6–10] Indeed, presentations of newly determined protein structures often involve the incorporation of information deduced from sequence analogues of the protein to signal functionally important amino acids. To this end, one usually estimates the level of residue conservation directly from multiple sequence alignment (MSA) of the protein homologues. A key problem with this approach is that in many cases the homologues do not evenly sample the sequence space, e.g. eukaryotes may be overly represented as compared to prokaryotes, or *vice versa*. Thus, a method that properly weights the level of conservation by the evolutionary distance of the proteins from one another would be desirable.

Cohen and co-workers developed such a method.[11–13] Their method, referred to as "The Evolutionary Trace Method", is based on constructing a phylogenetic tree from the MSA. A consensus sequence is then derived for the sequences at each node of the tree, and the level of residue conservation is derived from the variability of the consensus sequences and projected onto the protein surface. The evolutionary trace method was tested on the SH2 and SH3 modular signaling domains and the DNA binding domain of the nuclear hormone receptors[11] It was then used to explore G proteins[12] and zinc binding domains.[13] In all of these cases the method successfully identified surface patches, such as the peptide-binding pocket of SH2 domains, that are known to be functionally important.

The Evolutionary Trace Method was the first attempt to take into account the evolutionary history of a protein family, but despite its overall success in the mapping of functionally important residues on protein surfaces, its treatment of the evolutionary process is only approximate. For example, the phylogenetic tree is built using the UPGMA method (under the PILEUP sequence alignment tool[14,15]). This method is based on the assumption of equal rates of evolution along all branches of the phylogenetic tree, an assumption that had been repeatedly refuted in the past.[16] Following tree reconstruction, the aligned sequences at each node are compared to construct consensus sequences, a procedure that only takes into account identical amino acid residues at a position. The sequences derived from the nodes are, then, compared to form a general consensus sequence. This all-or-none consensus sequence-based method treats all columns with variable amino acid residues as non-conserved, regardless of the physicochemical similarity between them, and may affect the sensitivity of the Evolutionary Trace Method. This issue is considered in the Discussion below.

Results

We introduce here a novel method, referred to as conservation surface-mapping, or ConSurf, for mapping of evolutionarily conserved residues on protein surfaces. The method uses evolutionary trees that are consistent with the MSA, and takes into account the physicochemical distance between the replaced amino acids. It should therefore be more sensitive than the Evolutionary Trace Method.

After obtaining the MSA, ConSurf constructs evolutionary trees that are consistent with it, using the protein parsimony method,[17] which also allows deduction of the amino acid changes that occurred throughout evolution by tracking changes along the branches of each tree. Miyata *et al.*[18] assumed that the physicochemical properties of the residues that are crucial for maintaining the protein fold are conserved, and calculated a physicochemical similarity matrix of the amino acids. Our program evaluates each amino acid exchange by this matrix.

We apply this method on the well-studied SH2 and PTB domains. The SH2 domain (for Src homology 2) is a phosphotyrosine binding module that was located on various proteins involved in signal transduction.[19–21] SH2 domains have been exploited widely for structural-specificity studies.[21–23] Applying the method to SH2 domains also allows us to compare our results with those obtained by the Evolutionary Trace Method.

Another tyrosine phosphate recognizing module is the PTB (phosphotyrosine binding) domain, which is flexible compared to the SH2 domain in terms of peptide binding and recognition.[24–26] This flexibility is probably reflected in the higher sequence variability of the PTB domain family. We mapped the level of residue conservation on representatives from this family to examine how the differences in specificity and sequence variability are reflected in the results obtained by ConSurf.

SH2 domains

The structure of many SH2 domains is available, and we chose to present the conservation map on the structure of the domain of human Src, which is known both in complex with one of the native peptides[21] and in the context of the intact protein.[19] It is noteworthy that C^{α} rmsd between SH2 domains from different sources is small and the conservation pattern is, in essence, independent of the structure used for the mapping (see below). We collected homologuous sequences, aligned them, constructed a phylogenetic tree (Figure 1), calculated conservation for each position in the alignment and presented the conservation level on the molecular surface of the domain as described in Methodology, below. We also analyzed the changes in the conservation pattern on the molecular surface as more clades of the tree in Figure 1 were added. To this end, we used the phylogenetic tree drawn by CLUSTAL W,[27] using the neighbor joining method.[28]

The sequences comprising each clade of the phylogenetic tree of Figure 1 are listed in Table 1. The ConSurf results obtained by consecutively adding more branches to the analysis are presented in Figure 2. (Since the SH2 tree of Figure 1 is unrooted, including more branches does not necessarily indicate an increase in the mean evolutionary distances among the sequences.) It is evident from the Figure that as the clade size is increased, the observed conservation patch decreases in size until about two-thirds of the phylogenetic tree is included in the calculations and the conservation pattern converges (Figure 2G). A comparison of the converged conservation pattern (Figure 3(a)) and the binding contact, determined from the structure of the complex (Figure 3(c)), shows a nearly perfect fit between the highly conserved patch and the area of intensive contact (both marked in dark red).

Figure 3(a) and (b) present a comparison of the residue conservation pattern obtained for SH2 domains of two different structures. Figure 3(a) was obtained using the structure of human Src SH2 in complex with a native peptide,[21] and Figure 3(b) was obtained using the SH2 domain from the structure of the intact Src.[19] It is evident from the Figure that the conservation pattern is very similar despite the different structures that were used.

The C-terminal tail of Src contains a sequence that resembles the ligand motif, and upon phosphorylation it may bind to the SH2 domain, thus limiting its access to Src signaling partners. The contact projection obtained for the peptide (Figure 3(c)) and for the pseudo-peptide (Figure 3(d)) are fairly similar, and the evolutionary conservation pattern of Figure 3(a) and (b) fits them well. Interestingly, the conservation pattern of the SH2 domains fits better with the contact projection obtained for the peptide than for the pseudo-peptide, which may be related to the higher affinity of the peptide compared to the pseudopeptide to SH2 domains.[22]

Figure 4 shows the ''back-side'' of the SH2 domain from intact Src,[19] which is known to interact with the SH3 and catalytic domains of Src. The patch that is in contact with the SH3 domain (highlighted by the green circle in Figure 4(c)) is mirrored by the conservation pattern (Figure 4(a)). However, the area that is in contact with the catalytic domain (marked in magenta) shows only average conservation, despite its presumed functional relevance for interaction between the catalytic and SH2 domains of Src, by oppositely charged amino acids.[19] This inconsistency problem is solved by surface mapping of residue conservation, taking into account the Src clade of the tree only (Figure 1, A; Table 1, row A). This clade consists of 24 sequences, including Fyn, Hck, Lck, Yes, and Fgr. The conservation map obtained for the Src clade (Figure 4(b)) matches the contact area at the backside of the SH2 domain very well. Cys245 (circled in white) is the contact site for the SH2-linker loop, which seems to function in SH3 regulation[19] rather than in stabilizing the SH2 interaction with the whole protein. This may explain the discrepancy with its low level of conservation.

The fact that the conservation at the back-side of Src is only seen in the Src clade and not throughout the entire family may indicate that while peptide binding at the ''front'' end is typical for all (or most) of the members of the family, the interactions of Src SH2 domains with the SH3 and catalytic domains are limited to members of the Src clade only. It further suggests that the Src inhibitory mechanism, involving the internal contacts between the SH2 domain and the SH3 and cataly-

Table 1. Clades of the phylogenetic tree of the SH2 domain

Clade index[a]	Number of sequences[b]	From	To	Average conservation[c]
A	24	SRC_HUMAN	FGR_MOUSE	0.837
B	35	SCR1_DROME	BLK_HUMAN	0.811
C	40	SCR1_DROME	SRK1_SPOLA	0.784
D	42	SCR1_DROME	CSK_HUMAN	0.716
E	59	SCR1_DROME	ABL_FSVHY	0.648
F	89	SCR1_DROME	GTPA_HUMAN	0.675
G	97	YKF1_CAEEL	GTPA_HUMAN	0.659
H	108	SRM_MOUSE	GTPA_HUMAN	0.702
I	108	NCK_HUMAN	GTPA_HUMAN	0.710
J	127 (all)			0.685

[a] The clade index used in Figure 1.
[b] The number of sequences in the clade.
[c] The average conservation calculated for the clade.

