Diversity in the SH2 domain family phosphotyrosyl peptide binding site

S.J.Campbell and R.M.Jackson

School of Biochemistry and Molecular Biology, Garstang Building, University of Leeds, Leeds LS2 9JT, UK

1To whom correspondence should be addressed.
E-mail: jackson@bmb.leeds.ac.uk

Src homology 2 (SH2) domains are ~100 residue phosphotyrosyl peptide binding modules found in signalling proteins and are important targets for therapeutic intervention. The peptide binding site is evolutionarily well conserved, particularly at the two major binding pockets, pTyr and pTyr + 3. We present a computational analysis of diversity within the peptide binding region and discuss molecular recognition beyond the conventional binding motif, drawing attention to novel conserved ligand interaction sites which may be exploitable in ligand binding studies. The peptide binding site is defined by selecting crystal contacts and domains are clustered according to binding site residue similarity. Comparison with a classification based on experimental peptide screening reveals a high level of qualitative agreement, indicating that the method is able independently to generate functional information. A conservation scoring method reveals extensive patches of conservation in some groups not present across the whole family, challenging the notion that the domains recognise only a linear phosphopeptide sequence. Conservation difference maps determine group-dependent clusters of conserved residues that are not seen when considering a larger experimentally determined group. Many of these residues contact the peptide outside the pTyr to pTyr + 3 motif, challenging the conventional view that this motif is largely responsible for ligand recognition and discrimination.

Keywords: drug design/evolutionary trace/molecular recognition/protein interaction/residue conservation

Introduction

SH2 domains were identified through sequence similarities in the N-terminal catalytic regions of Src related protein tyrosine kinases (Sadowski et al., 1986). They are highly structurally conserved motifs of ~100 amino acids, found in a large number of proteins involved in signal transduction. They exhibit a number of functions, including the ability to transmit signals and act as adapters between receptors. They regulate kinase activity, affecting a range of cell responses such as proliferation, apoptosis, growth and regulation of enzyme activity [reviews are available (Sawyer, 1998; Grucza et al., 1999)]. These domains recognise the phosphorylation state of tyrosine residues, specifically binding phosphotyrosine (pTyr) containing protein motifs with high affinity.

A large number of structures have been determined for the SH2 family of domains, including src, hck, fyn, lck, syk, abl, p85, plc and shp-2, revealing that many of the domains exhibit a conserved architecture, containing a five-stranded anti-parallel β-sheet core, sandwiched between two α-helices, additionally extended by a β-strand and small triple-stranded β-sheet (Waksman et al., 1992). The preferred sequence of the pTyr-containing peptide ligand is pTyr–X1–X2–X3, where X1, X2 and X3 refer to the first (pTyr + 1), second (pTyr + 2) and third (pTyr + 3) positions within the peptide sequence. Structural studies show that the phosphorylated ligand binds perpendicularly to the 5-stranded anti-parallel β-sheet core, characteristically interacting with two well-defined pockets, which play a role in ligand recognition. The first of these pockets is the basic pTyr-binding pocket, containing several positively charged residues including a vital arginine (Arg 32) from the SH2 ‘signature’ motif, Phe–Leu–Val–Arg–Glu–Ser (FLVRES). It has been shown in src that mutation of Arg32 prevents virtually all binding of pTyr-containing ligands to the domain (Bibbins et al., 1993). Even so, it has been demonstrated that the phosphate is also involved in hydrogen-bonding interactions with Ser34, Glu35 and Thr36 and makes hydrophobic contacts with the side chain of Lys60. Previously, the specificity of the interaction has largely been attributed to the hydrophobic second binding pocket, which is formed by two loop regions and accommodates the X3 residue. Indeed, mutations to the residues that make up the second cavity have been shown to result in changes to ligand binding specificity and activity (Marengere et al., 1994; Bradshaw and Waksman, 1998), although biochemical studies of the pTyr–Glu–Glu–Ile (pYEEl) phosphopeptide binding to Src-like domains have shown that the X1 and X2 also remain important for high-affinity recognition of the peptide (Gilmer et al., 1994). There is also evidence to suggest that the pTyr + 4 position may contribute to the binding affinity in the N-terminal SH2 domain of shp-1 (Beebe et al., 2000).

Studies of the interactions of SH2 domain-containing proteins provide an insight into the complex mechanisms and functions of signal transduction. However, the understanding of the cellular mechanisms has been somewhat hindered, perhaps by the lack of known inhibitors of SH2 domains which are effective in cell-based assays (Sawyer, 1998). Additionally, mechanistic studies have been curbed by the difficulty in making quantitative measurements of the blockage of signal transduction pathways via such inhibitors and the associated downstream readouts. Despite some recent successes (Shakespeare et al., 2000; Vu, 2000), progress in the field of SH2 domain structure-based drug design has also been slower than first anticipated. This is to some degree because of the lack of bioavailability, but also perhaps as a result of the high degree of similarity between the domains, particularly in the region which binds the conventional pTyr to pTyr + 3 motif, which has provided the chief focus of many drug design strategies. Experiments involving screening of randomized phosphopeptide libraries (Songyang et al., 1993) led to the idea that SH2 interactions maintain specificity by interacting
with particular amino acid sequences contained within the phosphorylated peptide. However, it has also been suggested that while the selected preferential motifs are those with the highest affinity, the difference in affinity between a ‘specific’ and ‘non-specific’ interaction is not necessarily high, amounting to less than two orders of magnitude in affinity (Ladbury and Arold, 2000). It has been questioned whether this is sufficient to guarantee mutual exclusivity in signalling pathways in cells with more than one type of SH2 domain. Indeed, in the same study, an investigation into the surfaces of four SH2 domains in terms of charge and polarity revealed that there is little to distinguish between the binding sites of src, p85, syp and grb (Ladbury and Arold, 2000). Nevertheless, SH2 domains remain a highly desirable drug target, owing to the range of potential applications, for example inhibitors of src with respect to regulating bone resorption, zap-70 inhibitors regarding immune suppression and inhibitors of grb2, a component of the oncogenic Ras pathway.

