J. Mol. Biol. (1996) **257**, 342–358

JMB

An Evolutionary Trace Method Defines Binding Surfaces Common to Protein Families

Olivier Lichtarge[1,2], Henry R. Bourne[1] and Fred E. Cohen[1,2]*

[1]*Departments of Cellular and Molecular Pharmacology and Medicine and* [2]*Department of Pharmaceutical Chemistry University of California San Francisco CA 94143-0450, USA*

**Corresponding author*

X-ray or NMR structures of proteins are often derived without their ligands, and even when the structure of a full complex is available, the area of contact that is functionally and energetically significant may be a specialized subset of the geometric interface deduced from the spatial proximity between ligands. Thus, even after a structure is solved, it remains a major theoretical and experimental goal to localize protein functional interfaces and understand the role of their constituent residues. The evolutionary trace method is a systematic, transparent and novel predictive technique that identifies active sites and functional interfaces in proteins with known structure. It is based on the extraction of functionally important residues from sequence conservation patterns in homologous proteins, and on their mapping onto the protein surface to generate clusters identifying functional interfaces. The SH2 and SH3 modular signaling domains and the DNA binding domain of the nuclear hormone receptors provide tests for the accuracy and validity of our method. In each case, the evolutionary trace delineates the functional epitope and identifies residues critical to binding specificity. Based on mutational evolutionary analysis and on the structural homology of protein families, this simple and versatile approach should help focus site-directed mutagenesis studies of structure-function relationships in macromolecules, as well as studies of specificity in molecular recognition. More generally, it provides an evolutionary perspective for judging the functional or structural role of each residue in a protein structure.

Keywords: SH2 domains; SH3 domains; functional motifs; protein evolution; protein-protein interaction

Introduction

In order to understand how proteins recognize ligands and form multimeric complexes, and to identify important functional interfaces to serve as targets for pharmaceutical design, we need to locate these interfaces and evaluate the contribution of individual residues to the overall free energy of binding. Unfortunately, some macromolecular complexes do not easily yield X-ray quality co-crystals, and these complexes are frequently beyond the current limits of multidimensional NMR spectroscopy. Thus in proteins, structural knowledge is often limited to lone domains, which by themselves lack explicit binding site information. Even when the structure of a complex is available, it has proved difficult to deduce the relative contribution of each individual residue to the total binding energy. In fact the binding site, or functional epitope, is a distinct subset of the contact site, or structural epitope (Schreiber & Fersht, 1995; Wells, 1994). Exhaustive mutational analysis therefore remains a mainstay of binding site characterization. Often, a wealth of data is already available in databases, where sequences homologous to the protein of interest record mutation ''experiments'' that have passed the test of natural selection. Our aim is to extract for a protein of interest, the proband, the mutational data imbedded in those sequences and infer which residues are likely to be important to its function.

This task is guided by two observations. First, protein structures descending from a common ancestor are remarkably similar, with backbone deviations remaining within 2 Å even when sequence identity falls to 25% (Chothia & Lesk, 1986). Second, because active site residues are under

Abbreviations used: ET, evolutionary trace; MSA, multiple sequence alignment; PIC, position identity cutoff; DBD, DNA binding domain; ZnF, zinc finger.

0022–2836/96/120342–17 $18.00/0

evolutionary pressure to maintain their functional integrity, they undergo fewer mutations than less functionally important amino acids (Zvelebil *et al*., 1987). These observations imply that evolutionarily related sequences can be compared with one another to extract structural and functional data (Baldwin, 1993; Benner, 1989; Livingstone & Barton, 1993; Sternberg & Cohen, 1982; Zvelebil & Sternberg, 1988). Indeed, useful strategies for protein secondary and tertiary structure prediction based on sequence alignment have been proposed (Barton, 1990; Benner *et al*., 1994; Crawford *et al*., 1987; Sander & Schneider, 1991).

For the distinct purpose of extracting functionally important residues, we extend these observations with two hypotheses. First, common functions retained or evolved from an ancestral protein will take place in descendant proteins at, or near the same structural site. Such a site is well defined because the protein backbone variation is small within the evolutionary family. Furthermore, if no functional divergence occurs within subgroups in the family, the mutation rate at that site will be very low across that subgroup. Second, when a functional difference is observed between evolutionarily related proteins, we assume it arises mostly from mutations at, or near, residues performing that function (we chose to ignore possible indirect effects). Once they have occurred, such mutations now define new functional branches in a protein family, and are themselves under selective pressure not to mutate, lest their critical role be compromised. In summary, a protein family should not only retain its fold but also: (i) conserve the location of functional sites; and (2) have a distinctly lower mutation rate at these sites, punctuated by mutation events that cause divergence.

The novelty of the evolutionary trace (ET) method consists in forming a direct connection between mutations that alter function and patterns of residue conservation in aligned sequences. In turn this allows, first, identification of the position of functionally important residues. Second, the functional resolution (defined below) of the evolutionary trace can be adjusted to maximize either specificity or sensitivity, and thereby shift the focus of the trace from residues that are functionally essential, to residues that regulate specific functional features. Third, because the proband structure is known, we can further distinguish functionally important residues that are internal, and presumably contribute to the structural integrity of the protein, from those that are external, and more likely to be directly linked to binding sites or enzymatic activity. Our approach has been automated but remains transparent: the computer tracks amino acid types at each position in every sequence of a multiple sequence alignment, MSA. Each step is conceptually straightforward and can be readily checked from the original MSA. These points offer a contrast to a vectorial method proposed by Casari *et al*. (1995) for defining functionally important residues in proteins without knowledge of the structure. These authors formulated similar assumptions, but then rooted their approach in a multivariate representation of protein sequences that uses vector analysis to correlate sequence profiles, eigenvector directions and specific residue differences between sequences.

Results

SH2-SH3

SH2 and SH3 domains are small modular elements of proteins typically involved in intracellular signal transduction proteins (Cicchetti *et al*., 1992; Mayer *et al*., 1988; Sadowski *et al*., 1986). They play critical roles in recruitment and assembly of signaling complexes through their specific recognition and binding to target proteins at sites characterized by peptide segments with phosphorylated tyrosine residues, SH2, or a high density of proline residues, SH3 (Cantley *et al*., 1991; Koch *et al*., 1991; Ren *et al*., 1993; for a review, see Pawson, 1994). For both SH2 and SH3 domains, several structures of their complexes with high-affinity peptide ligands are now available (Bibbins *et al*., 1993; Eck *et al*., 1994, 1993; Guruprasad *et al*., 1995; Lee *et al*., 1994; Musacchio *et al*., 1992; Waksman *et al*., 1992, 1993; Wu *et al*., 1995). The distinct binding affinity of each SH2 or SH3 domain toward a specific ligand target sequence (Songyang *et al*., 1993) is essential to limit crosstalk between distinct signaling pathways, but how each residue on the structural epitope influences specificity remains unclear (Birge & Hanafusa, 1993). The SRC_SH2 structure bound to the pYEEI high-affinity ligand solved at 2.7 Å (Waksman *et al*., 1993) was chosen as our first proband. The SH3_ABL protein-peptide complex structure is known at high resolution (Musacchio *et al*., 1992) and served as a second proband.

Dendrogram and partition

In the absence of functional information for every sequence of a large evolutionary family, clustering proteins by sequence identity will produce reasonable groups containing proteins with similar functions. In the case of the SH2 domains, 85 sequences of approximately 100 residues each were identified, and aligned to generate a sequence identity dendrogram, using standard techniques (see Methods). The dendrogram branches are not uniformly populated (see Figure 1). The availability of sequences among divergent SH2 domains ranges from 13, in the ksrc group of sequences (identity greater than 90%), to many single representatives of distinct evolutionary branches such as shc, gagc and kcsk. The grouping of each SH2 domain parallels the evolution of the protein to which it belongs. Thus, we find distinct SH2 branches corresponding to: non receptor protein tyrosine kinases (e.g. ksrc, kfgr, kfyn, kyes, klck and, more distantly, kabl); to