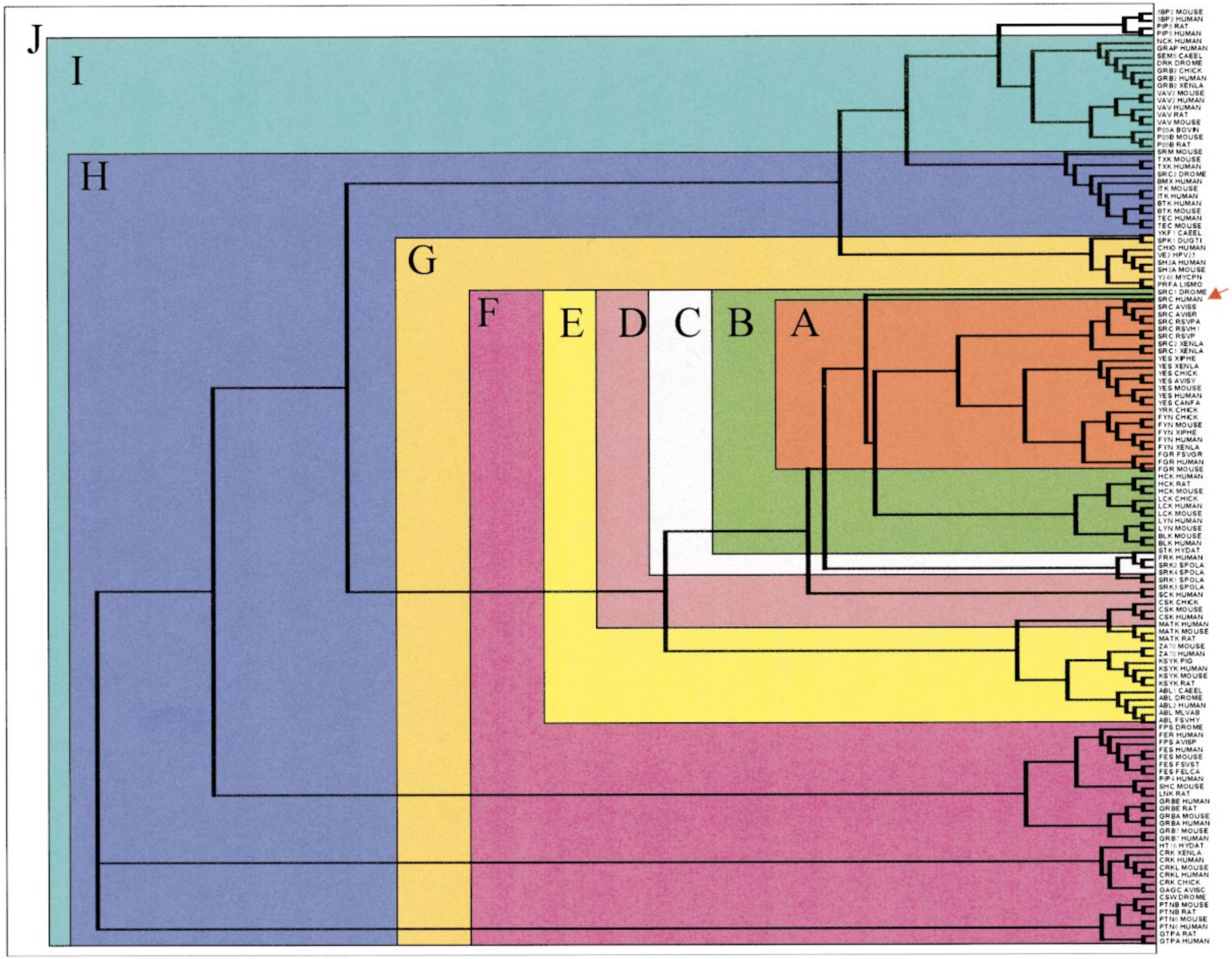

Figure 1 (*legend shown opposite*)

tic domains, may be unique for the Src clade of the tree.

Overall, the conservation pattern obtained for clade A (Figures 2A and 4(b)) is much wider than that of the entire phylogenetic tree (Figures 2J and 4(a)). This is presumably in part because the SH2 domains in clade A are functionally more related to each other than to other SH2 domains and in part because of insufficient divergence time within clade A.

PTB domains

The human adapter protein Shc contains both an SH2 domain, and another tyrosine phosphate binding module, the ''PTB domain''.[24] Despite the fact that the two domains recognize mainly protein fragments containing tyrosine phosphate, they differ significantly in their sequence and structural topology.[22]

PTB domains share very low levels of sequence similarity[24] and aligning them properly is difficult. Therefore, we used the MSA,[29] as described in Methodology, below, to construct the phylogenetic tree of Figure 5, from which we calculated conservation grades for each position in the alignment. The residue conservation grades obtained for three PTB domains of known 3D structure are presented in Figure 6. Despite the significant sequence divergence in the family, ConSurf was able to map the peptide-binding site in these domains. Again, it is evident from the Figure that the pattern of phylogenetic conservation depends very weakly on the structure; a patch of conserved residues was mapped onto the peptide binding pocket in the three different PTB structures tested here.

X11 is a neuron-specific protein containing a PTB domain, which was found to bind to the cytoplasmic domain of the β-amyloid precursor protein (β-APP) with high affinity and specificity.[30] As in other PTB domains, ligand binding to the X11 protein involves residues C-terminal to the phosphotyrosine, especially in the NPxY motif.[31] Figures 7 and 8(b) display the residue conservation mapping of this domain. It is evident from the Figures that residues, such as Leu413 and Ile416 that are hydrogen bonded to Asn (−3) of the peptide, Tyr483 that is linked to the Pro (−2) of the beta turn, and Ser417 that is linked to the tyrosine itself (Figure 8(b)), all closely involved in binding to the NPXY motif,[31] were also found to be highly conserved by ConSurf. (The numbers of peptide residues refer to the tyrosine residue of the NPxY motif, which was denoted at residue 0.) Figure 7 also shows that the area that binds Tyr (−5) was mapped as highly to moderately conserved. This position seems to be conserved due to its hydrophobicity,[31] as was described for the Shc PTB domain.[32] Our data show slightly above average conservation for residues that provide a hydrophobic surface, against which the aromatic rings of Phe (+2) and (+3) can pack. Mutating each of these residues to alanine decreases peptide affinity by about tenfold, but they are not strictly conserved throughout the PTB family.[31] Conservation mapping, using ConSurf, of the X11 subfamily shows a high level of conservation for this area (data not shown).

Although we noticed a close contact between the peptide and the domain in the N terminus region of the peptide, and mutational data indicate that they are important, their contact partners are only partially conserved throughout the family. This lack of conservation of the protein surface is in agreement with the fact the physicochemical nature of the complementary peptide surface is variable. Thus, the polar nature of the peptide residue at position (−7) in APP peptide (the X11 ligand) stands in contrast to high affinity peptides for Shc and IRS-1, which tend to contain hydrophobic residues at this position.[33]

Discussion

We developed a new method, referred to as ConSurf, for mapping evolutionarily conserved regions on the surface of proteins of known 3D structure. ConSurf aligns sequence homologues of the protein whose structure is known, and uses the alignment to construct phylogenetic trees. The trees are then used to infer the presumed amino acid exchanges that occurred throughout the evolution of the protein. Each exchange is then weighted by the physicochemical distance between the exchanged amino acid residues. By mapping these grades onto the surface of the SH2 and PTB domains, we showed that the patches of conserved residues correlate well with the known functional regions of the domains. In the following, we compare the conservation pattern of these two protein modules in light of the difference in their functions. We then discuss the implications and limitations of ConSurf.

Comparison of the tyrosine binding site of SH2 and PTB domains

The SH2 and the PTB domains were initially characterized as phosphotyrosine binding modules in signaling proteins.[22,34] These modules vary significantly in sequence and structure; PTB domains adapt a pleckstrin homology-like fold[31,35,36] that is

Figure 1. Phylogenetic reconstruction for sequences of SH2 domains. The phylogenetic tree was built using the neighbor joining method in the CLUSTAL W package[27] for the same MSA data used for the tree's reconstruction by Protpars. Clades of the tree are marked from A to J. The sequence names in each clade are listed in Table 1. The red arrow marks the position of human Src, analyzed here, in the tree.