SH2 domains have been used as model systems in several studies where techniques for prediction of protein–protein interfaces have been developed. Casari et al. (1995) represented entire proteins and sequence residues as vectors in a generalized ‘sequence space’ to predict residues involved in protein function. To assess the validity and accuracy of an evolutionary trace method that defines binding surfaces common to protein families, Lichtarge et al. (1996) identified functional epitopes and residues within the SH2 domain family critical to binding. In testing an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information, Armon et al. (2001) demonstrated general surface conservation of the SH2 domain binding site. It was observed that the ‘typical’ SH2 domain binding site was well conserved and the conservation ‘patch’ decreased in size as the clade size from a phylogenetic tree was increased. However, while these studies predict the general location of the phosphotyrosyl binding site with a good degree of accuracy, they do not directly describe the differences between SH2 domains in terms of specific interactions with peptides.

Here, we address the issue of ligand recognition and discrimination by SH2 domains in the context of crystallographic and peptide screening data. A conservation scoring method (Valdar and Thornton, 2001) is used to analyse the degree of conservation present in and around the SH2 domain phosphotyrosyl binding site. The domains are analysed within the context of entire proteins and sequence residues as vectors in a generalized ‘sequence space’ to predict residues involved in protein function. Calculation and display of residue conservation

**Materials and methods**

**SH2 domain 3D-dataset and sequence alignment**

Coordinates of SH2 domains were retrieved from the Protein Databank (Berman et al., 2000) following sequence searches in Sequences Annotated by Structure (SAS) (Milburn et al., 1998) and text-based searches in PDBSum (Laskowski et al., 1997). Where there was more than one SH2 domain per PDB file, structural alignments were performed on the domain C atoms coordinates using a geometric hashing structural alignment algorithm (R.M.Jackson, unpublished work) to select the structure with the highest level of atomic similarity to the rest of the dataset. In the case where only one of the domains contained a ligand, that domain was chosen. All other domains were removed from each PDB file. PREPI (Islam, 1995) was used to transform the coordinates of the c-src representative structure so that the surface of the peptide binding site was oriented on the ‘y’-plane. The ‘z’-axis represents the depth of the binding site. The selected domains were then structurally aligned to the newly orientated c-src domain, resulting in the transformation of the domains to the same orientation, allowing superimposition of multiple ligands on a representative template. Sequences of structurally unresolved SH2 domains were obtained using text-based and sequence-based searches in SwissProt (Bairoch and Apweiler, 2000) and BLAST (Altschul et al., 1990) and converted to FASTA format where necessary. A complete sequence alignment of all the domains was created using ClustalX (Thompson et al., 1997) and edited using Jalview (Clamp, 1999).

**Definition and clustering of binding site residues**

Binding site residues were defined by determining protein–ligand contacts in the crystal structures. For a given SH2 domain, all PDB files containing ligands were used in the calculation and determination of protein–ligand atomic contacts. Protein residues with at least one atom within a calculated atomic distance of 5 Å or less from any ligand atom were included in the binding site definition. The ligand contacting residues were mapped to the full sequence alignment of SH2 domains. The binding site is defined as any position at which at least 50% of the domains with bound ligands have at least one contact, resulting in the inclusion of 17 positions in the binding site definition. A second definition of the binding site is made, where any position with at least 35% of the SH2 domains with bound ligands had at least one contact. This resulted in 24 alignment positions in the second binding site definition.

Binding site similarity scores were calculated across the whole SH2 domain family in an ‘all-against-all’ comparison, using a binary scoring method. This systematically selects pairs of domains from a given alignment for comparison and scores identical pairs of residues at each alignment position. A total score is then calculated, producing a matrix of scores that represents how similar each SH2 domain-binding site is to the binding sites of the rest of the family. Cluster analysis was performed on these scores with the OC cluster analysis program (Barton, 1993), using the the means linkage method (UPGM). This was carried out using Scorecons (Valdar and Thornton, 2001).
Scorecons calculates residue conservation at each position within a multiple sequence alignment. A value \( \text{Cons}(i) \) between 0 and 1 is assigned to each alignment position, where 0 represents a position that is not conserved and 1 represents a completely conserved position. The Pairwise Exchange Table (PET91) of Jones et al. (Jones et al., 1998) is used to assess the diversity of residues at each alignment position. A weighted sum of pairwise similarities between residues at a given alignment position is then calculated. The function \( \text{Cons}(i) \) for position \( i \) within the alignment is defined as

\[
\text{Cons}(i) = \frac{\sum_{j}^{N} \sum_{k}^{N} W_j W_k \text{Mut}(s_j(i), s_k(i))}{\sum_{j}^{N} \sum_{k}^{N} W_j W_k}
\]

where \( N \) is the number of aligned sequences; \( s_j(i) \) and \( s_k(i) \) are the residues at alignment position \( i \) of sequences \( s_j \) and \( s_k \); \( \text{Mut}(a,b) \) measures the similarity between residues \( a \) and \( b \) according to the mutation data matrix. \( W_j \) is the average evolutionary distance between \( s_j \) and the other aligned sequences. For further details, see Valdar and Thornton (Valdar and Thornton, 2001).