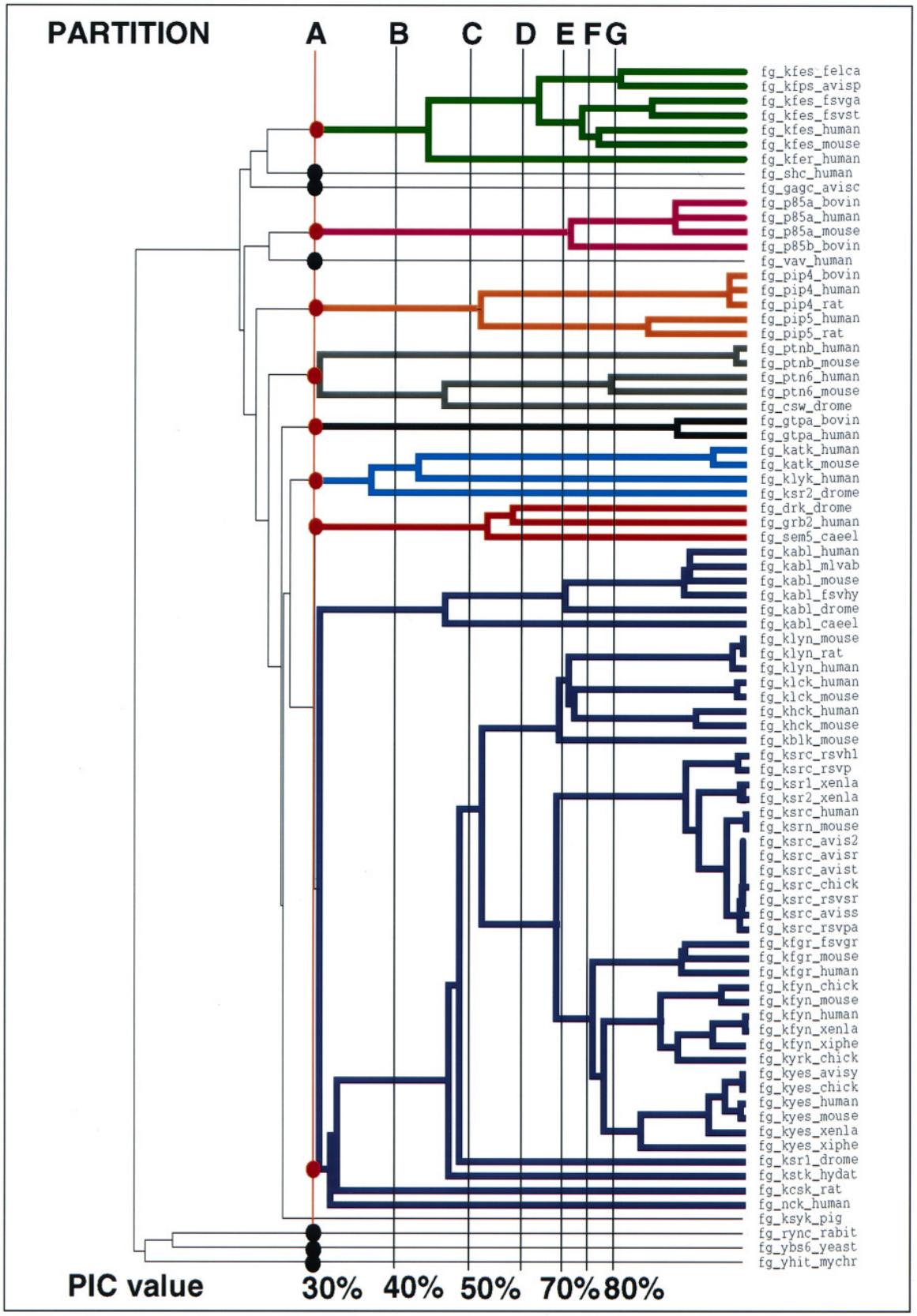

Figure 1. Sequence identity dendrogram of SH2 domains obtained as described in Methods. Vertical lines A to G represent the different partition identity cutoffs (PICs) that are used to define partitions. For each PIC, a partition of the entire family is generated by grouping together sequences that branch off from a node, shown in red, to the right of a PIC line. As PIC increases, from A to G, partitions comprise more groups, each with fewer sequences. Thick colored lines are used in this example to display the different groups in partition A. Six nodes, in black, give rise to singular groups and effectively drop out of the evolutionary trace analysis. ET analysis was performed over this entire tree, see Figure 3, and over the subtree shown in purple, see Figure 4.

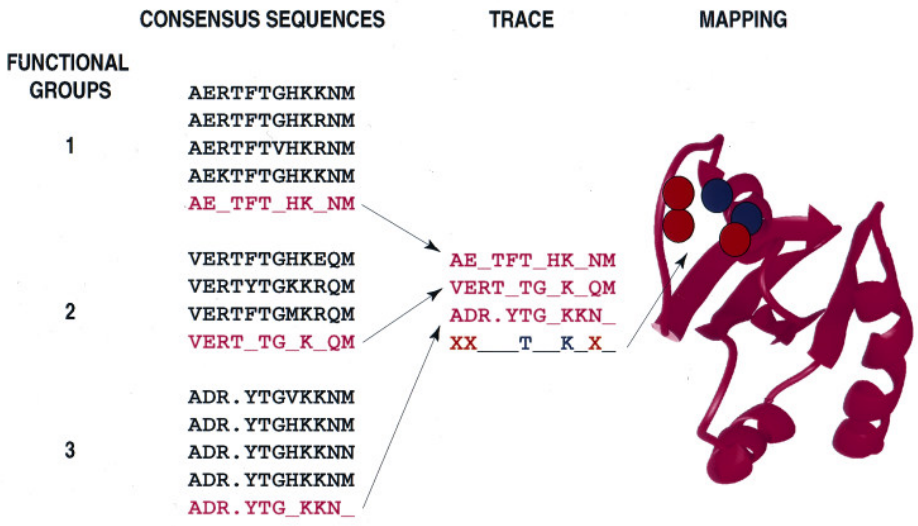

Figure 2. Derivation of the evolutionary trace. The left panel shows sequences from a protein family that have been partitioned into three groups. A consensus sequence is constructed, in magenta, for each group by labeling positions that are invariant in the multiple sequences alignment by its amino acid, and leaving variable positions blank. These invariant residues are next compared between consensus sequences, middle panel. There the evolutionary trace records whether a position in the multiple sequence alignment is conserved (contains an invariant residue conserved in the entire family, blue); or is class-specific (contains invariant residues that change between groups, X in red); or is neutral (fails to be invariant in at least one group). Conserved and class-specific positions are then mapped onto the structure, right panel, to generate a three-dimensional map.

phospholipid metabolism regulation (p85a, pip); and to guanine nucleotide regulatory protein signaling (grk, sem5 and grb2; Pawson, 1994).

To generate an evolutionary trace, a functional clustering is produced by exploiting the correlation between sequence identity groupings and functional characteristics. As can be seen in Figure 1, new mutations emerge constantly during evolution and give rise to new groups that branch off from nodes in the tree. For our purpose, we define a cluster as the set of all the sequences that originate from a common node. An early node (far to the left in Figure 1) defines a large cluster, and a later node (further right) defines a smaller cluster. By considering clusters defined by the first node to the right of a vertical line, the entire dendrogram can be partitioned into clusters (see Figure 1). The minimum percentage identity within a cluster is bounded by the partition identity cutoff (PIC; Du & Alkorta, 1994).

We further define the functional resolution as the number of different clusters generated by a partition of a given family. In Figure 1 the systematic variation of the vertical PIC boundary from right to left, generates distinct partitions. At low PIC, partitions consist of a few large clusters that represent entire classes of proteins. Sequences grouped within the same cluster may vary substantially and the functional resolution is low. Partitions obtained at higher PICs achieve higher functional resolution as large clusters become fragmented into smaller ones. Proteins that remain in the same cluster are increasingly alike in sequence and presumably in function as well. Concurrently, differences between the many emerging smaller clusters narrow, so that finer functional nuances are segregated into distinct clusters at higher PIC.

When sufficient functional knowledge exists for a large number of related sequences, alternative partitions can be chosen without reference to a sequence identity dendrogram. Such partitions can group together sequences based on any number of functional characteristics and need not follow the groupings we use here, which are closely related to patterns of divergent evolution.

Evolutionary trace

For each partition, an evolutionary trace can be constructed. First a consensus sequence is assembled for each group, as illustrated in Figure 2, right panel. A consensus sequence position can be neutral and left blank, if residues for that position vary within the group. Otherwise, it is invariant and assigned its conserved residue type. A group with few sequences will have fewer observed mutations and so its consensus sequence will have relatively few neutral positions.

All consensus sequences are then aligned to obtain the evolutionary trace for the entire partition. In the trace, a position is neutral (identified by an underscore character) if it is neutral in any of the consensus sequences. Otherwise, it is either conserved, if all consensus sequences have the same invariant residue at that position, or class-specific, if its type varies between consensus sequences. The middle panel of Figure 2 summarizes the steps in the construction of an evolutionary trace. Gaps are counted as an extra residue type and confer

neutrality on the trace at positions where they occur in the MSA. Finally, the spatial relationship of functionally important residues defined by ET can be assessed by color coding (mapping) the protein structure to reflect the character of each position: conserved, class-specific or neutral.

Because the tree from which partitions are generated will vary depending on the magnitude of the sequence database, we computed the evolutionary trace for SH2 domains over the full dendrogram, and over a smaller subset restricted to the ksrc family (kabl, klyn, klck, kfyn and kyes). In each case the trace for increasingly fine partitions was mapped onto the SH2-SRC structure in Figures 3 and 4.

Structural cluster identification

For both trees, the evolutionary traces mapped onto the protein structure identify a single, dominant spatial cluster of conserved and class-specific residues. Figure 3 offers a structural context for the SH2 domain. Successive rows represent higher PIC values and hence increasing functional resolution. Cluster 1, on face A consists of buried residues (less than 30% of their side-chain is solvent-accessible), surrounded by a set of solvent exposed residues. Residues are color-coded by ET type, conserved (dark blue) or class-specific (red) internal residues; and conserved (cyan) and class-specific (yellow) external residues. Cluster 1 is visible at the coarsest resolution, partition A1, and steadily expands in a nearly concentric fashion as contiguous conserved and class-specific residues appear at successively finer partitions. On the other molecular faces, an initial lack of signal gives way to a non-coherent scattering of residues that appears only at the finest functional resolution, partition G1. These residues do not display the regular, contiguous expansion pattern seen in face A. Rather, they appear individually and lack contact with other conserved or class specific residues. The same observations hold for the evolutionary trace obtained from the smaller tree, although scattered signals appear earlier on the other molecular faces, beginning with partition D2 (see Figure 4). The pattern of emergence of cluster 1 is characteristic of a functionally significant region, where residues are conserved within each cluster, but can vary between them.