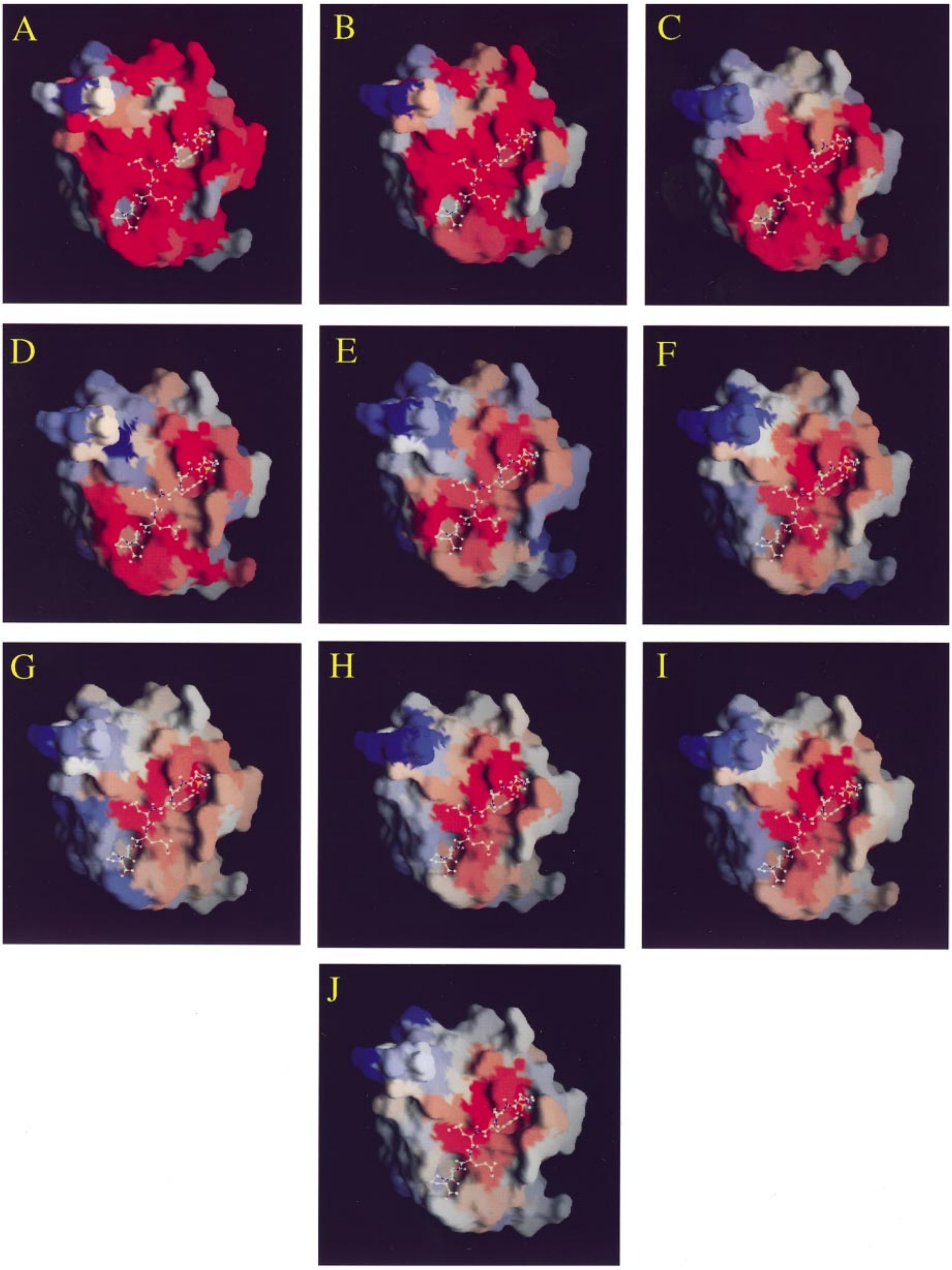

Figure 2 (*legend shown opposite*)

distinct from that of SH2 domains. SH2 and PTB domains also differ from each other in their mode of ligand recognition and ligand fold. While the SH2 ligand binds in extended conformation[19,20,22] the NPxY motif of the SHC ligand forms a β-turn structure prior to binding.[35,36] This fold was found also in other PTB-ligand complexes.[31,37,38] In addition, while tyrosine phosphorylation is crucial for ligand recognition in most SH2 domains detected so far, it appears to be much less crucial in PTB domains.[30,31,39] The *Drosophila* Numb PTB domain, which binds phosphorylated as well as unphosphorylated NPxY-containing peptide, also binds YIGPYL with high affinity,[40] and therefore provides an example of the flexibility of the PBT domain in ligand recognition. Other members of the PTB domain family do not require posphorylation or even the presence of tyrosine on their binding ligands.[26,31]

Overall, sequence similarity among PTB domains is significantly lower than among SH2 domains. In fact, the similarity is so low that structural information was used for the alignment and discovery of members of PTB domain family.[41] The sequence variability in PTB domains is probably related to the fact that they recognize a diversity of peptides including unphosphorylated peptides and the variance in the conformation of the bound peptides (e.g. Figure 6).

The differences in ligand selectivity between the SH2 and PTB domains are clearly reflected in the residue conservation patterns obtained for these domains (Figure 8(a) and (b)). The deep phosphotyrosine binding pocket is highly conserved in SH2 domains (Figure 8(a)), which correlates well with the high affinity of the domains to phosphotyrosine-containing peptides.[23,42] It may also explain why tyrosine phosphorylation is a prerequisite for peptide-binding in the vast majority of SH2 domains.[22] In contrast, we detected only average conservation in the tyrosine-binding site of the X11 PTB domain, which confers with the fact that phosphorylation is not obligatory for peptide-binding in PTB domains.[30,31,39]

Comparisons with other methods

The main problem we sought to solve in conservation grading is the multiplicity of sequences in sequence alignment. This results from the uneven representation of amino acid sequences in protein databases.[43,44] Thus, counting the differences between each pair of amino acids in the alignment may lead to a misestimation of replacement frequency, even after weighing each pair comparison by its sequence distance.[44] Using a tree-based method, we can infer the branches in which specific amino acid changes occurred. We, thus, solve the problem of non-independent sampling due to plesiomorphy, i.e. similarity due to maintenance of the ancestral state. This ''filtering'' allows us to use sequence alignments without being bothered by sample size and evenness of sampling.

Another tree-based approach is the Evolutionary Trace Method.[11] However, this method should be inferior to ConSurf, since it uses phylogenetic trees that were built based on the assumption of equal rates of evolution in all branches, and since it is an all-or-none consensus sequence-based method, as mentioned above. The use of the average conservation along the sequence in ConSurf enabled us to normalize according to the number of sequences in the tree or clade. The qualitative nature of the evolutionary trace method does not admit the use of average conservation. Comparing the results obtained using the two methods, we discover a similar mapping of the conservation of the peptide binding face (compare Figure 2J here with Figure 3(a) of Lichtarge *et al.*[11]). Rotating the molecule by 135 °, ConSurf detects the contact area of the SH2 domain with the other domains of the Src protein (Figure 4), which were not identified by the Evolutionary Trace Method (Figure 3(c) of Lichtarge *et al.*[11]).

Implications and limitations

We show here how conservation mapping based on evolutionary tree reconstruction, maximum parsimony tracing, and physicochemical grading, can assist in identifying functional regions on protein surfaces. However, the quality of the results depends on the quality of the sequence alignment and the phylogenetic tree reconstruction. When applied to protein families of widely diverse functions, or when the MSA input includes proteins from different families, the ConSurf analysis might show a mosaic of traits from different evolutionarily related proteins. Alternatively, if the ConSurf analysis is carried out using MSA containing sequences of limited diversity, the picture obtained

Figure 2. Mapping of evolutionary conservation on the molecular surface of SH2 domains: the peptide binding face of the domain. The conservation map is presented on the structure of the SH2 domain in complex with the high affinity peptide Y(p)EEI,[21] taking into account different number of sequences from the evolutionary tree in Figure 1. (a) The conservation map obtained using clade A of the phylogenetic tree in Figure 1. (b) The conservation map obtained using clades A and B of the phylogenetic tree in Figure 1. (c) The conservation map obtained using clades A to C of the phylogenetic tree in Figure 1. (d) The conservation map obtained using clades A to D of the phylogenetic tree in Figure 1. (e) The conservation map obtained using clades A to E of the phylogenetic tree in Figure 1. (f) The conservation map obtained using clades A to F of the phylogenetic tree in Figure 1. (g) The conservation map obtained using clades A to G of the phylogenetic tree in Figure 1. (h) The conservation map obtained using clades A to H of the phylogenetic tree in Figure 1. (i) The conservation map obtained using clades A to I of the phylogenetic tree in Figure 1. (j) The conservation map obtained using clades A to J of the phylogenetic tree in Figure 1, i.e. the whole tree. Residue conservation is color-coded onto the molecular surface of the domain: dark blue corresponds to maximal variability, white corresponds to average conservation level and dark red to maximal conservation. The peptide is shown as bond lines. The picture was drawn using GRASP.[56]