Conservation scores can be displayed via a given representative PDB file for which the sequence is an exact match to one within the multiple alignment. Representative template structures for each group of domains were selected based on the method used to solve the structure, the resolution of the structure, stereochemical quality of the structure using Procheck (Laskowski et al., 1993) and whether ligands were bound. Conservation scores at each position in the alignment were mapped to the corresponding residue in the representative PDB file and the resulting files were viewed using Grasp (Nicholls et al., 1991), colouring the molecular surface by residue conservation. The colour scheme presented here ranges from blue, through white, to red, where blue represents maximal conservation and red zero conservation. To ensure that all surfaces are coloured on the same scale, two ‘dummy’ atoms are added to each PDB file, with values of 1 and 0, representing the maximum and minimum conservation score, respectively.

Analysis of residue conservation

Mean conservation scores were calculated by adding scores for each position within the full alignment or part of an alignment (i.e. binding site) and dividing by the number of positions. The sequence alignment was annotated according to (i) residue conservation, (ii) residue accessibility, and (iii) the presence of residues in the binding ‘face’ of the domain. This involves (i) residues with a conservation value of \( \geq 0.65 \), (ii) relative residues surface accessibility of \( \geq 11\% \) was calculated using Naccess (Hubbard and Thornton, 1993), and (iii) a ‘\( z \)’ coordinate cut-off of 10 Å from the residue with the highest positive ‘\( z \)’ coordinate to denote presence in the binding ‘face’ of the domain. This combination is used to highlight residues that fitted all three criteria. It defines highly conserved residues in the binding face of the domain.

Comparison of parent and group conservation

Differences in residue conservation scores between a group alignment, \( \text{Cons}_G(i) \), and the parent alignment, \( \text{Cons}_P(i) \), are given by subtracting the parent conservation score from the group score at each alignment position. Since conservation scores range from 0 to 1, it follows that the difference between two conservation scores potentially ranges from –1 to 1. These differences in conservation scores between alignments at each residue position are normalized on a scale of 0 to 1, using

\[
\frac{[\text{Cons}_G(i) - \text{Cons}_P(i)] + 1}{2} = \text{Normalised Value}
\]

and displayed on a representative PDB structure as described above. A value from 0 to \(<0.5\) represents a loss in conservation of the group relative to the parent. This is coloured grey through to white. A value from \( >0.5 \) to 1 represents a gain in conservation. This is coloured white through to green. To ensure that all surface difference maps are coloured on the same scale, two ‘dummy’ atoms with values of 0 and 1 were added to each PDB file. Calculated differences in conservation score between a group and parent were mapped back to the sequence alignment. Residues with relative accessibilities \( \geq 11\% \) showing a gain in conservation score of \( \geq 20\% \) relative to the parent group that are also present on the surface of the binding ‘face’ of the domain were highlighted.

Results

SH2 domain–ligand contacts

All SH2 domain–ligand contacts based on the available data are shown in Figure 1, mapped to the full sequence alignment. The general binding site is highlighted and is defined as positions at which at least 50% of the ligand-containing structures have at least one protein–ligand contact. Additional sites are highlighted where at least 35% of the ligand-containing structures have at least one protein–ligand contact. Thus, the SH2 domain binding site can be described by 17 (\( \geq 50\% \) having contacts) or 24 (\( \geq 35\% \) having contacts) positions within the complete sequence alignment. The locations of most of the residues involved in protein–ligand contacts appear to be consistent throughout the dataset. Additionally, the identities of many of the residues themselves are sequence conserved, in particular the G/S-L motif at alignment position 118–119. The so-called SH2 domain ‘signature’ FLVRES sequence is at positions 35–40, although it appears that within the given cut-off distance of 5 Å, only the second half of the motif (RES) is directly involved in the ligand interaction as defined here. Two additional amino acids at positions 41–42, usually E and S/T that follow the RES motif, are also frequently involved in the interaction. The highly conserved arginines at positions 13 and 38 can also be seen. The greatest diversity in terms of the binding sites seems to be between positions 105 and 126 (Figure 1) where there is inconsistency in the number of residues involved in the interaction and the specific location of these contacting residues within the alignment.

Clustering of SH2 domain binding sites

Based on the \( \geq 50\% \) contact binding site definition (see Materials and methods), the family of SH2 domains has been divided into five major groups according to amino acid similarity. Figure 2 demonstrates the clustering of the domains into groups A–E and subgroup A’. A study of mean conservation scores (Table I) indicates that when divided into these groups, whole SH2 domain sequences show a higher degree of conservation (0.72, 0.37, 0.43, 0.42 and 0.40 for groups A’–E) than the mean conservation of the entire dataset (0.33). The conservation becomes much more pronounced for the \( \geq 50\% \) and \( \geq 35\% \) contact definitions of the binding site. The highest level of mean conservation is seen in group A’ and
Fig. 1. Full SH2 domain sequence alignment showing contact residues and defined binding regions. Crystal contacts between an SH2 domain and a ligand with atomic distances of ≤ 3.5 Å are underlined. Regions with a black background are positions at which >50% of the ligand-containing structures have at least one protein–ligand contact. Regions with a grey background represent additional contacts at the 35% contact level. Residues marked in bold are conserved within each group at a threshold of >0.65 with solvent accessibilities ≤ 11% and are present on the binding ‘face’ of the domain. The alignment has been divided into groups according to binding site clustering (see text).
Diversity in the SH2 domain peptide binding site

Fig. 2. Clustering of SH2 domain phosphotyrosyl binding sites based on residue similarity. The clustering procedure produces five major groups, A–E. Miscellaneous sequences are excluded from the shaded boxes. Subgroup A’, a group of highly related binding sites is marked. Classification according to Songyang et al. (Songyang et al., 1993) is shown in brackets alongside each domain (see Discussion).