The extent of cluster 1 is difficult to define unambiguously since it grows with each finer partition. To delineate this region, we use the emergence in the trace of scattered, non-clustering signal, as occurs in partition G1 of Figure 3 (PIC $\sim$85%), as a gauge that groups are becoming too small to reliably distinguish neutral drift from functional divergence given the relative paucity of sequences. To improve specificity at the possible expense of sensitivity, we seek the PIC threshold that minimizes the scattered background signal assessed visually. This corresponds to E1 in Figure 3 or C2 in Figure 4. In both partitions, cluster 1 clearly stands out as a region of coalescing conserved and class specific residues on a background that is otherwise free of signal.

Comparison of cluster 1 defined by these two traces, establishes that evolutionary tracing is not exquisitely dependent upon the size of the protein sequence family. Cluster 1 comprises 11 residues in partition E1, {R32, S34, H58, Y59, L94, I71, Y87, K57, K60, R74, C42} and 13 residues in partition C2, {R32, S34, H58, Y59, L94, I71, Y87, K57, K60, R74, R12, G93, C95}. These sets overlap over a core of ten residues, of which seven are internal (<30% of side-chain is solvent accessible) {R32, S34, H58, Y59, L94, I71, Y87}, and three are external {K57, K60, R74}. This overlap constitutes 91% of cluster 1 as defined by partition E1 and 77% as defined by C2. Differences occur at R12, G93 and C42. The first two residues are invariant in the larger src family and thus conserved in C2, but they are neutral in E1, which includes the distant kfes group (where they mutate). E1 has a higher PIC and greater functional resolution than C2. Thus the src family is broken into finer groups in E1. The variation of C42 in the greater src family confers neutrality in C2, but becomes class-specific in E1. All three residues extend or complete the common cluster 1 core without significantly altering its overall location (see Figure 5). Thus cluster 1 is specified clearly whether an extensive tree is used (E1) or a more limited approach is taken (C2).

A comparison of the evolutionary trace and crystallographic results shows that the binding site of the SH2 domain is located specifically and accurately (see Figure 5). The structural epitope, defined by residues that are within 4 Å of the ligand, consists of 16 residues (Figure 5, in cyan): {R12*, R32*, S34*, E35, T36, C42, K57*, H58*, Y59*, K60*, I71*, T72, Y87*, D92, G93*, L94*}. It agrees with 10 of 11 cluster 1 residues (underlined) defined by E1 (Figure 5), yielding 91% specificity over the entire tree. Over the smaller tree, it overlaps at 11 of 13 (asterisk) residues from the C2 trace. The specificity remains high at 85%. Ten of 16 positions in the structural epitope are covered by cluster 1 defined by E1 (63% sensitivity), and 11/16 when C2 is used (69% sensitivity).

In the family of SH3 domains, evolutionary tracing identifies a single functionally important surface structural cluster that matches the ligand binding site. As was observed in SH2 domains, the SH3 domain dendrogram mirrors the functional divergence of the parent protein. Figure 6 shows the partitions generated by steadily increasing PIC. The corresponding evolutionary traces are mapped onto the crystal structures of the ABL SH3 domain (Musacchio *et al*., 1992) in Figures 7 and 8. Little extraneous signal occurs. From a low PIC of 25% to a high of 80%, a single dominant cluster arises on face A (cluster 1), which displays a steady, expanding pattern reminiscent of that seen in the SH2 domain analysis. A second signal appears later on face D at partition Dl, but by partition E1 it has merged with and extended cluster 1. The actual

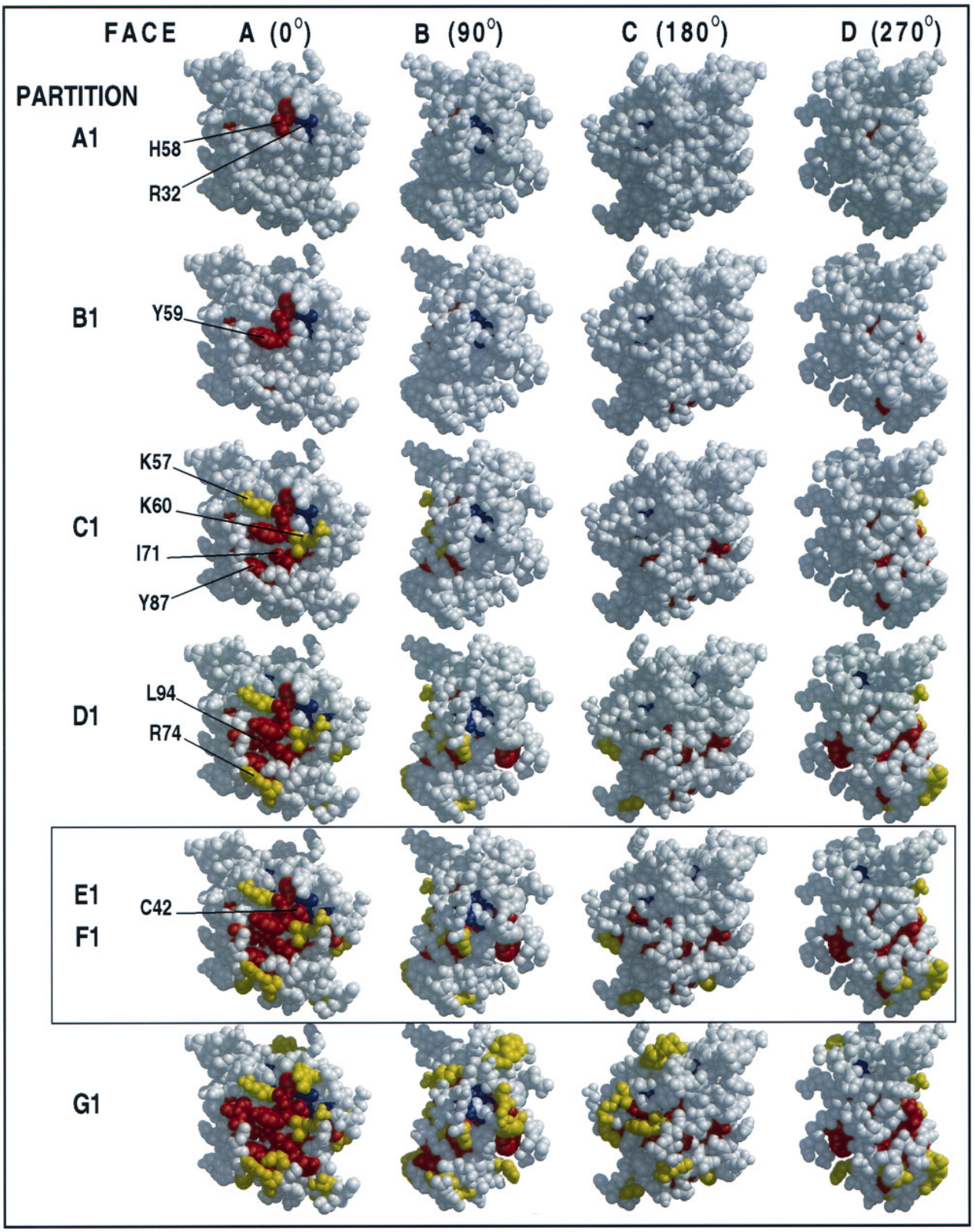

Figure 3. Evolutionary trace of the complete family of SH2 domains mapped onto the 2.7 Å resolution structure of Src SH2 domain (Waksman *et al*., 1993). Each row displays the protein in sequential 90% rotations about the *z*-axis, and successive rows show traces that arise from partitions A1 to G1, as defined in Figure 1. The internal positions that are conserved (dark blue) and class-specific (red), and the external class-specific (yellow) and conserved (cyan, though none in this Figure) positions are distributed inhomogeneously and form a single localized cluster on face A. At low PIC, partitions A and B, the trace is specific and points to only three residues, R32, H58 and Y59. To these core residues K57, K60, I71 and Y87 are added contiguously in partition C, L94 and R74 in partition D and C42 in partition E. The other faces remain essentially free of trace signal until partition G1. There, the scattered appearance of trace positions suggests that the useful limit of functional resolution has been reached at that higher PIC value. Partition F1 is not shown because it is equivalent to partition E1.