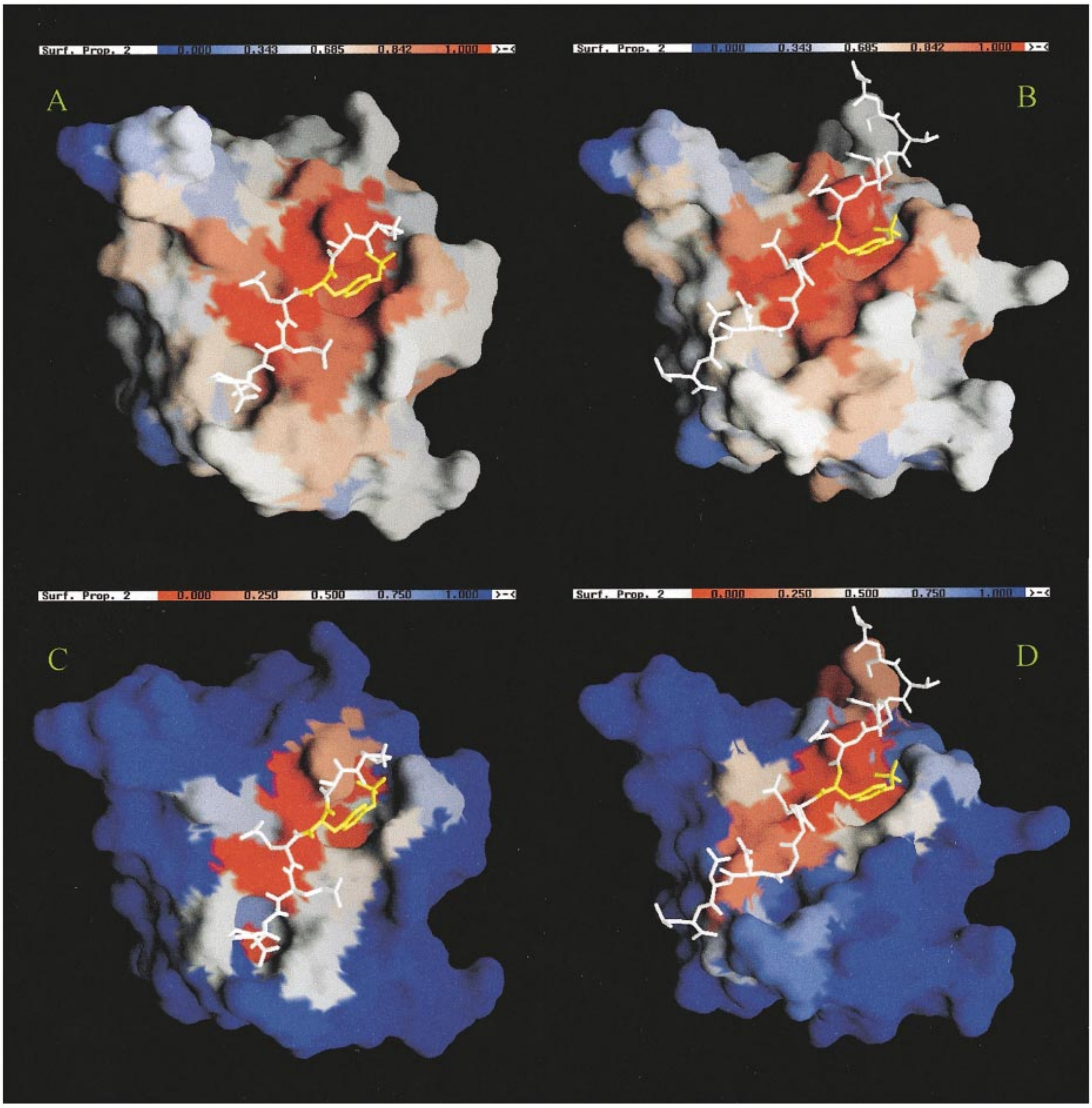
Surf. Prop. 2 0.000 0.343 0.685 0.842 1.000 >-<
A
Surf. Prop. 2 0.000 0.343 0.685 0.842 1.000 >-<
B
Surf. Prop. 2 0.000 0.250 0.500 0.750 1.000 >-<
C
Surf. Prop. 2 0.000 0.250 0.500 0.750 1.000 >-<
D

Figure 3. Conservation ((a) and (b)) and contact ((c) and (d)) maps of the peptide ((a) and (c)) and pseudo-peptide ((b) and (d)) ligands of SH2 domains. (a) Conservation pattern for the peptide. (b) Conservation pattern for the pseudo peptide. (c) Contact mapping for the peptide. (d) Contact mapping for the pseudo-peptide. The structure of human SH2 domain in complex with a native peptide[21] was used in (a) and (c). The structure of human Src[19] was used in (b) and (d) but for clarity, only the SH2 domain and C-terminal pseudo-peptide are displayed. For contact mapping, residues with no contact with the peptide are blue. Residues that their water accessible surface area changes by 50 % are white and residues that are completely buried upon complex formation are red. The picture was drawn using GRASP[56] and conservation is color-coded as in Figure 2. The peptide is shown with bond lines and the tyrosine residue is in yellow.

is a mixture of functionally important residues and shortness of divergence time (e.g. Figure 2A). This puts a limit on the usefulness of ConSurf for the analysis of sub-clades of highly conserved families such as the SH2 domains, but provides a clue on differentiation to specific functions.

In its current implementation, ConSurf uses the amino acid similarity matrix that was derived by Miyata *et al.*[18] based on the physicochemical relations between the amino acids. However, this matrix can be readily replaced by empirical amino acid replacement matrices, such as that by Dayhoff *et al.*[45] Notwithstanding the use of physicochemical distances, evolutionary replacement matrixes are incorporated into ConSurf in an indirect manner, since both the search for homologous sequences and the MSA reconstruction involve an intensive use of such matrices. A statistically significant

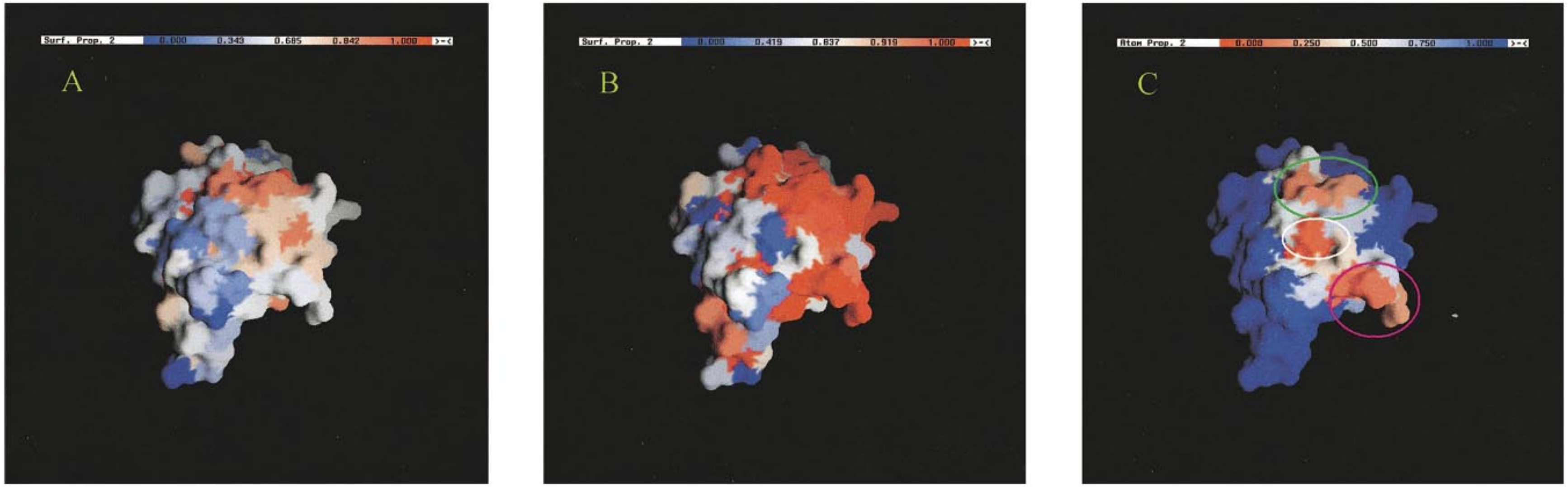

Figure 4. Mapping of evolutionary conservation on the molecular surface of SH2 domains: interaction with the SH3 and catalytic domains. (a) The conservation map obtained using the entire SH2 phylogenetic tree in Figure 1. (b) Same as (a) but using the sequences in the Src clade of the tree (Figure 1 A; Table 1, row A). (c) Contact mapping between the SH2 domain and the SH3 and catalytic domains. The structure of intact Src was used[19] and the protein was rotated by 135° around the *x*-axis from the orientation of Figure 2. The picture was drawn using GRASP[56] and conservation is color-coded as in Figure 2.