<table>
<thead>
<tr>
<th>Group</th>
<th>Mean conservation score for whole SH2 domain</th>
<th>Average conservation score for 35% conserved binding residues</th>
<th>Average conservation score for 50% binding residues</th>
</tr>
</thead>
<tbody>
<tr>
<td>All SH2 domains</td>
<td>0.33</td>
<td>0.53</td>
<td>0.56</td>
</tr>
<tr>
<td>Group A’</td>
<td>0.72</td>
<td>0.90</td>
<td>0.91</td>
</tr>
<tr>
<td>Group B</td>
<td>0.37</td>
<td>0.65</td>
<td>0.68</td>
</tr>
<tr>
<td>Group C</td>
<td>0.43</td>
<td>0.67</td>
<td>0.70</td>
</tr>
<tr>
<td>Group D</td>
<td>0.42</td>
<td>0.69</td>
<td>0.75</td>
</tr>
<tr>
<td>Group E</td>
<td>0.40</td>
<td>0.65</td>
<td>0.72</td>
</tr>
<tr>
<td>Group 1a</td>
<td>0.50</td>
<td>0.87</td>
<td>0.88</td>
</tr>
<tr>
<td>Group 1b</td>
<td>0.33</td>
<td>0.54</td>
<td>0.57</td>
</tr>
<tr>
<td>Group 3</td>
<td>0.35</td>
<td>0.53</td>
<td>0.57</td>
</tr>
<tr>
<td>Group 4</td>
<td>0.28</td>
<td>0.45</td>
<td>0.47</td>
</tr>
</tbody>
</table>

the lowest in group B. Importantly, the mean conservation score is much greater in the defined binding sites than complete sequences, confirming that these sites are more conserved than the rest of the domain.

Group A, containing fgr, fyn, blk, yes, v-src, c-src, d-src, lyn, hck, lck, abl and the nck domains, is the largest of these and appears to be the group with the least diverse binding sites. Although the nck domains, abl and d-src have been included in group A, they appear to exhibit greater diversity from the rest of the group. Exclusion of these domains results in a distinct subgroup (A’). Indeed, the alignment (Figure 1) shows the 17-residue binding sites of fgr, fyn and yes to be identical. Group B, contains csk and the N-and C-terminal domains of zap-70 and syk. Within this cluster, the N- and
Fig. 3. Conservation maps produced using Scorecons (Valdar and Thornton, 2001). (A) Conservation of the interface for the whole dataset (37 domains) mapped to a representative structure (c-src, PDB file 1f2f) and rotations about the ‘y’-axis through 90°, 180° and 270°. (B) Conservation of the interface within groups for the five main clusters, A–E, mapped to representative structures as follows: group A, c-src, PDB file 1f2f; group B, syk, PDB file 1a81; group C, shptp2 N-terminal domain, PDB file 2shp; group D, p85a C-terminal domain, PDB file 1qad; group E, plc-g1 C-terminal domain, PDB file 2pld. (C) Conservation of the interface mapped to representative structures for the Songyang et al. (Songyang et al., 1993) classification of SH2 domains, groups 1a, 1b and 3. Representative structures are as follows: group 1a, c-src, PDB file 1f2f; group 1b, syk, PDB file 1a81; group 3, plc-g1 C-terminal domain, PDB file 2pld. Conservation maps are not shown for group 2 (a group with a single domain, vav) and group 4 (miscellaneous). Where appropriate, all ligands from available structures belonging a particular group have been superimposed onto the template structure.

C-termini of syk and zap-70 group together. Group C is the second largest of the clusters and is comprised of the N- and C-terminal SH2 domains of shp1, shp2 and csw. Again, the domain termini are seen to group together rather than clustering according to protein type. This pattern is also seen within group D (the p85a and p85b N-and C-terminal domains) and group E (the phospholipase C-γ1 and γ2 N- and C-terminal domains). Shc, atk, xlp, cbl and sem5 are classed as miscellaneous. However, it can also be seen that shc and atk branch at an early point from the same node. Although sem5 and xlp appear to be related to the group A domains, they have been treated separately owing to the branching distances.

The grouping described above is supported by clustering based on the 35% contact conservation level (not shown). Only minor differences are apparent between the 35 and 50% contact conservation trees. For example, csk, which is perhaps
Diversity in the SH2 domain peptide binding site

Fig. 4. Diversity within the phosphotyrosyl binding site: comparison of groups C, D and E with an experimentally determined parent group (see text). Left column: each image represents the residue conservation of the interface for the whole parent group mapped to a representative structure of each of groups C, D and E. Representative structures for groups C, D and E are the same as described in the legend for Figure 3. Middle column: conservation of the interface for groups C, D and E. Right column: difference map (see text for details). The parent group contains all of the domains from groups C, D and E.

an outlier of group B, branches with abl in group A. Group E (the phospholipase C-γ1 and γ2 N- and C-terminal domains) branches into two pairs where the N-termini domains are clustered within group A. However, despite these discrepancies, the groups described herein can be considered stable, as all other SH2 domains remain within the groups on both trees. There are also similarities between the clustering of 50% contact conserved binding regions and that of full SH2 domain sequences (not shown). For example, groups C and D and most of group A remain intact in the full-sequence tree. This
might be expected since it has been noted in a related study (S.J. Campbell and R.M. Jackson, unpublished work) that sequence identities and root mean square deviations relating to structural comparisons tend to be higher within groups than between groups. However, using full sequences in the clustering process failed to group each of the zap-70 and syk domains with csk (group B). Most notable is group E, where the plc domains are once again grouped as two separate pairs within the tree, suggesting that group E is the least stable of the groups. The nck domains, which might be considered to be outliers of group A were also grouped separately within the tree.