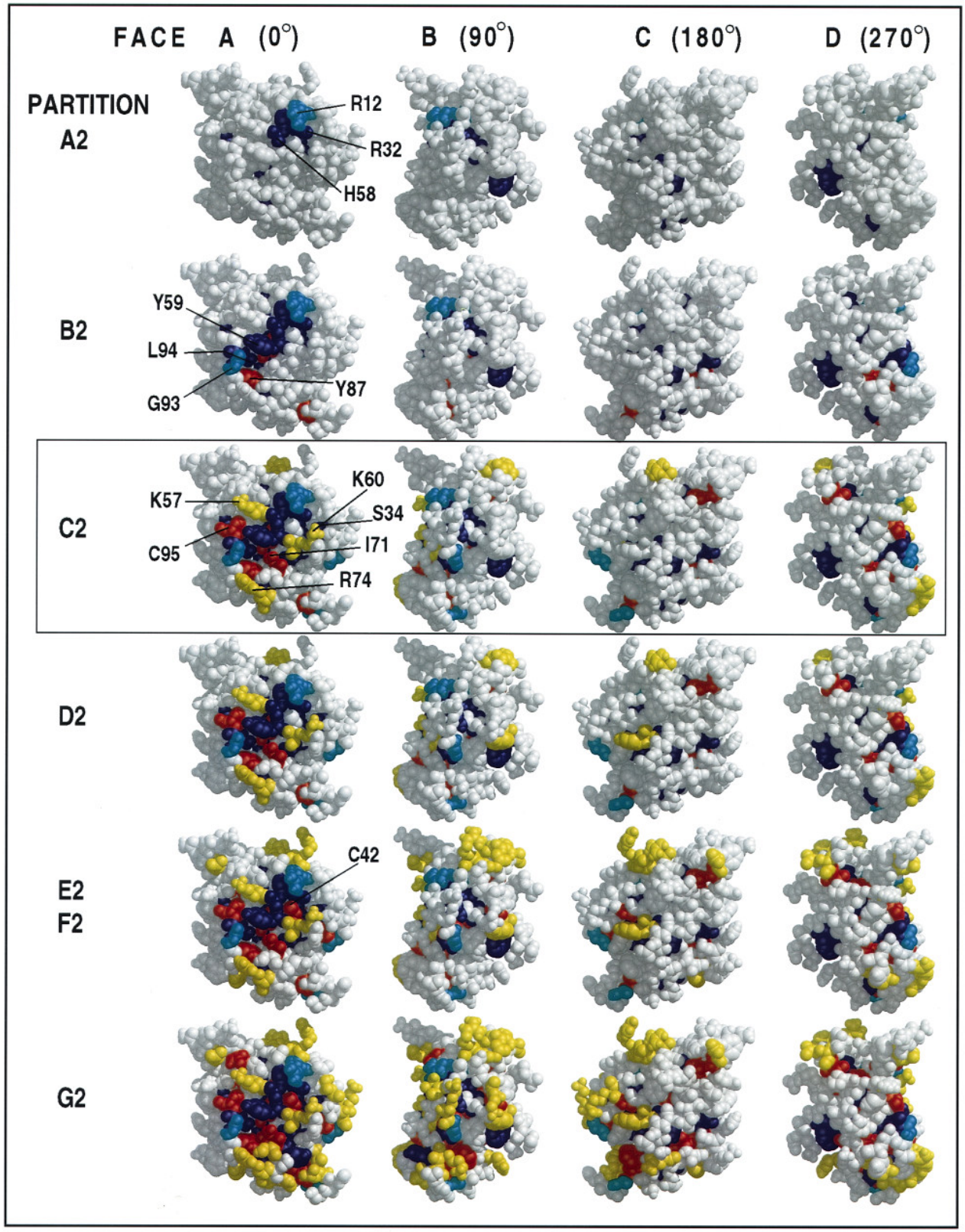

Figure 4. Results of an ET analysis performed over a subfamily of SH2 domains, consisting of the sequences between fg_kabl and fg_nck in Figure 1. PIC values, orientation of the Src SH2 domain and color codes are as in Figure 3. The trace finds the same spatial region as it did in Figure 3 to be functionally important, though full definition of the cluster (partition C2) and the appearance of scattered signal (partition D) occur earlier, at a lower PIC value.

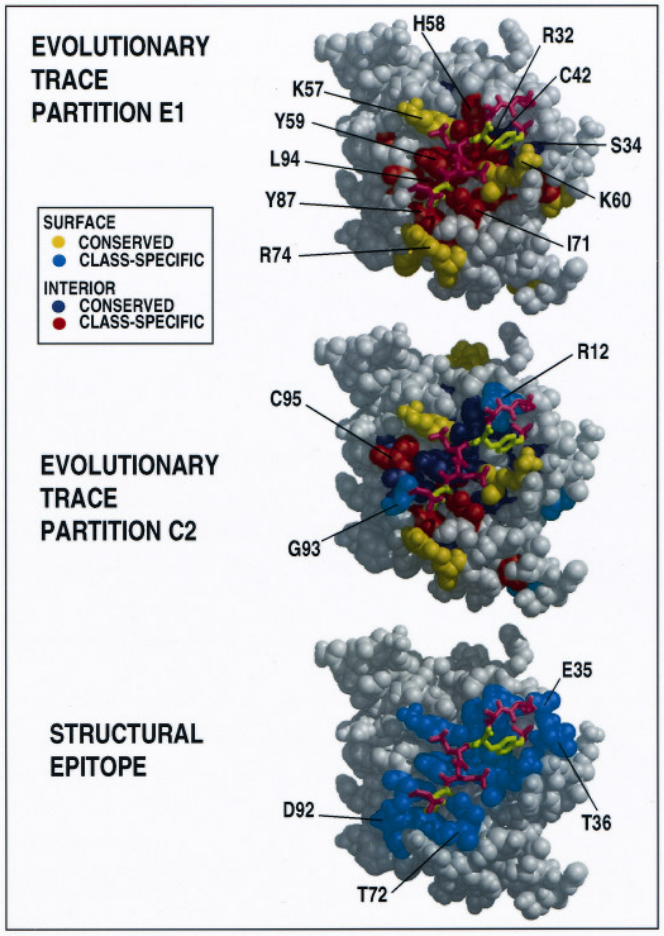

Figure 5. Three equivalent views of Src SH2 with the bound PQ(pY)EEI ligand (Waksman *et al*., 1993) documents the extensive overlap between functionally important regions determined by ET over the complete tree (partition E1, Figure 3, top), and over the restricted tree (partition C2, Figure 4, middle), with the structural epitope shown in cyan (residues that are within 4 Å of the ligand, bottom). The color scheme is shown, and two key ligand residues are in yellow: phosphotyrosine (pY) and isoleucine at pY + 3.

ligand binding site is shown in Figure 9, and highlights the overlap between cluster 1 and the actual binding site.

In partition E1, cluster 1 comprises T100, W120, F93, P133, Y136 and Y91, and reaches around to face 4 to include L90. By comparison the structural epitope (4 Å cutoff) is {**Y136**, **W120**, **Y91**, N135, **P133**, E119, N115, **T100**, W131, D98, S96}, overlap shown in bold. Thus the specificity of cluster 1 is 71% (5/7) and it covers 45% (5/11) of the structural epitope at partition level E1. At a coarser partition, cluster 1 is 100% specific and comprises the two residues closest to the ligand, W120 and Y136.

When ET analysis is restricted to the smaller tree composed by the {ksrc, kabl} family only, the results remain essentially unchanged, although G106, A89 and K105 are added to the face D extension of cluster 1. The contiguity of these residues to the expanding cluster, the complete invariance of K105, and the lack of a background signal surrounding argue that these residues are significant functional extensions of cluster 1 (see Figure 8). K105 is neutral in the larger tree because of non-conservative mutations (R105 → C105 → K105) in the distant katk group of sequences. Specificity at partition D2 falls to 44% (4/9) and sensitivity to 36% (4/11), largely due to the extension onto face D.

Even in the smaller, more sensitive tree, structural epitope residues S96, D98, W131 and N115 remain neutral. But discrepancies between the structural epitope and ET analysis do not necessarily undermine the approach. Rather, it may be that S96, D98, W131 and N115 do not play critical energetic roles in ligand binding. This interpretation is supported by mutations at W131 and N115, which did not affect binding (Lim & Richards, 1994), and the observation that over the entire family of SH3 domains, S96 and D98 are extremely variable (ten and nine different residue types appear at their respective positions). On the other hand, a mutation at F93, absent from the structural epitope but present in our traces, did affect binding (Lim & Richards, 1994). Hence, with these caveats in mind, specificity of the SH3 analysis rises in the small {kabl,ksrc} tree to 55% (5/9) and the sensitivity increases to 50% (5/10).

Nuclear hormone receptor DNA binding domains (DBD)

Nuclear hormone receptors are an entirely distinct class of proteins on which to test ET. In response to a hormone ligand, proteins in this family bind DNA at the hormone response element, and thus initiate downstream DNA transcription (Chambon, 1995; Evans, 1988; Rusconi & Yamamoto, 1987; Yamamoto, 1985). Binding activity to DNA has been extensively characterized by X-ray crystallography (Luisi *et al*., 1991; Newcomer *et al*., 1993; Rastinejad *et al*., 1995; Schwabe *et al*., 1993), and resides in zinc-finger domains (ZnF) that dimerize in different configurations to recognize palindromes or double-repeats. ET analysis was carried out over 80 homologous sequences of ZnF domains gathered from the sequence database, to test whether the DNA binding site common to all ZnF domains could be identified.