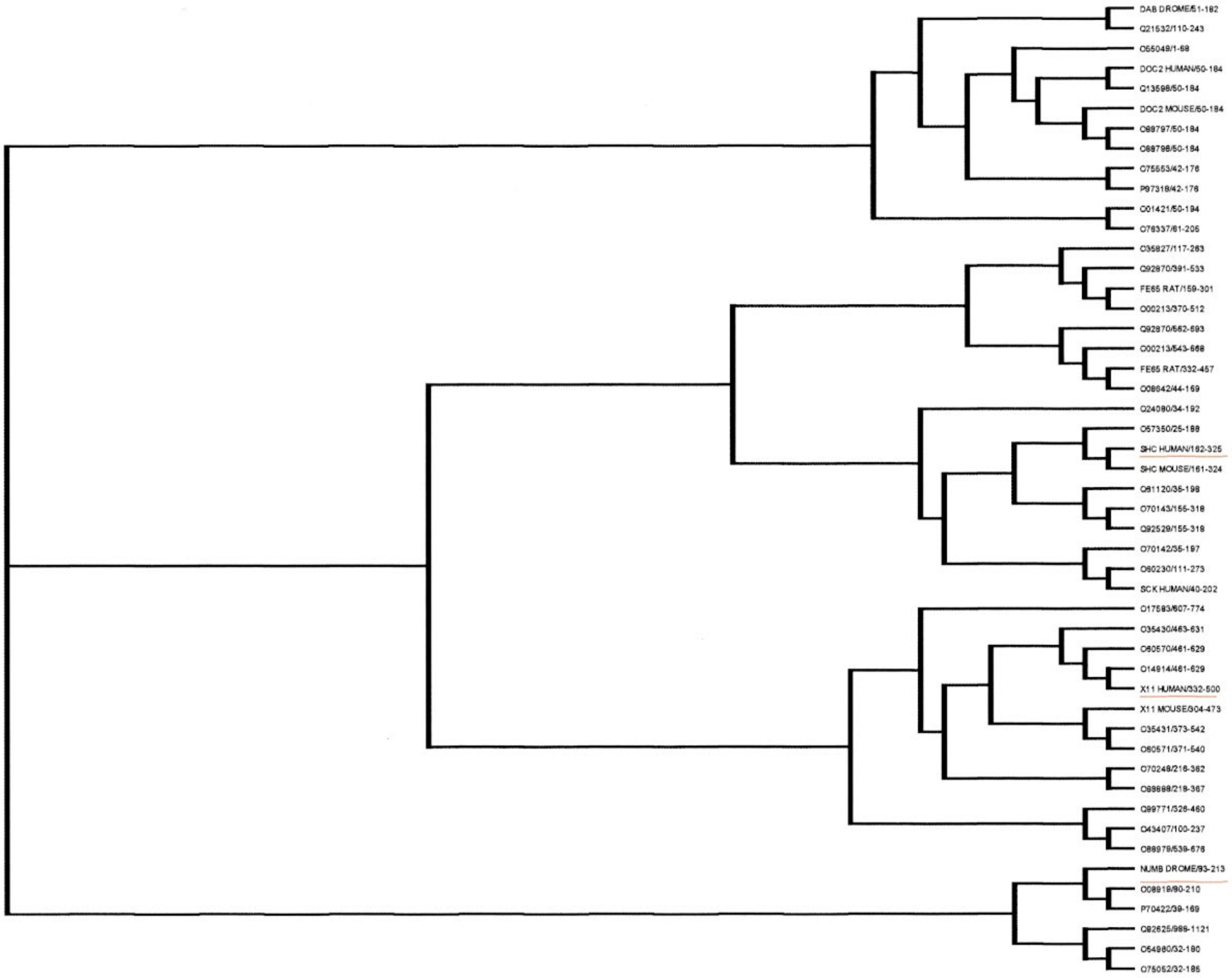

Figure 5. A representative phylogenetic tree for the PTB domains. The proteins for which we present conservation maps are underlined in red.

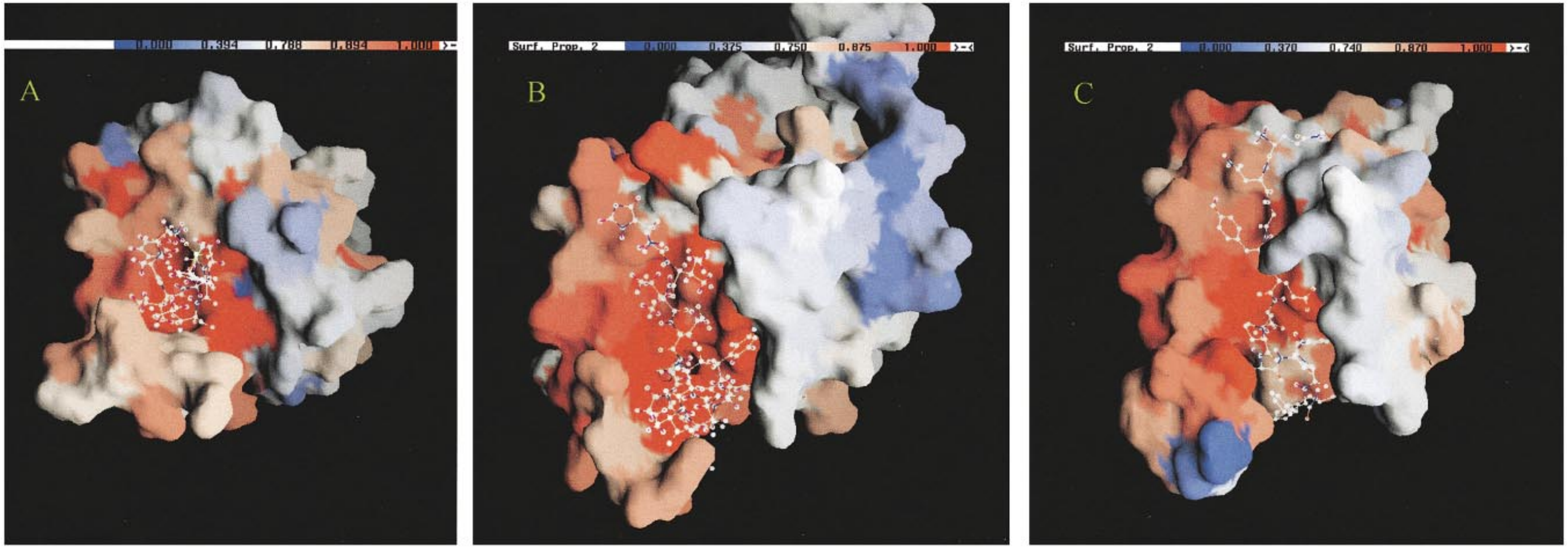
0.000 0.394 0.788 0.894 1.000 >-<
A
Surf. Prop. 2 0.000 0.375 0.750 0.875 1.000 >-<
B
Surf. Prop. 2 0.000 0.370 0.740 0.870 1.000 >-<
C

Figure 6. Conservation maps of the PTB domains from: (a) the *Drosophila* numb protein;[40] (b) human adapter protein SHC;[35] and (c) human neuron-specific protein, X11.[31] The structures were superimposed using InsightII (MSI, San Diego, CA), based on coordinates from the HOMSTRAD structural alignment database[59] The picture was drawn using GRASP[56] and conservation is color-coded as in Figure 2. The peptide is shown as a ball and stick model.

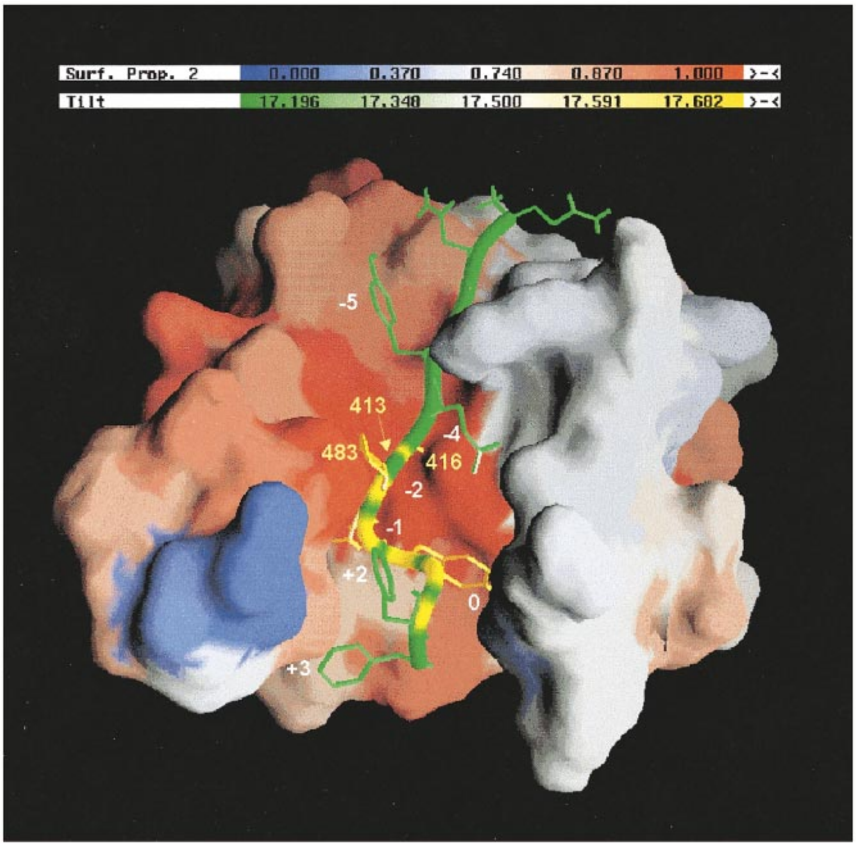

Figure 7. The peptide binding site of the PTB domain of X11.[31] Amino acid conservation grades are color-coded on the molecular surface or the domain, and the peptide is colored according to its tilt, to emphasize the β-turn structure (yellow). The numbers correspond to amino acids of the PTB domain (yellow) and the peptide (white).

number of co-crystallized binary protein complexes exists (several thousand, including dimers; the number depends on the criteria and varies in different studies). We plan to use these complexes to derive an amino acid replacement matrix specifically for inter-protein interfaces.