Residue conservation within groups

Mapping surface conservation to the molecular surface shows the degree of similarity between binding sites within the five main groups A–E (Figure 3). Figure 3a shows the conservation patterns when all 37 of the domains from the dataset are included in the scoring. The rotations through 90°, 180° and 270° demonstrate that when the whole family is considered, the main region of conservation is restricted to the area immediately proximal to the pTyr binding pocket, rather than covering a greater proportion of the molecule.

By comparing the conservation of all domains (Figure 3a) with those of groups A–E (Figure 3b), it is apparent that binding sites are more highly conserved within the groups than across the whole SH2 family. If five SH2 domains are picked arbitrarily from each of the five groups to form a new random grouping, a similar conservation pattern to Figure 3a (0°) is observed (results not shown). When considering all 37 SH2 domains, the most highly conserved area is around the pTyr binding pocket. However, the region of conservation is seen to extend further across the surface of the domains when individual groups are considered. The group A’ domains exhibit the largest conserved surface area, as might be expected from the high level of sequence similarity (Table 1). The majority of the face of the molecule containing the binding site is highly conserved, in addition to the residues in immediate contact with the ligand, demonstrating the high degree of similarity between the binding sites. In contrast, group B shows a more defined conservation pattern that corresponds closely to the path taken by the bound peptide fragments across the ‘face’ of the molecule. However, in groups C and E the highly conserved regions spread further away from the ligand-contacting regions across the peptide binding site. In the case of group E, there is a strong degree of similarity between the PLC SH2 domain binding sites, despite the apparent group diversity when clustering the domains according to the whole sequences (see above). Group D shows a slightly different pattern, where the pTyr pocket and the area C-terminal to the peptide binding site are well conserved, but the surface between exhibits a lack of conservation. However, when split into two subgroups containing the p85 C-terminal and the N-terminal SH2 domains (not shown), the conserved area increases, suggesting that group D is comprised of two distinct pairs of binding sites, as can be seen in the branching pattern in Figure 2. It should be noted that as with the whole SH2 domain family, it is only the area on the phosphopeptide binding face of the groups of domains that is highly conserved.

The sequence alignment (Figure 1) has been annotated in terms of evolutionary conservation, solvent accessibility and inclusion in the binding ‘face’ of the domain (see Materials and methods). Thus, when considering the three criteria together, the annotations show the distribution of surface residues on the binding ‘face’ of the molecule within each group that are at least 65% conserved. Positions that display all three characteristics are shown in bold. Many of the protein–ligand contacts (underlined) are found to be within these regions, showing that using surface conservation in conjunction with accessibility data can accurately predict functional residues. It follows that the regions that fit the three criteria correspond highly with the ≥50% and ≥35% binding site definitions, highlighted in black and grey. However, there also appear to be positions that are not involved in the binding site definitions. Those of note (see Figure 1) include alignment position 4 in groups B and D, position 6 in groups A, B and E, 14–15 in groups B and C, 20 in groups C and E, 45–48 in groups A, B and C, 60 in groups A and E, 122–123 in group D and alignment position 124 in group C. These might be potential candidate sites for investigation of binding by site-directed mutagenesis.

Locating diversity within the binding site

Binding site diversity between groups and a larger ‘parent’ group (e.g. all SH2 domains) can be investigated using difference maps of surface conservation. Difference scores between a group and its parent are given by subtracting the parent conservation score from the group conservation score at each alignment position and normalizing the resulting difference. Here we investigate differences between groups C, D and E by comparison of each with a parent group consisting of all three (Figure 4). These groups were selected to form a parent group because all except two of the sequences (shptp 2 and corkscrew C-terminal SH2 domains) are described as related in a study by Songyang et al. (Songyang et al., 1993), where the SH2 domain family was classified according to in vitro phosphopeptide recognition (see Discussion). The results of the study (Figure 4) reveal the surface location of regions which are more conserved in the groups than in the parent, where green represents a gain in conservation of a group alignment position relative to the parent (i.e. a positive difference in conservation score), grey represents a loss in conservation (i.e. a negative difference in conservation score) and white represents no change. It is apparent that within each of the groups C, D and E, the size of the conserved region is larger than in the parent group, showing as expected that the groups are more conserved that their parent background. The difference maps reveal the surface locations of residues that are more (or less) conserved within the groups than the parent group and provides an indication of the extent to which there is a difference in conservation. The phosphotyrosine binding pocket is coloured white in all three difference maps, indicating that there is little change between group and parent in what is already a well conserved binding pocket. However, in all three groups, areas immediately adjacent to this site are more conserved than the parent background. An increase in conservation is also clearly present in residue clusters that surround the main area of intensive ligand contact. This is particularly evident in group E. At these locations each group contains residues which are more highly conserved than in the combined group, suggesting that regions surrounding the area of phosphopeptide contact may be important for functional discrimination in the different groups. This pattern challenges the notion that SH2 domains recognise only a linear phosphopeptide sequence, as it is evident that the conserved regions are more extensive than those regions involved in binding the pTyr–X₁–X₂–X₃ motif, which is generally considered most important for high affinity recognition.
Diversity in the SH2 domain peptide binding site

Fig. 5. Difference in conservation between groups and experimentally determined parent group (see text) mapped to sequence alignment. Residues in uppercase are surface residues with solvent accessibilities 11%, those shown in italics make up the binding ‘face’ of the domain and those in bold type show an increase in conservation score 20% relative to the parent alignment. Residues highlighted in grey fit all three criteria. Protein–ligand contacts are underlined and represented by the following symbols: * = contact is between protein and peptide pTyr, pTyr + 1, + 2 or + 3 position; – = contact is N-terminal to peptide pTyr position; + = contact is C-terminal to peptide pTyr + 3 position.