Figure 10 shows three views of ET results mapped onto the glucocorticoid receptor ZnF homodimer bound to its palindromic DNA target (Luisi *et al*., 1991). There is little ET signal except on the surface that binds DNA (Figure 10) where the evolutionary trace forms two clusters, one for each ZnF domain. These identical clusters contain eleven residues, {D445, H451, Y452, G458, S459*, K461, K465, R466, R489, K490, P493*} that directly face the DNA ligand. Mutations at S459 and P493 cause the nuclear hormone receptor to interact with other transcriptional activators in the absence of specific DNA binding (Lefstin *et al*., 1994). Furthermore, Thomas and Yamamoto have

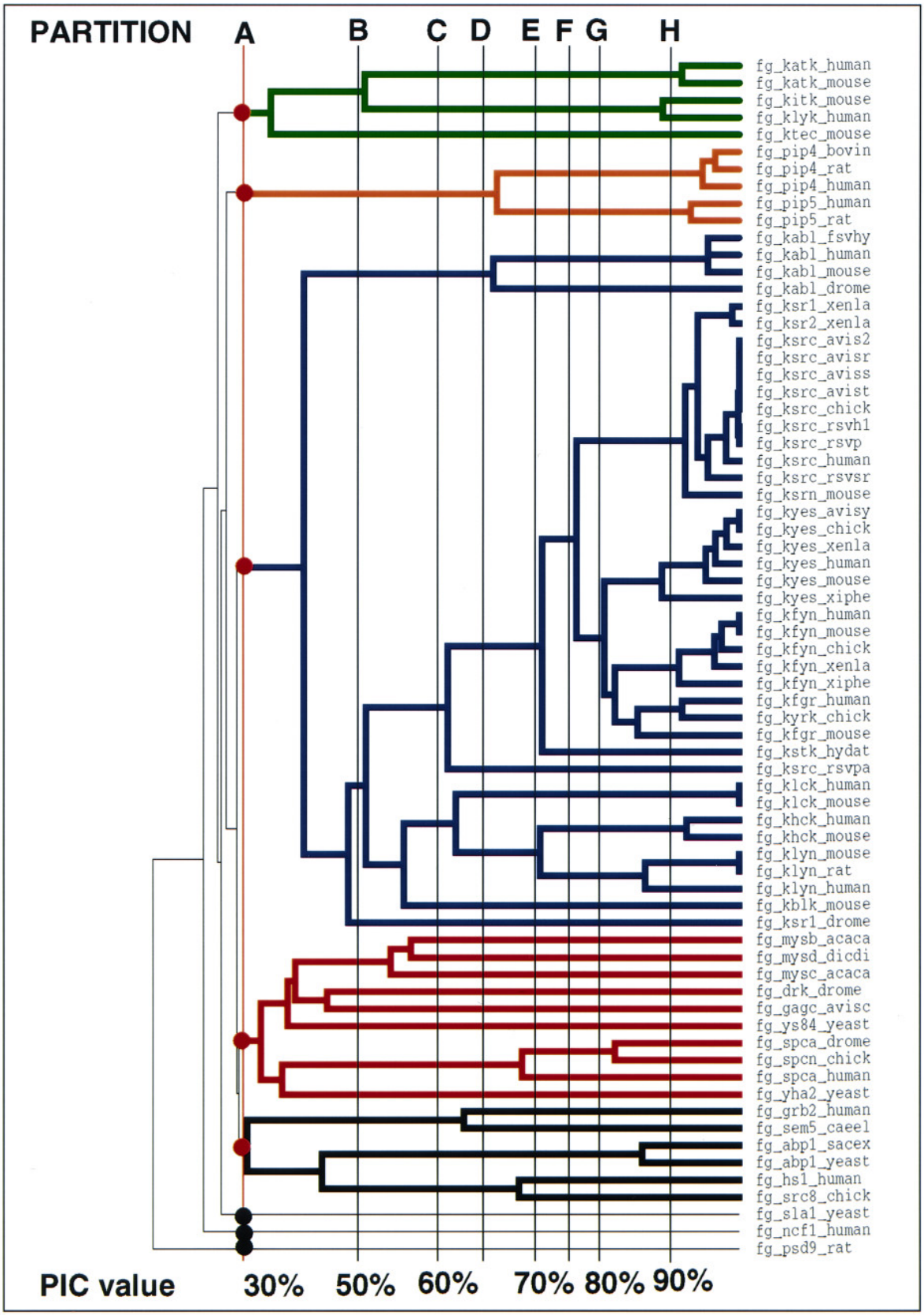

Figure 6. Sequence identity dendrogram of the SH3 domain family generated by PILEUP of sequences extracted from the Swiss protein databank with a FASTA search for kfyn_human related sequences (see Methods). As described for Figure 1, PIC lines A to G were used to define partitions. The entire tree was used in the results depicted in Figure 7, and a restricted tree, shown in purple, spanning sequences kabl_fsvhy to ksrt1_drome was used for the analysis in Figure 8.

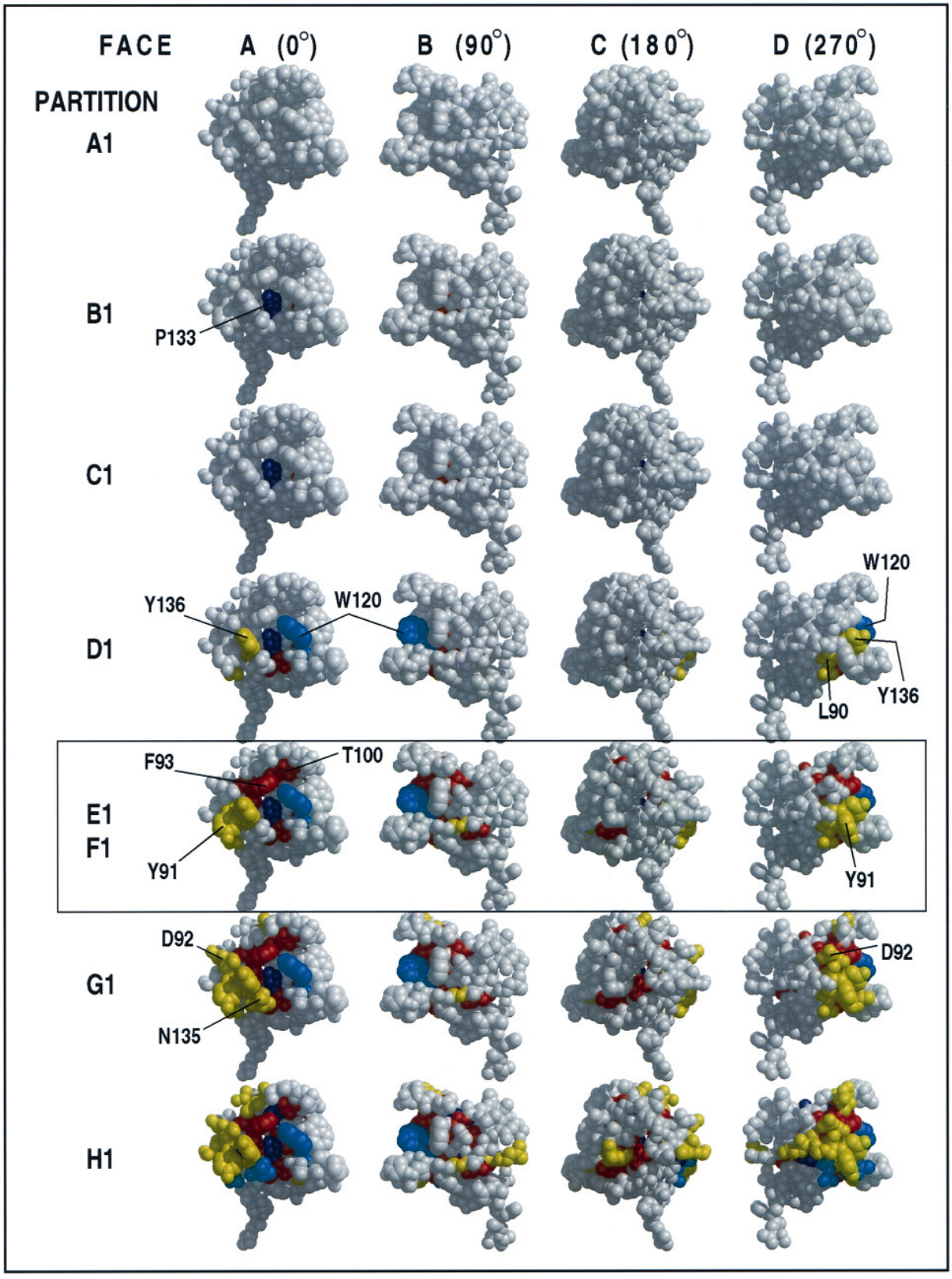

Figure 7. Evolutionary trace of the SH3 domain family mapped onto the 2.8 Å resolution structure of Abl SH3 domain (Musacchio *et al*., 1994), conventions follow those of Figure 3 and the traces correspond to the PIC values in Figure 6. A single domain is defined by evolutionary tracing to be functionally important, its core comprises position P133 (B1 and C1), and the adjacent W120, Y136 (D). The cluster is well defined in partitions E1 and G1, and scattered signal appears in partition H1. Faces B1 and D1 show side views of this cluster.