Since the tree we use is unrooted, tracing the pattern of differentiation on the protein surface by conservation mapping may identify not only evolutionarily relevant clade-specific patterns, but also evolutionarily irrelevant neighbor-specific patterns. This shortcoming may be rectified by using rooted trees in cases where the root is known. Reassuringly, the conservation pattern obtained for the SH2 domains using ConSurf is very similar to that obtained when the conservation grades were calculated using the maximum likelihood principle and a single rooted phylogenetic tree (Pupko *et al*., unpublished results).

One of the main limitations of ConSurf is its low resolution. Since it is based on phylogenetic information on residue conservation, it cannot provide information at atomic resolution. Moreover, since there are slight differences in function even between closely related proteins, it may be risky to use ConSurf to identify specific residues that are functionally important for a specific protein. ConSurf should best be regarded as a tool for the identification of functionally important patches of residues on the surface of proteins of known 3D structure, some of which are discovered without known function or ligand.[46] The conservation pattern detected by ConSurf should guide detailed experimental work towards the identification of the exact residues involved in molecular recognition between the protein at hand and its ligands. The low resolution of ConSurf may, however, be regarded as an advantage, since it allows for the use of ConSurf even in cases where only low-resolution or model structures of the protein are available.

In any event, the analysis provided by ConSurf only provides information on the functionally important surface regions of the protein under study and does not reveal the nature of the binding partner(s). In some cases, however, the type of amino acids conserved on the surface of the protein may provide some clues on the binding partner(s).

ConSurf, which is based on CLUSTAL W alignment, Protein Parsimony tree reconstruction, and tracing changes in the sequence positions, is freely available for academic use at www.ashtoret.tau.ac.il/ ~ rony It is written to allow users to either align a set of homologous proteins, the sequence of which is taken as input to ConSurf, or to provide an external sequence alignment as input to ConSurf.

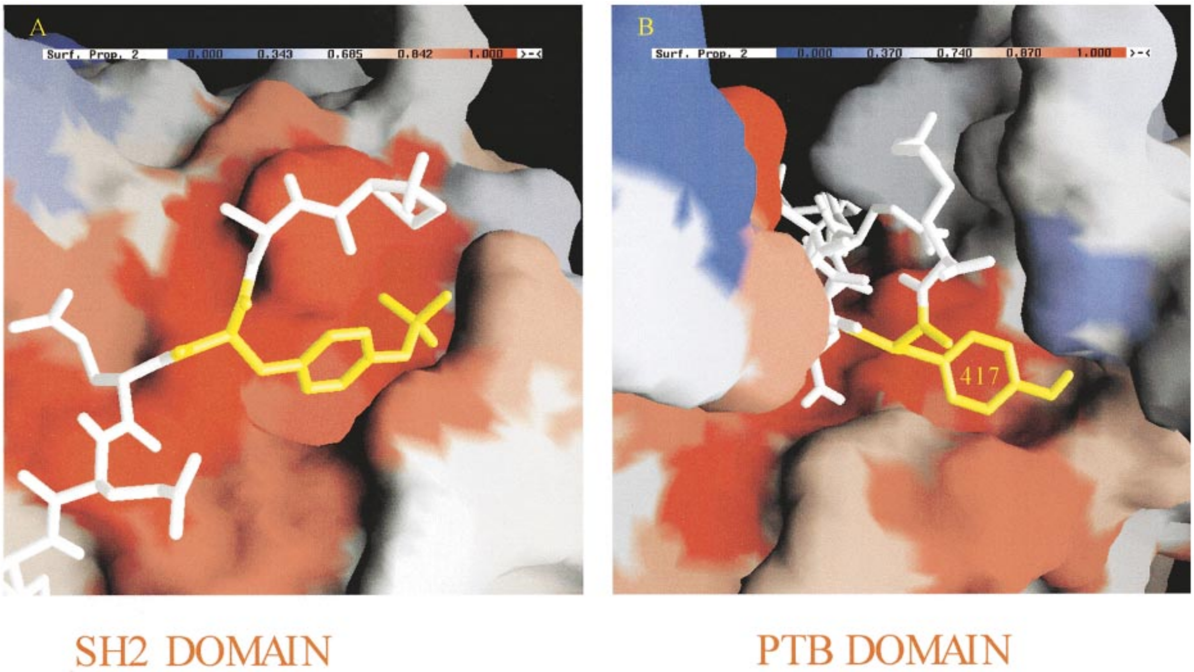

Figure 8. A close look at the amino acid conservation grades in the tyrosine-binding pocket of (a) the human Src SH2[21] and (b) x11 PTB[31] domains. The picture was drawn using GRASP[56] and conservation is color-coded as in Figure 2. The peptide is shown as bond lines and Ser417 is marked in yellow.

Methodology

Searching for homologous sequences

We used the Smith & Waterman[47] algorithm, with default exchange matrix, gap opening penalty of 10 and gap extension penalty of 0.5, to collect sequence homologues of the protein of known 3D structure from the SwissProt database.[48] Homology search using this non-heuristic procedure (i.e. comparing the query protein to all sequences in the database) was found to be superior in terms of sensitivity and selectivity compared to popular heuristic alternatives such as FASTA and BLAST.[49] We limited our collection to sequences associated with *E* score values lower than 0.05, as an indicator for biologically significant matches.[50] The homologous sequences were then collected from the search output and identical sequences were filtered out before further analysis. The initial methionine was cut out from each sequence.

Generating multiple sequence alignment (MSA)

The homologous sequences gathered in the previous stage were formatted for FASTA before alignment. We mainly used the CLUSTAL W method that uses different amino acid replacement matrices depending on the sequence similarity.[27] These matrices, mainly from the PAM and BLOSSUM series, are based on counting the frequencies of amino acid changes in confirmed alignment.[51] They allow us to weight each amino acid exchange between aligned sequences by the estimated probability of obtaining it in relative sequences of a given evolutionary distance. For each evolutionary distance there is an optimal matrix in the mentioned series. Since the collected sequences were of variable evolutionary distances from one another, we found it comfortable to use CLUSTAL W, which exchanges the use of specific matrices in relation to the distances between sequences.

Phylogenetic reconstruction

After obtaining the MSA, we constructed evolutionary tree(s) consistent with it using the protein parsimony method.[17] This method also allows us to deduce the amino acid changes that occurred throughout evolution by tracking along the branches of the tree.[52] In practice, we used the PROTPARS program from the PHYLIP package.[17]

Grading amino acid exchanges

Each exchange between amino acids i and j ($i = 1,2,\ldots 20$; $j = 1,2,\ldots 20$) was multiplied by a weight factor according to the physicochemical distance between the amino acids[18] (Table 2). Miyata *et al.*[18] have estimated the physicochemical distance (M_{ij}) between each pair of amino acids by:

$$M_{ij} = [(\Delta p_{ij}/\sigma_{\mathrm{p}})^2 + (\Delta v_{ij}/\sigma_{\mathrm{v}})^2]^{(1/2)} \qquad (1)$$

where Δp_{ij} and Δv_{ij}, calculated by Grantham,[53] are the differences in polarity and volume between the

Table 2. The amino acid pair distance used in this study

	C	P	A	G	S	T	Q	E	N	D	H	K	R	V	L	I	M	F	Y	W	–	X	Z	B
C	0	1.33	1.39	2.22	1.84	1.45	2.48	3.26	2.83	3.48	2.56	3.27	3.06	0.86	1.65	1.63	1.46	2.24	2.38	3.34	6	6	2.87	3.15
P		0	0.06	0.97	0.56	0.87	1.92	2.48	1.8	2.4	2.15	2.94	2.9	1.79	2.7	2.62	2.36	3.17	3.12	4.17	6	6	2.2	2.1
A			0	0.91	0.51	0.9	1.92	2.46	1.78	2.37	2.17	2.96	2.92	1.85	2.76	2.69	2.42	3.23	3.18	4.23	6	6	2.19	2.08
G				0	0.85	1.7	2.48	2.78	1.96	2.37	2.78	3.54	3.58	2.76	3.67	3.6	3.34	4.14	4.08	5.13	6	6	2.63	2.17
S					0	0.89	1.65	2.06	1.31	1.87	1.94	2.71	2.74	2.15	3.04	2.95	2.67	3.45	3.33	4.38	6	6	1.85	1.59
T						0	1.12	1.83	1.4	2.05	1.32	2.1	2.03	1.42	2.25	2.14	1.86	2.6	2.45	3.5	6	6	1.47	1.73
Q							0	0.84	0.99	1.47	0.32	1.06	1.13	2.13	2.7	2.57	2.3	2.81	2.48	3.42	6	6	0.42	1.23
E								0	0.85	0.9	0.96	1.14	1.45	2.97	3.53	3.39	3.13	3.59	3.22	4.08	6	6	0.42	0.88
N									0	0.65	1.29	1.84	2.04	2.76	3.49	3.37	3.08	3.7	3.42	4.39	6	6	0.92	0.33
D										0	1.72	2.05	2.34	3.4	4.1	3.98	3.69	4.27	3.95	4.88	6	6	1.18	0.33
H											0	0.79	0.82	2.11	2.59	2.45	2.19	2.63	2.27	3.16	6	6	0.64	1.51
K												0	0.4	2.7	2.98	2.84	2.63	2.85	2.42	3.11	6	6	1.1	1.95
R													0	2.43	2.62	2.49	2.29	2.47	2.02	2.72	6	6	1.29	2.19
V														0	0.91	0.85	0.62	1.43	1.52	2.51	6	6	2.55	3.08
L															0	0.14	0.41	0.63	0.94	1.73	6	6	3.11	3.8
I																0	0.29	0.61	0.86	1.72	6	6	2.98	3.68
M																	0	0.82	0.93	1.89	6	6	2.71	3.39
F																		0	0.48	0.11	6	6	3.2	3.99
Y																			0	1.06	6	6	2.85	3.67
W																				0	6	6	3.75	4.64
–																					0.5	6	6	6
X																						0	6	6
Z																							0	1.05
B																								0