Table II. Key residues in each group that are more conserved than the experimentally determined parent group, containing all of the domains from groups C, D and E

<table>
<thead>
<tr>
<th>Group</th>
<th>Residuea</th>
<th>Increase in Cons. score (%)</th>
<th>Ligand contact?b</th>
<th>Residuea</th>
<th>Increase in Cons. score (%)</th>
<th>Ligand contact?b</th>
<th>Residuea</th>
<th>Increase in Cons. score (%)</th>
<th>Ligand contact?b</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>S118</td>
<td>24</td>
<td>y</td>
<td>G14</td>
<td>46</td>
<td>y</td>
<td>R18</td>
<td>36</td>
<td>y</td>
</tr>
<tr>
<td></td>
<td>G119</td>
<td>36</td>
<td>y</td>
<td>R18</td>
<td>36</td>
<td>y</td>
<td>P29</td>
<td>30</td>
<td>y</td>
</tr>
<tr>
<td></td>
<td>L125</td>
<td>38</td>
<td>y</td>
<td>R26</td>
<td>28</td>
<td>y</td>
<td>R30</td>
<td>20</td>
<td>y</td>
</tr>
<tr>
<td></td>
<td>S142</td>
<td>32</td>
<td>y</td>
<td>E37</td>
<td>22</td>
<td>y</td>
<td>N40</td>
<td>30</td>
<td>y</td>
</tr>
<tr>
<td></td>
<td>H143</td>
<td>36</td>
<td>y</td>
<td>G42</td>
<td>28</td>
<td>y</td>
<td>G53</td>
<td>50</td>
<td>y</td>
</tr>
<tr>
<td></td>
<td>P144</td>
<td>36</td>
<td>y</td>
<td>D51</td>
<td>36</td>
<td>y</td>
<td>K54</td>
<td>48</td>
<td>y</td>
</tr>
<tr>
<td></td>
<td>G145</td>
<td>38</td>
<td>y</td>
<td>E70</td>
<td>28</td>
<td>y</td>
<td>R59</td>
<td>50</td>
<td>y</td>
</tr>
<tr>
<td></td>
<td>G163</td>
<td>20</td>
<td>y</td>
<td>V167</td>
<td>34</td>
<td>y</td>
<td>L69</td>
<td>38</td>
<td>y</td>
</tr>
<tr>
<td></td>
<td>T168</td>
<td>46</td>
<td>y</td>
<td>V91</td>
<td>30</td>
<td>y</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>H169</td>
<td>20</td>
<td>y</td>
<td>Q92</td>
<td>54</td>
<td>y</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>V170</td>
<td>24</td>
<td>y</td>
<td>H93</td>
<td>46</td>
<td>y</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>G184</td>
<td>28</td>
<td>y</td>
<td>N94</td>
<td>74</td>
<td>y</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>E204</td>
<td>40</td>
<td>y</td>
<td>S96</td>
<td>34</td>
<td>y</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>G207</td>
<td>72</td>
<td>y</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>T208</td>
<td>20</td>
<td>y</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>V209</td>
<td>20</td>
<td>y</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Residues numbers refer to PDB structures as follows: group C, shptp2 N-terminal domain, PDB file 2shp; group D, p85a C-terminal domain, PDB file 1qad; group E, plc y C-terminal domain, PDB file 2pdl.

The exact location of such groups of residues can be seen in Figure 5, which shows the distribution of surface residues on the binding ‘face’ of the molecule within each group that show an increase in conservation relative to the parent background. Residues shown in uppercase are solvent-accessible surface residues, those in italics make up the binding ‘face’ of the domain and those in bold type show an increase in conservation score 20% relative to the parent alignment. Residues highlighted in grey fit all three criteria. Protein–ligand contacts are underlined and represented by the following symbols: * = contact is between protein and peptide pTyr, pTyr + 1, + 2 or + 3 position; – = contact is N-terminal to peptide pTyr position; + = contact is C-terminal to peptide pTyr + 3 position.

Table II. Key residues in each group that are more conserved than the experimentally determined parent group, containing all of the domains from groups C, D and E