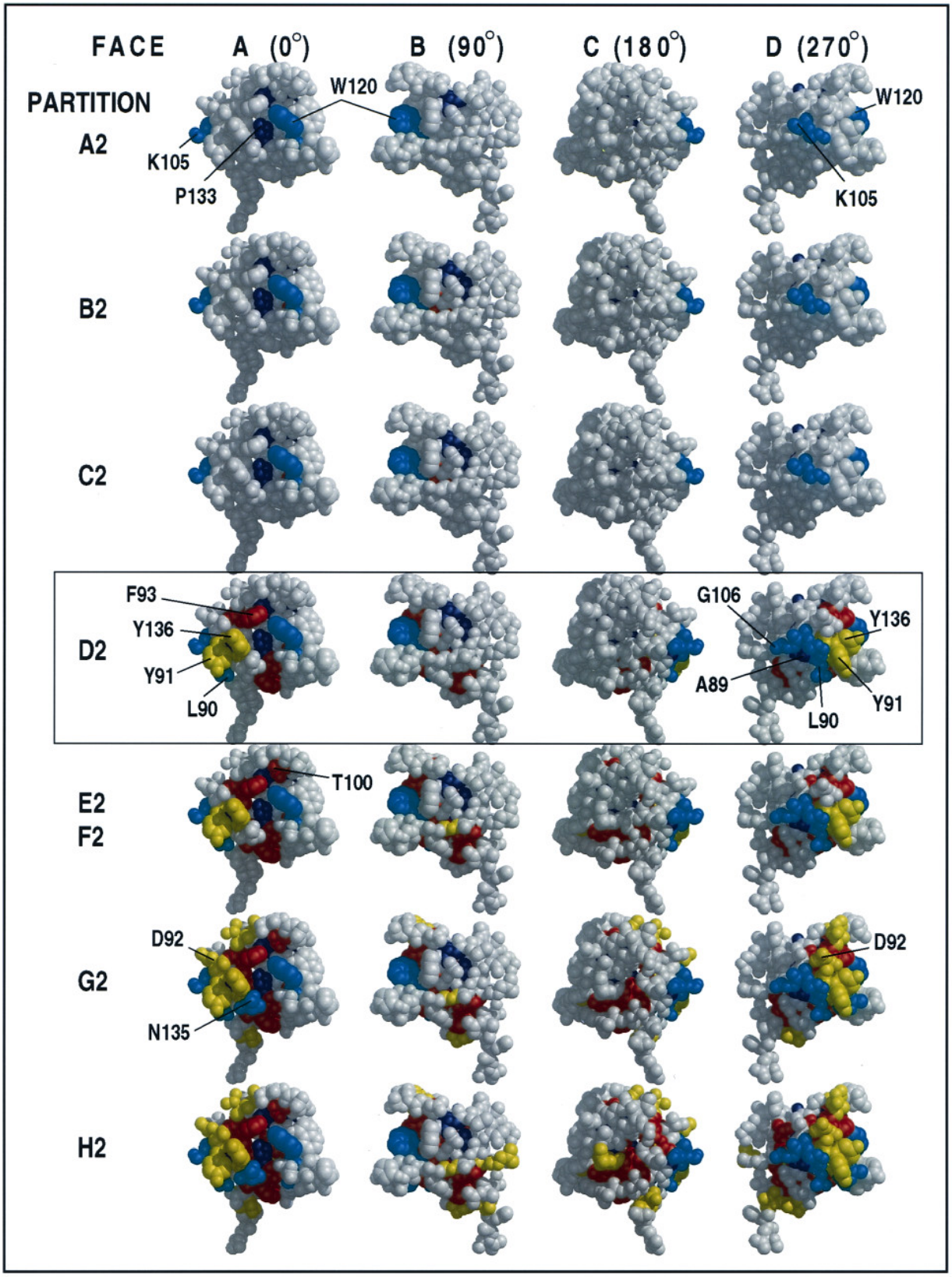

Figure 8. ET analysis carried out over a restricted SH3 domain family (see Figure 6). PIC choices, Abl SH3 orientation and color codes are unchanged from Figure 7, and the same region, on face A, is picked out by ET as functionally important. The face A cluster now prominently extends onto face D, consistent with a possible functional role distinct from ligand binding, in SH3 domains from tyrosine kinases.

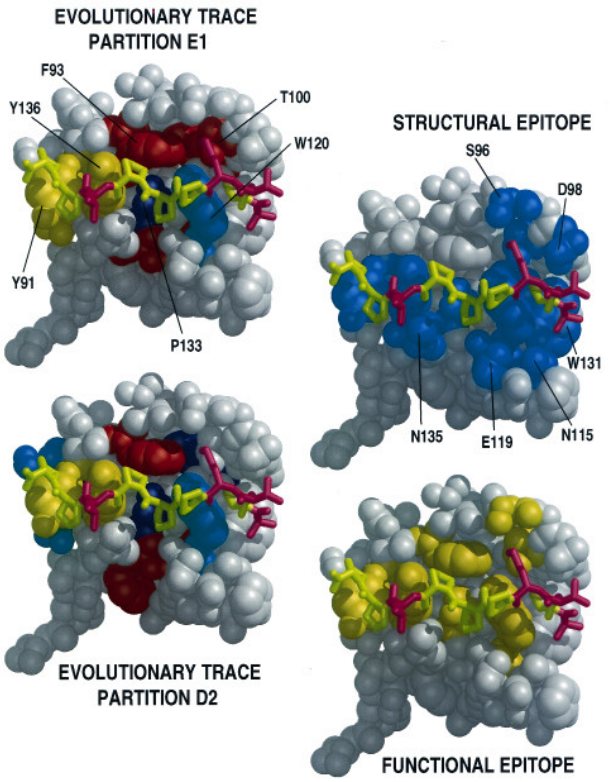

Figure 9. Four identical views of the ligand binding site of Abl SH3, the proline residues from the ligand are colored yellow (Musacchio *et al*., 1994). The left panels show the functional site picked by ET over the large and small SH3 domain families, respectively. There is extensive overlap with the structural epitope (top right) but the match is best with the functional epitope (bottom right), where yellow positions mark residues where mutations caused impaired function (Lim & Richards, 1994).

independently mutated the remaining residues and found seven positions, underlined, which severely compromise DNA binding (Thomas, 1993). V462, which is surrounded by functionally important residues and has been shown to be functionally important to binding, prominently fails to appear in the trace, due to a single mutation, A462S, in the thyroid hormone receptor group. The dimerization interface between Znf domains also fails to appear in the trace. This is expected, since different ZnF domains dimerize either as homodimer or heterodimers, and each mode involves distinct, non-overlapping regions. Thus, positions in the homodimer interface, shown in Figure 10, appear neutral because they are neither conserved nor class-specific among zinc fingers that heterodimerize.

Discussion

The ET method exploits the information inherent in a family of homologous proteins by dividing it to maximize functional similarity within groups and functional variation between groups. This functional partition of the protein family ensures that neutral positions within a group are less likely to occur at functional sites. By jointly mapping the neutral positions of each group onto the protein, the union of positions at which residues are unlikely to play a functional role can be contrasted with the remaining positions (which form the intersection, over all groups, of invariant positions). These latter positions either display mutations that correlate with functional divergence, or no sequence record exists in our databases to suggest they have ever mutated. Although these positions may be conserved or class-specific by coincidence, they are more likely to be functionally important. No clear inference can be made about these positions if they are scattered throughout the protein structure. If they cluster spatially however, they form a site characterized by an unusually low rate of mutation, all of which occur in concert with functional divergence. We believe such spatial clusterings of conserved and class-specific residues identify evolutionarily privileged functional sites, which represent ancestral functional regions that have remained common to all protein descendants.

Proper functional partitioning is essential in ET analysis. In the absence of detailed biochemical insight into functional variation, we rely additionally on sequence identity clustering to partition a family into functional subgroups. The PIC parameter defines the extent of sequence similarity within each group, and by varying PIC the functional resolution of the evolutionary trace can be controlled. At low PIC, a few large clusters separate sequences into broad functional groups. Conserved and class-specific residues identified at such a low functional resolution indicate the highest degree of evolutionary preservation and hence should correlate best with critical functional properties. At high PIC many more clusters appear; this refines the partitioning and allows identification of residues that contribute to finer functional nuances.

The SH2 and SH3 domain and the nuclear hormone receptor DBDs are useful test cases for the ET approach. First, all have known crystal structures and well-characterized ligand binding sites at which the variation of key residues regulates specificity. Thus correct detection of their binding site does not depend on recognizing the simple invariance of key residues, but requires the more subtle detection of a site, or sites, where most residues vary in a class-specific pattern that correlates with functional distinctions. In addition, a sufficient number of crystal structures from various family members is available and provides explicit evidence of the preservation of structural similarity over a wide range. Thus the assumption that proteins in the dendrogram can be mapped onto a single prototypical structure is, not surprisingly, justified.

It is logical to investigate the specificity and sensitivity of an evolutionary trace as a function of PIC. As shown in Figures 5 and 9, the predicted

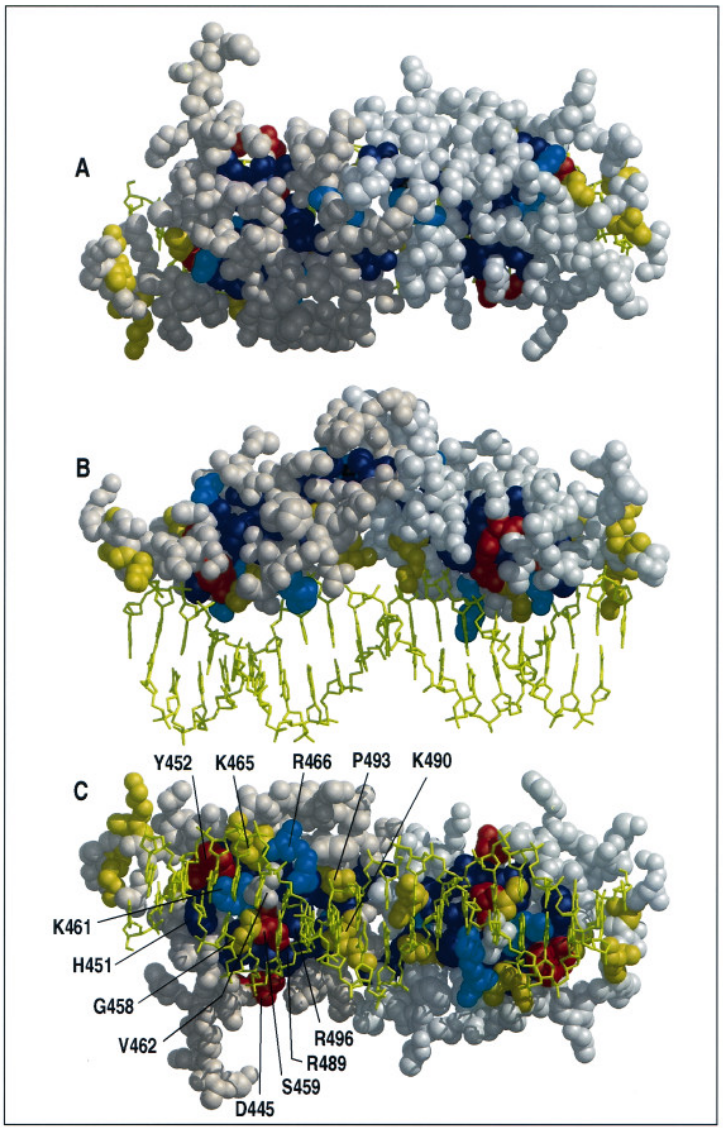

Figure 10. ET analysis on the nuclear hormone receptor family of DNA binding domains (zinc fingers). Two zinc fingers homodimerize and are complexed to a palindromic DNA hormone response element, in yellow (Luisi *et al*., 1991). A to C Consecutive 90° rotations about the *x*-axis. ET mapping follows the usual color code. No functionally important cluster occurs except in C, the DNA binding surface, where conserved and class-specific positions form a single cluster per ZnF domain. Upon homodimerization these clusters create a large area of functionally important positions at the DNA binding interface.