The values were taken from Miyata *et al.*[18]
See the text for details.

amino acids, and σ_p and σ_v are the corresponding standard deviations. An implicit assumption here is that the physicochemical differences between the amino acids are primarily due to differences in their polarity and volumes, which is a commonly accepted principle.[54]

Columns B and Z were not a part of the original matrix of Miyata *et al.*[18] B represents an ambiguous case, in which the sequencing procedure gave glutamate or glutamine as possible amino acids at a certain position, and Z stands for uncertainties between aspartate and aspargine. The exchange grades of any amino acid with B were calculated as the average between the exchange grade of that amino acid with glutamate and that amino acid with glutamine. Likewise, the exchange grades of any amino acid with Z were calculated as the average between the exchange grade of that amino acid with aspartate and that amino acid with aspargine.

Thus, the physicochemical conservation grade, P_k at position k in the alignment, was calculated as:

$$P_k = \sum_{m=1}^{N} (A_{ij}^{m}(k) M_{ij}) \qquad (2)$$

where A_{ij}^{m} is a matrix of elements 0 and 1 describing amino acid replacements, M_{ij} is the replacement value taken from Table 2, and N is the number of sequences in the alignment. In cases where multiple phylogenetic trees were generated by Protpars, a grade was determined for each tree and the final grade was taken as the average of all trees.

Dealing with gaps

Multiple gap positions were given low conservation scores; the higher the number of gaps was, the lower were the scores. "Exchange" of any amino acid with a gap, which actually means insertion or deletion, reduces the conservation grade of the position by six (Table 2; the – and X symbols). Since it is assumed that gaps appear in low conserved spots, we choose this value as it is above, yet close to the highest amino acid distances (Table 2). If for position k, both father and son sequences contain a gap, ConSurf subtracts 0.5 point from the conservation score for that position. This score was chosen after higher scores biased the average conservation too much.

Average conservation

After grading each position in the alignment, ConSurf chooses the positions in the alignment that are ungapped in the query protein (the protein of known structure) and computes their average conservation ($\langle P_k \rangle$) and standard deviation (σ_k). Thus, gapped positions are allowed to influence the overall grade of each position only if they are ungapped in the query protein. This is because gapped positions in the query protein appear due to the adding of distant relatives to the alignment and we wish to compute the average conservation only for its existing positions, which are under the pressure of protein evolution. ConSurf then normalizes each grade as:

$$P_k = (P_k - \langle P_k \rangle)/_k \qquad (3)$$

and the normalized conservation grades, replacing the *B* (or temperature) factors in the pdb file of the protein, are ready to be mapped onto the protein surface.

Normalization was done for two main reasons. First, it provides a reference state for the level of conservation; a residue detected by ConSurf to be maximally conserved in a protein is as conserved as a residue that is buried in the protein core. We assumed that surface residues that maintain bimolecular linkages in essential cellular function should be as conserved as the internal ones. The second reason is methodological. When dealing with the SH2 domains, we compared the conservation of the protein surface using increased levels of evolutionary distances between sequences around our query (Figures 2 and 3). In general, as we increase the number of sequences we also increase the evolutionary distance of the group. This increase is supposed to change the level of conservation of each position, and thus change the average conservation. The normalization using equation (3) is a means to diminish this effect and to compare the conservation grades obtained for the different branches of the phylogenetic tree on the same grounds.

Conservation mapping by increasing evolutionary distances

To map the conservation according to increasing evolutionary distance for the SH2 domain (Figure 2), we collected sequences defined as clades (or sub-trees) of the tree shown in Figure 1, around the sequence of the human Src protein. After collecting the sequences we aligned them and used the alignment as input for another round of ConSurf to generate conservation grades for each position.

A shorter path

For the PTB domain we chose an already existing alignment that was done using a hidden Markov model and appears on the pfam database.[29] This is because the three PTB domains that we investigated cannot be aligned using CLUSTAL W due to the low level of sequence similarity between them. The alignment was then processed using PROTPARS to build a phylogenetic tree and to derive the conservation grades. It should be noted that the IRS-1 PTB domain, the structure of which is known,[36] is missing in the alignment and was therefore not included in our analysis.

Viewing the conservation grades

The temperature (*B*) factors in the input coordinate file (in pdb format) were replaced with the conservation grades of the residues. Thus, any 3D-protein viewer, such as RASMOL,[55] which is capable of presenting the *B* factors, is suitable for viewing the conservation map of proteins. We used GRASP[56] for mapping the conservation grades on the molecular surface of the SH2 and PTB domains.

Comparison of conservation and contact maps

The contact (or projection) maps were deduced from those changed in the water accessible surface area of the amino acids in the liganded and unliganded domain. We calculated the water accessible surface area with a modified Shrake-Rupley[57] algorithm, implemented in the SURFV computer program.[58] The contact map presents the ratio between the water accessible surface area obtained for a given residue with ligand and without ligand.

Acknowledgments

We are thankful to Burkhard Rost for his comments on the manuscript. This work was supported by the Israel Cancer Association and by fellowships from the Wolfson and Alon Foundations to N.B.-T. D.G. was supported by the Magnet ''Da'at'' consortium of the Israel Ministry of Industry and Trade.

References

1. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). The protein data bank. *Nucl. Acids Res.* **28**, 235-242.
2. Sali, A. (1998). 100,000 protein structures for the biologist. *Nature Struct. Biol.* **5**, 1029-1032.
3. Rost, B. (1998). Marrying structure and genomics. *Structure,* **6**, 259-263.
4. Branden, C. & Tooze, J. (1999). *Introduction to Protein Structure*, 2nd edit., Garland Publishing, Inc., New York.
5. Bogan, A. A. & Thorn, K. S. (1998). Anatomy of hot spots in protein interfaces. *J. Mol. Biol.* **280**, 1-9.
6. Meyer, T. E., Tollin, G & Cusaovich, M. A. (1994). Protein interaction sites obtained *vis* homology. The site of complexation of electron transfer portions of cytochrome *c* revealed by mapping amino acid substitution onto three-dimensional protein surfaces. *Biochimie,* **76**, 480-488.
7. Pazos, F., Helmer-Citterich, M., Ausiello, G. & Valencia, A. (1997). Correlated mutations contain information about protein-protein interaction. *J. Mol. Biol.* **271**, 511-523.
8. Lockless, S. W. & Ranganathan, R. (1999). Evolutionary conserved pathways of energetic connectivity in protein families. *Science,* **286**, 295-299.
9. Kisters-Woike, B., Vangierdegom, C. & Müller-Hill, B. (2000). On the conservation of protein sequences in evolution. *Trends Biochem. Sci.* **25**, 419-421.
10. Gallet, X., Charloteaux, B., Thomas, A. & Brasseur, R. (2000). A fast method to predict protein interaction sites from sequences. *J. Mol. Biol.* **302**, 917-926.
11. Lichtarge, O., Bourne, H. R. & Cohen, F. E. (1996a). An evolutionary trace method defines binding sur-

faces common to protein families. *J. Mol. Biol.* **257**, 342-358.