<table>
<thead>
<tr>
<th>Group</th>
<th>Residuea</th>
<th>Increase in Cons. score (%)</th>
<th>Ligand contact?b</th>
<th>Residuea</th>
<th>Increase in Cons. score (%)</th>
<th>Ligand contact?b</th>
<th>Residuea</th>
<th>Increase in Cons. score (%)</th>
<th>Ligand contact?b</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>S118</td>
<td>24</td>
<td>y</td>
<td>G14</td>
<td>46</td>
<td>y</td>
<td>R18</td>
<td>36</td>
<td>y</td>
</tr>
<tr>
<td></td>
<td>G119</td>
<td>36</td>
<td>y</td>
<td>R18</td>
<td>36</td>
<td>y</td>
<td>P29</td>
<td>30</td>
<td>y</td>
</tr>
<tr>
<td></td>
<td>L125</td>
<td>38</td>
<td>y</td>
<td>R26</td>
<td>28</td>
<td>y</td>
<td>R30</td>
<td>20</td>
<td>y</td>
</tr>
<tr>
<td></td>
<td>S142</td>
<td>32</td>
<td>y</td>
<td>E37</td>
<td>22</td>
<td>y</td>
<td>N40</td>
<td>30</td>
<td>y</td>
</tr>
<tr>
<td></td>
<td>H143</td>
<td>36</td>
<td>y</td>
<td>G42</td>
<td>28</td>
<td>y</td>
<td>G53</td>
<td>50</td>
<td>y</td>
</tr>
<tr>
<td></td>
<td>P144</td>
<td>36</td>
<td>y</td>
<td>D51</td>
<td>36</td>
<td>y</td>
<td>K54</td>
<td>48</td>
<td>y</td>
</tr>
<tr>
<td></td>
<td>G145</td>
<td>38</td>
<td>y</td>
<td>E70</td>
<td>28</td>
<td>y</td>
<td>R59</td>
<td>50</td>
<td>y</td>
</tr>
<tr>
<td></td>
<td>G163</td>
<td>20</td>
<td>y</td>
<td>V167</td>
<td>34</td>
<td>y</td>
<td>L69</td>
<td>38</td>
<td>y</td>
</tr>
<tr>
<td></td>
<td>T168</td>
<td>46</td>
<td>y</td>
<td>V91</td>
<td>30</td>
<td>y</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>H169</td>
<td>20</td>
<td>y</td>
<td>Q92</td>
<td>54</td>
<td>y</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>V170</td>
<td>24</td>
<td>y</td>
<td>H93</td>
<td>46</td>
<td>y</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>G184</td>
<td>28</td>
<td>y</td>
<td>N94</td>
<td>74</td>
<td>y</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>E204</td>
<td>40</td>
<td>y</td>
<td>S96</td>
<td>34</td>
<td>y</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>G207</td>
<td>72</td>
<td>y</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>T208</td>
<td>20</td>
<td>y</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>V209</td>
<td>20</td>
<td>y</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Contacts calculated for ligands as follows: group C, Pdgrf-1009; group D, Tyr751 phosphopeptide from the pdgf receptor; group E, P03 and phosphopeptide from pdgf receptor 1018–1029 (D–N–D–pTyr–I–I).

The exact location of such groups of residues can be seen in Figure 5, which shows the distribution of surface residues on the binding ‘face’ of the molecule within each group that show an increase in conservation relative to the parent background. Residues shown in uppercase are solvent-accessible surface residues, those in italics make up the binding ‘face’ of the domain and those in bold type show an increase in conservation score >20% relative to the parent alignment. Positions that display all three characteristics are highlighted in grey. Group C has the highest number of residues (18) which show an increase in conservation relative to the background, group D has 15 residues and group E 10 residues. These key residues are summarized in Table II, where they are recorded by residue number from a representative PDB file. The corresponding percentage increases in conservation score are also shown. Of these, residues which have been shown to contact ligands in the crystal structures have been indicated, showing that a significant number of protein–ligand contacts occur in regions of the binding site that are conserved in the clustered groups but not in the larger experimentally determined ‘parent’ group.

Many of these conserved regions are involved in contacts at the peptide pTyr to pTyr + 3 positions (see Figure 5), suggesting the presence of some group-dependent residue clusters located at these known binding sites. However, Figure 5 also shows a significant number of contacts outside these
sites that are distinct from the pTyr to pTyr + 3 motif, normally considered the key determinant for recognition. Many of these contacts correspond to group-dependent clusters. In particular this is seen in group D at positions 120–125, where the contact between the p85α C-terminal SH2 domain and the peptide is predominantly C-terminal to the pTyr + 3 position. Group C shows a similar pattern at positions 89–90 and 121–127, where there is group-dependent conservation in a region which is in contact with the peptide C-terminal to the normal binding motif. In the context of the full alignment (Figure 1), these areas (group C 121–127; group D 120–125) incidentally correspond to a period of insertion relative to groups A and B. Additionally, in several cases in groups D (positions 88 and 112) and E (positions 88 and 121), conservation is seen within the pairs of C-terminal and N-terminal SH2 domains (see Figure 5). Again, these sites involve contacts with the peptide C-terminal to the pTyr + 3 binding position. Such group-dependent clusters are different from many of the pTyr to pTyr + 3 binding positions (13, 38, 39, 40, 48, 49, 50, 52, 70, 72, 73, 119) which are generally well conserved across two or more of the groups in Figure 5.

Discussion

Relationship between binding site clustering and full sequence phylogeny

Since the intention was to group the family according to similarity of phosphopeptide binding sites, we used crystallographic structures and contact data to study the SH2 domain binding site. The observed differences in grouping full sequences in exactly the same way as binding sites merits the study of the domains in terms of binding site residues only. However, this asks the question of the extent to which the two differ. The Pfam (Bateman et al., 2002) phylogenetic tree of SH2 domains shows that most of the group B, C and D SH2 domains cluster together. Many of the group A domains cluster on the Pfam tree, but this is not unexpected since we have already shown that many of the group A sequences are highly similar. The observed similarities between the clustering using binding sites only and either the full sequence clustering or Pfam database shows that to a first approximation whole sequence is sufficient to give similar groups in the case of the SH2 domains. Clearly, it remains to be seen if this is a general phenomenon.