ligand binding sites in both SH2 and SH3 domains correlate with experiment. At low PIC the signal is highly specific and identifies the functionally important residues. In the SH2 domain, the signal appears only at the center of the binding site location (H58, R32 in A1) and then it expands to include Y59 (B1) and K57, K60, I71, Y87 (C1). Waksman *et al*. (1993) and others (Eck *et al*., 1993) have shown that residues 58, 59 and 60 form a pocket in which the phosphotyrosine aromatic ring binds. R32 makes a key hydrogen bond to the phosphate group (Bibbins *et al*., 1993) and I71 and Y87 line the hydrophobic recess where the side-chain of the third residue C-terminal to pY (pY + 3) binds. K57, though not in contact with the ligand, lies within 4 Å of pY + 2, a ligand subsite that frequently contains a negatively charged side-chain (E in the YEEI high-affinity peptide of src). Proteins containing mutations at R32, C42 and H58 have no detectable binding (Bibbins *et al*., 1993; Marengere & Pawson, 1992; Mayer *et al*., 1992; Yoakim *et al*., 1994). Remarkably, R12 which makes four hydrogen bonds to pY and thus seems to be functionally critical, is conspicuously missing from the cluster identified using the most sensitive cutoff (El). When this residue was mutated to E, A, P or K, however, ligand binding was not affected (Bibbins *et al*., 1993; Marengere & Pawson, 1992; Mayer *et al*., 1992; Yoakim *et al*., 1994). Thus, the residues uncovered by low-resolution evolutionary tracing precisely outline the essential elements of the binding site from the phosphotyrosine residue to the pY + 3 binding site. The residues that appear only at higher PICs are not as critical to function.

Similar conclusions hold in the SH3 domain, where P133 appears first (at partitions B1, C1, Figure 7), followed by Y136, W120 and L190 (D1) and then F93, T100, Y91 (E1). F93, W120, P133 and Y136 form the core binding pocket where the ligands core, P6 and P7, lie. W120 also forms the edge of the P2 contact site, while T100 is within 4 Å of the ligand's methionine residue M4. Finally, Y91 and Y136 define the ligand contact area of P9 and P10 (Musacchio *et al*., 1992).

The distinction between structural and functional epitopes, and the correlation between the evolutionary trace with the functional epitope are particularly clear in the SH3 domain example. Lim & Richards (1994) demonstrated that mutations of F93, W120, Y136, P133 and Y91 resulted in significant changes in binding affinity. They also showed that the coupled mutation of S96 and T100 affected ligand binding. These residues closely match the trace in E1. The single false positive is L90, which was predicted to play a functional role yet did not affect binding upon mutation. Mutations at V114, N115 and D92 had little effect, and W131 had none. Both W131 and N115 are neutral in E1 and are not in the functional epitope. Thus, in SH3 domains, residues identified at low PIC values form the outline of key functional contacts, while positions that were not in the trace were tolerant to mutations.

Internal and surface residues evolve under distinct sets of mutational pressures. To mutate, internal residues must satisfy the packing constraints created by neighboring residues or must participate in a stable repacking of the protein core. By contrast, alterations of external side-chains are largely free of structural restrictions. Functional constraints on surface residues operate only in specialized regions of the molecule. In the SH2 and SH3 domains, and in the nuclear hormone receptor DNA binding domains, the subset of residues detected by ET are dominated by positions inaccessible to the solvent. The distinction between internal and external locations blurs at cleft sites where residues that have over 70% of their side-chain buried have fewer neighbors and substantially contribute to surface electrostatics (Honig & Nicholls, 1995). By plotting the ET residues onto a space-filling model of the protein three-dimensional structure, the functionally important solvent-accessible and cleft positions are highlighted. For example, in the SH2 domain, cluster 1 consists of an entire cleft plus surface residues at its edge, as seen in E2. If solvent accessibility was used as the only guide to distinguish structurally from functionally important positions, the unique functional characteristics of some protein clefts would be masked and the contiguity of the functional epitope would be disrupted. The same issues arise in the ET analysis of the functional epitope in the SH3 domain (see Figure 9), and at the DNA interface of the ZnF domains (see Figure 10). Since clefts are a common feature of active sites, we expect this finding to hold for many proteins.

As proteins undergo neutral drift during evolution, it is important to assess the signal to noise problems that are inherent in sequence analysis. We have investigated strategies for distinguishing a true positive signal from a false positive noise. When a small number of sequences are available, it is likely that some positions that should be neutral will not demonstrate sufficient variation to achieve this designation. These positions should however occur at random and should be scattered throughout the structure. By contrast, it is unusual for a residue to be functionally important regardless of its sequence and structural context. It is more likely to be surrounded by spatially proximal residues that are themselves functionally important and hence more easily recognized. We currently rely on visual inspection to distinguish signal from noise and some subjectivity is inevitable. Explicit clustering algorithms are being developed to automate this part of the sequence-structure analysis.

The basic distinction between signal and noise is borne out in the SH2 and SH3 domains over the $\{$src, abl$\}$ set and in the ZnF domains (data not shown). The earliest residues to appear are those that are most essential to function. When noise appears, it is scattered and does not involve contiguous internal and surface residues (see Figures 3 and 4, partitions G1, D2 and subsequent ones).

These criteria apply also in any tree partitioned with a large PIC, chosen so as to maximize the sensitivity of the evolutionary trace. When PIC is large, subgroups become more numerous but necessarily contain fewer sequences. Thus each consensus sequence is obtained over fewer, increasingly similar proteins. As the number of sequences per group falls, invariance within each group is a less stringent test of evolutionary pressure, and in the limit it becomes meaningless when groups are reduced to singlets (such singular groups provide no basis for identifying neutral positions, cannot contribute to site localization and effectively drop out of the analysis). Indeed, a strong increase in scattered signal is seen in G1 for SH2 at a PIC value of 80%, where many previously neutral residues now emerge as class specific. The loss of specificity can be remedied by considering new signals only if they are near a previously recognized cluster. This allows the detection of residues that play a lesser role in binding specificity. Such residues would be expected to be variable, until the PIC is large enough that proteins with subtle difference in their binding sites become segregated in distinct groups.

The three residues detected in this manner in the SH2 domain, C42, T72 and R12, prove to be modulators of function. C42 and T72 are conspicuously neutral ''holes'' in cluster 1, filled in only at high PIC (E1 for C42 and G1 for T72). R12 appears even later in G1 in a direct extension of the core H58 and R32. Position 42 is neutral in the E1 partition where it is cysteine in src, and serine in its subgroup partners $\{$Lck, Hck$\}$ and $\{$Yes, Fyn, Fgr$\}$. In D1, these subsets divide onto separate branches of sequences and position 42 becomes class-specific. Of note, this mostly buried residue neighbors the aromatic ring of the phosphotyrosine residue and must influence binding specificity, perhaps indirectly through an effect at pY + 1. A single mutation is described at that position, T → I in the p85 family with a tenfold loss of affinity (Yoakim *et al*., 1994). R12 remains neutral until PIC is high enough to separate the mouse (W12) and human

(R12) kfes sequences. This residue contacts pY through four hydrogen bonds, yet G12 is a functional substitution in both the ptnb and Syp sequences. The crystal structure of the latter illuminates how a rotation of the phosphate group suffices to restore the overall number of hydrogen bonds formed by the ligand (Lee *et al*., 1994). As already mentioned, its mutation decreased its binding affinity, but did not abolish it. T72 is a variable (eight residue types) surface residue that is part of a small loop, EF, which delimits the pY + 3 side-chain binding pocket. Its mutations have been shown to switch the binding specificity of src-sh2 to that of grb2-sh2 (Yoakim *et al*., 1994).

As currently implemented, neutrality can be misleadingly assigned, and lead to a loss of information in two ways. First, we treat all mutations equally so that conservative substitutions are not distinguished from non-conservative ones. Thus in the SH2 domain, position G93 is in the structural epitope but is labeled as neutral in the trace because it can be R or K in the kfes cluster. In effect, by applying this overly restrictive criterion, and counting as fully neutral this charge-preserving mutation, we enhance specificity at the expense of sensitivity. This is a sensible strategy in trees with relatively few sequences or in families with low overall mutation rates where most positions appear conserved or class-specific in the trace. Greater leniency in defining neutrality would, in this setting, cause too great a loss of specificity to warrant the sensitivity gain. As an increasing number of sequences become available in databases, it should become possible to relax the definition of neutrality.