12. Lichtarge, O., Bourne, H. R. & Cohen, F. E. (1996b). Evolutionarily conserved G-alpha-beta-gamma binding surfaces support a model of the G protein-receptor complex. *Proc. Natl Acad. Sci. USA*, **93**, 7507-7511.
13. Lichtarge, O., Yamamoto, K. R. & Cohen, F. E. (1997). Identification of functional surfaces of the zinc binding domains of intracellular receptors. *J. Mol. Biol.* **274**, 325-337.
14. Feng, D. F. & Doolittle, R. F. (1987). Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J. Mol. Evol.* **25**, 351-360.
15. Higgins, D. G. & Sharp, P. M. (1989). Fast and sensitive multiple sequence alignments on a microcomputer. *Comput. Appl. Biosci.* **5**, 151-153.
16. Graur, D. & Li, W.-H. (2000). *Fundamentals of Molecular Evolution*, 2nd edit., Sinauer Associates, Sunderland, MA.
17. Felsenstein, J. (1996). Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. *Methods Enzymol.* **266**, 418-427.
18. Miyata, T., Miyazawa, S. & Yashunaga, T. (1979). Two types of amino acid substitutions in protein evolution. *J. Mol. Evol.* **12**, 219-236.
19. Xu, W., Harrison, S. C. & Eck, M. J. (1997). Three-dimensional structure of the tyrosine kinase c-Src. *Nature*, **385**, 595-602.
20. Sicheri, F., Moarefi, I. & Kuriyan, J. (1997). Crystal structure of the Src family tyrosine kinase Hck. *Nature*, **385**, 602-609.
21. Waksman, G., Shoelson, S. E., Pant, N., Cowburn, D. & Kuriyan, J. (1993). Binding of a high affinity phosphotyrosyl peptide to the Src SH2 domain: crystal structure of the complexed and peptide-free forms. *Cell*, **72**, 779-790.
22. Songyang, Z., Shoelson, S. E., Chaundhuri, M., Gish, G., Pawson, T. & Haser, W. G. *et al.* (1993). SH2 domains recognize specific phosphopeptide sequences. *Cell*, **72**, 767-778.
23. Kimber, M. S., Nachman, J., Cunningham, A. M., Gish, G. D., Pawson, T. & Pai, E. F. (2000). Structural basis for specificity switching of the Src SH2 domain. *Mol. Cell*, **5**, 1043-1049.
24. Harrison, S. C. (1996). Peptide-surface association: the case of PDZ and PTB domains. *Cell*, **86**, 341-343.
25. Siegal, G. (1999). The surprisingly flexible PTB domain. *Nature Struct. Biol.* **6**, 7-10.
26. Forman-Kay, J. D. & Pawson, T. (1999). Diversity in recognition by PTB domains. *Curr. Opin. Struct. Biol.* **9**, 690-695.
27. Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucl. Acids. Res.* **22**, 4673-4680.
28. Saitou, N. & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406-425.
29. Bateman, A., Birney, E., Durbin, R., Eddy, R. D., Howe, K. L. & Sonnhammer, E. L. L. (2000). The Pfam protein families database. *Nucl. Acids Res.* **28**, 263-266.
30. Borg, J. P., Ooi, J., Levy, E. & Margolis, B. (1996). The phosphotyrosine interaction domains of X11 and FE65 binf to distinct sites on YENPTY motif of amyloid precursor protein. *Mol. Cell. Biol.* **16**, 6229-6241.
31. Zhang, Z., Lee, C., Mandiyan, V., Borg, J. P., Margolis, B., Schlessinger, J. & Kuriyan, J. (1997). Sequence-specific recognition of the internalization motif of the Alzheimer amyloid precursor protein by the X11 PTB domain. *EMBO J.* **16**, 6141-6150.
32. Trub, T., Choi, W. E., Wolf, G., Ottinger, E., Chen, Y., Weiss, M. & Shoelson, S. E. (1995). Specificity of the PTB domain of Shc for beta turn-forming pentapeptide motifs amino-terminal to phosphotyrosine. *J. Biol. Chem.* **270**, 18205-18208.
33. Wolf, G., Trub, T., Ottinger, E., Groninga, L., Lynch, A. & White, M. F., *et al.* (1995). PTB domains of IRS-1 and SHC have distinct but overlapping binding specificities. *J. Biol. Chem.* **270**, 27407-27410.
34. Kavanaugh, W. M. & Williams, L. T. (1994). An alternative to SH2 domains for binding tyrosine-phosphorylated proteins. *Science*, **266**, 1862-1865.
35. Zhou, M., Ravichandran, K. S., Olejniczak, E. T., Petros, A. M., Meadows, R. P. & Sattler, M. *et al.* (1995). Structure and lignad recognition of the phosphoytrosine binding domain of SHC. *Nature*, **378**, 584-592.
36. Eck, M. J., Dhe-Paganon, S., Trüb, T., Nolte, R. T. & Shoelson, S. E. (1996). Structure of the IRS-1 PTB domain bound to the juxtamembrane region of the insulin receptor. *Cell*, **85**, 695-705.
37. Zhou, M.-M., Huang, B., Olejniczak, E. T., Meadows, R. P., Shuker, S. B. & Miyazaki, M., *et al.* (1996). Structural basis for IL-4 receptor phosphopeptide recognition by the IRS-1 PTB domain. *Nature Struct. Biol.* **3**, 388-393.
38. Dho, S. E., Jacob, S., Wolting, C. D., French, M. B., Rohrschneider, L. R. & McGlade, C. J. (1998). The mammalian numb phosphotyrosine-binding domain. Characterization of binding specificity and identification of a novel PDZ domain-containing numb binding protein LNX. *J. Biol. Chem.* **273**, 9179-9187.
39. Howell, B. W., Lanier, L. M., Frank, R., Gertler, F. B. & Cooper, J. A. (1999). The disabled 1 phosphotyrosine-binding domain binds to the internalization signals of transmembrane glycoproteins and to phospholipids. *Mol. Cell. Biol.* **19**, 5179-5188.
40. Li, S., Zwahlen, C., Sebastien, J. F., Vincent, C., ., McGlade, J., Kay, L. E., Pawson, T. & Forman-Kay, J. D. (1998). Structure of a numb PTB domain-peptide complex suggests a basis for diverse binding specifity. *Nature Struct. Biol.* **5**, 1075-1083.
41. Bork, P. & Margolis, B. (1995). A phosphotyrosine interaction domain. *Cell*, **80**, 693-694.
42. Bradshaw, J. M. & Waksman, G. (1999). Calorimetric examination of high-affinity Src SH2 domain-tyrosyl phosphopeptide binding: dissection of the phosphopeptide sequence specificity and coupling energetics. *Biochemistry*, **38**, 5147-5154.
43. Otsuka, J., Fukuchi, S. & Kikuchi, N. A. (1997). Theoretical method for evaluating the relative importance of positive selection and neutral drift from observed base changes. *J. Mol. Evol.* **45**, 178-192.
44. Schneider, R. & Sander, C. (1996). The HSSP database of protein structure-sequence alignments. *Nucl. Acids Res.* **24**, 201-205.
45. Dayhoff, M. O., Schwartz, R. M. & Orcutt, B. C. (1978). A model of evolutionary change in proteins. In *Atlas of Protein Sequence and Structure* (Dayhoff, M. O., ed.), vol. 5, suppl. 3, pp. 345-352, National Biomedical Research Foundation, Washington, DC.

46. Montelione, G. T. & Anderson, S. (1999). Structural genomics: keystone for a human proteome project. *Nature Struct. Biol.* **6**, 11-12.
47. Smith, T. F. & Waterman, M. S. (1981). Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195-197.
48. Bairoch, A. & Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucl. Acids Res.* **28**, 45-48.
49. Shpaer, E. G., Robinson, M., Yee, D., Candlin, J. D., Mims, R. & Hunkapiller, T. (1996). Sensitivity and selectivity in protein similarity searches: a comparison of Smith-Waterman in hardware to BLAST and FASTA. *Genomics,* **38**, 179-191.
50. Pearson, R. W. (1998). Empirical statistical estimates for sequence similarity searches. *J. Mol. Biol.* **276**, 71-84.
51. Durbin, R., Sean, R., Eddy, S. R., Krogh, A. & Mitchison, G. (1998). *Biological Sequence Analysis Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press, Cambridge, England.
52. Fitch, W. M. (1971). Toward defining the course of evolution: minimum change for a specific tree topology. *Syst. Zool.* **20**, 406-416.
53. Grantham, R. (1974). Amino acid difference formula to help explain protein evolution. *Science,* **185**, 862-864.
54. Schulz, G. E. & Schirmer, R. H. (1979). *Introduction to Protein Structure*, Springer-Verlag, New York.
55. Sayle, R. A. & Milner-White, E. J. (1995). RASMOL: biomolecular graphics for all. *Trends. Biochem. Sci.* **20**, 374-376.
56. Nicholls, A., Sharp, K. A. & Honig, B. (1991). Protein folding and association: insights from the interfacial and the dynamic properties of hydrocarbons. *Proteins: Struct. Funct. Genet.* **11**, 281-296.
57. Shrake, A. & Rupley, J. A. (1973). Environment and exposure to solvent of protein atoms. Lysozyme and insulin. *J. Mol. Biol.* **79**, 351-371.
58. Sridharan, S., Nicholls, A. & Honig, B. (1992). A new vertex algorithm to calculate solvent accessible surface area. *Biophys. J.* **61**, A174.
59. Mizuguchi, K., Deane, C. M., Blundell, T. L. & Overington, J. P. (1998). HOMSTRAD: a database of protein structure alignments for homologous families. *Protein Sci.* **7**, 2469-2471.

Edited by F. Cohen

(Received 9 October 2000; received in revised form 2 January 2001; accepted 3 January 2001)