Comparison with experimental screening

Songyang et al. classified the SH2 domain family, grouping the domains according to in vitro phosphopeptide recognition (Songyang et al., 1993). A series of experiments were performed in which phosphopeptide libraries consisting of randomized pTyr-containing peptides of the general sequence Gly–Asp–Gly–pTyr–X1–X2–X3–Ser–Pro–Leu–Leu–Leu were used to determine optimal binding sequences for specific SH2 domain binding sites. The X1, X2 and X3 positions were randomized and 22 recombinant SH2 domains were used to screen the library. It was found that the binding of the peptides is dependent on pTyr recognition and that different SH2 domains preferentially bind to different sequence motifs at X1, X2 and X3. SH2 domains were divided into four groups depending on the amino acid at the fifth residue in β-strand ‘D’ (alignment position 72 in Figure 1), which were shown experimentally to select phosphopeptides with similar sequences. The first group (group 1) preferentially bind a pTyr-hydrophilic-hydrophilic-Ile/Pro sequence and is further split into two subgroups, 1a (src, fyn, lck, fgr, lyn, yes, hck and d-src) and 1b (syk N- and C-termini, zap-70 N- and C-termini, atk, abl, csk, nck, sem5/grb2). Members of group 1a that were investigated selected phosphopeptides with the general motif pTyr-hydrophilic-hydrophilic-Ile/Pro. From this group, src, fyn, lck and fgr (from the src family) selected pTyr-Glu-Glu-Ile as the optimal peptide. Nck and abl and sem5/grb2 from group 1b all have an aromatic residue at bD5 but the other residues predicted to contact the ligand side chains are distinct from those in the src family. Crk, nck and abl selected Pro at the X3 binding position. Vav was listed alone in group 2. Group 3 SH2 domains (p85α and p85β N- and C-termini, plc-γ1 and plc-γ2 N- and C-termini, csk, cswt N-terminal, shptp1 N- and C-termini and shptp2 N-terminal SH2 domains) were shown to be selective for a pTyr–hydrophilic–X–hydrophilic sequence. Finally, group 4 (shptp2 C-terminal, csk, cswt C-terminal and shc) contains SH2 domains that exhibit distinct amino acids at the βD5 position.

The groupings of Songyang et al. can be directly compared with those presented here and have been included in brackets alongside each protein in Figure 2. We have also included for comparison surface conservation maps (Figure 3c) using the technique described previously. It is apparent that group A corresponds closely to group 1a of Songyang et al., with all group 1a domains included in the group A cluster. However, group A’ shows a higher degree of conservation as expected due to the more select nature of the subgroup. Group B is similar to group 1b of Songyang et al., both groups including the zap-70 and syk C- and N-termini and csk. The surface diagrams also display a similar region of conservation. However, three group 1b proteins (nck1, nck2 and abl) have been transposed to group A on the basis of binding residue similarity. Atk and sem5 have here been re-classified as miscellaneous. The most significant difference between the classification schemes is between groups C, D and E and group 3 of Songyang et al. Together, groups C, D and E correspond to group 3. However, from the results of the clustering, it appears that each of these is sufficiently distinct to warrant separation into different groups. This is evident in the conservation patterns, where the conserved region is seen to increase significantly in each of the groups relative to their parent group. The differences between group 3 and groups C, D and E are shown in Figure 4 and are discussed above. Cbl and shc remain in the miscellaneous class, while the shptp2 and csw C-termini have been transposed from miscellaneous group 4 to group C. These findings are confirmed by the latter half of the mean conservation scores study (see Table I). The groups which are equivalent to Songyang et al.’s scheme (i.e. A’ to group 1a, B to group 1b, C, D and E to group 3), all show higher mean levels of conservation, indicating that these groups are more robust. However, this level of qualitative agreement between the theoretical and experimental methods indicates that the method described here is able to generate important functional information about SH2 domain binding sites independently of experimental screening methods. This could be applied to other homologous families.

Diversity and specificity in the SH2 domain family

Classification of a protein family into groups provides a framework to study its members. It serves as a starting point for examining similarities and diversities within a large family. Conservation studies can reveal functionally important regions within protein structures, which relate to the interactions of
proteins within a wider assembly. Until a large family is divided into groups it is difficult to draw useful conclusions about the similarity or diversity between family members. The investigation of SH2 domains within the classification scheme described here demonstrates that there are conserved regions within groups that are not present throughout the family as a whole. This diversity, located in the known binding interface, may be important in the recognition of ligands.

It should be appreciated that in the present study, SH2 domain–ligand interactions have been investigated in isolation rather than in entire assemblies or protein complexes. It should also be noted that the results have been obtained using crystal structures of SH2 domains bound to peptides and other small ligands rather than whole proteins as occurs in vivo. Structures containing ligands were unavailable for the entire dataset and those that were available may not be representative of the rest of the SH2 domain family. Thus the clustering method described here is based upon SH2 domains with available ligand-bound structures and the residues where ligand atoms have bound. It is likely that some of the binding sites will have been defined more completely than others.

Nevertheless, several conclusions can be drawn from the study. The changing patterns of conservation between groups and lack of conservation throughout the whole binding region suggest that while the phosphoryltyrosine binding site is generally conserved across the family, regions proximal to this site can be considered diverse between groups. Indeed, the work of Songyang et al. probes only the three residue positions pTyr + 1 to pTyr + 3. Our study suggests that there is binding site residue conservation within groups of similar domains outside these areas that might correspond to the conservation of a protein–protein interface between the SH2 domain and a phosphorylated protein. This observation challenges the notion that SH2 domains recognize only a short linear phosphorylsyl peptide motif in vivo.

Acknowledgements

We thank William Valdar for advice and the use of Scorecons and related software. S.J.C. is funded by a BBSRC special studentship.

References


Hubbard,S.J. and Thornton,J.M. (1993) NACCESS, Department of Biochemistry and Molecular Biology, University College London.


Received September 30, 2002; revised December 16, 2002; accepted January 24, 2003