A second cause of false negatives will be sequence misalignment. To the extent that a misalignment occurs away from the active site, this has no impact on the value of the trace. Fortunately, sequences are most similar at their active sites (Zvelebil *et al*., 1987) so that we expect the trace will be relatively resistant to misalignment. Of course, if the degree of sequence identity is exceptionally low, it is possible that significant structural differences will arise that would undermine ET analysis. In practice, complete disappearance of signal suggests a false alignment or structures that are not truly homologous.

Conclusions

Evolutionary tracing is a new method for uncovering functionally important residues in proteins. We systematically segregate homologous sequences into functional groups and ask: in what way are the conservation patterns of each group similar to one another?

Our results show that in both the SH2 and SH3 domains, and in the DNA binding domains of nuclear hormone receptors, the evolutionary trace identifies the ligand binding site. Agreement, especially with the essential functional residues, is good and specific at lower PIC values. At larger PIC values, the trace has greater functional resolution and residues contributing to functional modulation round out a full picture of the functional epitope as defined experimentally. ET analysis has been applied to the family of Gα proteins and to functional subgroups within the ZnF family of nuclear hormone receptors (unpublished results). In all cases, class-specific positions are crucial to the recognition of the binding site. This demonstrates the value of functional partitioning as well as the basic soundness of an evolutionary analysis.

An inherent limitation lies in the availability of multiple homologous sequences. Hopefully this will diminish with time. More fundamental is the basic requirement for stability of the functional site structural motif through evolution. This may not be a severe limit on multifunctional proteins where the interplay of functions strongly limits tolerance for mutations. But it may be a problem with a protein that has a single interface with a partner equally free to evolve. In such a setting, the interfaces of evolving descendants could simply drift away from the ancestral site and from each other. Since the trace yields their intersection, it could miss large parts of the interface for any particular group. Refocusing the analysis on a smaller tree would palliate this problem, provided sequence availability permits subgroup analysis.

As more sequences become available, the full power of this approach can be developed further. A probabilistic treatment of sequence mutations could define an expected mutation rate throughout the structure and assign a formal statistical significance to deviations from it, at each position. This would permit one to sharpen the blunt criteria we use for neutrality so as to distinguish mutation types rather than simply record their occurrence. A quantitative probabilistic scoring of the structural clusters derived from the trace would also help to use this approach in high noise settings. In its current form, however, the evolutionary trace is already a practical tool that can usefully predict functionally important residues, and assist in targeting mutagenesis studies.

Methods

The standard FASTA tool (Pearson, 1988) from the GCG sequence analysis package (Devereux *et al*., 1984) was used to gather sequence fragments that matched the proband domain of interest. For the SH2 domains, a FASTA search over the Swiss Protein Databank was carried out using human src_sh2 as the proband sequence. The kfyn_human and gcr_rat sequences, respectively, were used for the analogous searches of SH3 domains and NHR DBDs. The lists were truncated when the proteins retrieved displayed identity over short stretches of sequence only and when their function became clearly unrelated. Sequence alignment and dendrogram construction was then carried out with the GCG multiple sequence alignment tool PILEUP (Feng & Doolittle, 1987; Higgins & Sharp, 1989), using default settings.

We implemented the evolutionary trace method in the program TRACE. Its input consist of the MSA from PILEUP, a reference X-ray or NMR structure and the name of this reference protein in the MSA. The evolutionary trace is built as illustrated in Figure 2, and the output is a list of conserved and class-specific residues broken down as internal or external positions on the protein structure. The latter distinction is defined by the percentage solvent exposure of the side-chain (or of the entire residue for glycine) as calculated by the ACCESS program (Lee & Richards, 1971). The TRACE output produces source files readable by the molecular graphics package MIDASPlus (Computer Graphics Laboratory, UCSF) with which all molecular Figures were generated (Ferrin *et al*., 1988; Huang *et al*., 1991). This work was performed on Silicon Graphics workstations.

Acknowledgements

We thank Alexis Falicov and Nomi L. Harris for their help in implementing the TRACE algorithm, and Scott R. Presnell for the use of the ACCESS program, which is a modification of the original Lee-Richards algorithm. We thank Keith Yamamoto for his comments. This work was supported by the American Heart Association, California Affiliate, and the National Institute of Health.

References

Baldwin, J. M. (1993). The probable arrangement of the helices in G protein-coupled receptors. *EMBO J.* **12**, 1693–1703.

Barton, G. J. (1990). Protein multiple sequence alignment and flexible pattern matching. *Methods Enzymol.* **183**, 403–428.

Benner, S. A. (1989). Patterns of divergence in homologous proteins as indicators of tertiary and quaternary structure. *Advan. Enzyme Regul.* **28**, 219–236.

Benner, S. A., Badcoe, I., Cohen, M. A. & Gerloff, D. L. (1994). *Bona fide* prediction of aspects of protein conformation. Assigning interior and surface residues from patterns of variation and conservation in homologous protein sequences. *J. Mol. Biol.* **235**, 926–958.

Bibbins, K. B., Boeuf, H. & Varmus, H. E. (1993). Binding of the Src SH2 domain to phosphopeptides is determined by residues in both the SH2 domain and the phosphopeptides. *Mol. Cell. Biol.* **13**, 7278–7287.

Birge, R. B. & Hanafusa, H. (1993). Closing in on SH2 specificity. *Science*, **262**, 1522–1524.

Cantley, L. C., Auger, K. R., Carpenter, C., Duckworth, B., Graziani, A., Kapeller, R. & Soltoff, S. (1991). Oncogenes and signal transduction. *Cell*, **64**, 281–302.

Casari, G., Sander, C. & Valencia, A. (1995). A method to predict functional residues in proteins. *Nature Struct. Biol.* **2**, 171–178.

Chambon, P. (1995). The molecular and genetic dissection of the retinoid signaling pathway. *Recent Prog. Hormone Res.* **50**, 317–332.

Chothia, C. & Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. *EMBO J.* **5**, 823–826.

Cicchetti, P., Mayer, B. J., Thiel, G. & Baltimore, D. (1992). Identification of a protein that binds to the SH3 region of Abl and is similar to Bcr and GAP-rho. *Science*, **257**, 803–806.

Crawford, I. P., Niermann, T. & Kirschner, K. (1987). Prediction of secondary structure by evolutionary comparison: application to. *Proteins: Struct. Funct. Genet.* **2**, 118–129.

Devereux, J., Haeberli, P. & Smithies, O. (1984). A comprehensive set of sequence analysis programs for the VAX. *Nucl. Acids Res.* **12**, 387–395.

Du, P. & Alkorta, I. (1994). Sequence divergence analysis for the prediction of seven-helix membrane protein structures. I. Comparison with bacteriorhodopsin. *Protein Eng.* **7**, 1221–1229.

Eck, M. J., Shoelson, S. E. & Harrison, S. C. (1993). Recognition of a high-affinity phosphotyrosyl peptide by the Src homology-2 domain of p56lck. *Nature*, **362**, 87–91.

Eck, M. J., Atwell, S. K., Shoelson, S. E. & Harrison, S. C. (1994). Structure of the regulatory domains of the Src-family tyrosine kinase Lck. *Nature*, **368**, 764–769.

Evans, R. M. (1988). The steroid and thyroid hormone receptor superfamily. *Science*, **240**, 889–895.

Feng, D. R. & Doolittle, R. F. (1987). Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J. Mol. Evol.* **25**, 351–360.

Ferrin, T. E., Huang, C. C., Jarvis, L. E. & Langridge, R. (1988). The Midas display system. *J. Mol. Graph.* **6**, 13–27.

Guruprasad, L., Dhanaraj, V., Timm, D., Blundell, T. L., Gout, I. & Waterfield, M. D. (1995). The crystal structure of the N-terminal SH3 domain of Grb2. *J. Mol. Biol.* **248**, 856–866.

Higgins, D. G. & Sharp, P. M. (1989). Fast and sensitive multiple sequence alignments on a microcomputer. *Comput. Appl. Biosci.* **5**, 151–153.

Honig, B. & Nicholls, A. (1995). Classical electrostatics in biology and chemistry. *Science*, **268**, 1144–1149.

Huang, C. C., Pettersen, E. F., Klein, T. E., Ferrin, T. E. & Langridge, R. (1991). Conic: a fast renderer for space-filling molecules with shadows. *J. Mol. Graph.* **9**, 230–236.

Koch, C. A., Anderson, D., Moran, M. F., Ellis, C. & Pawson, T. (1991). SH2 and SH3 domains: elements that control interactions of cytoplasmic signaling proteins. *Science*, **252**, 668–674.

Lee, B. & Richards, F. M. (1971). The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* **55**, 379–400.

Lee, C. H., Kominos, D., Jacques, S., Margolis, B., Schlessinger, J., Shoelson, S. E. & Kuriyan, J. (1994). Crystal structures of peptide complexes of the amino-terminal SH2 domain of the Syp tyrosine phosphatase. *Structure*, **2**, 423–438.

Lefstin, J. A., Thomas, J. R. & Yamamoto, K. R. (1994). Influence of a steroid receptor DNA-binding domain on transcriptional regulatory functions. *Genes Dev.* **8**, 2842–2856.

Lim, W. A. & Richards, F. M. (1994). Critical residues in an SH3 domains from Sem-5 suggest a mechanism for proline-rich peptide recognition. *Nature Struct. Biol.* **1**, 221–225.

Livingstone, C. D. & Barton, G. J. (1993). Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. *Comput. Appl. Biosci.* **9**, 745–756.

Luisi, B. F., Xu, W. X., Otwinowski, Z., Freedman, L. P., Yamamoto, K. R. & Sigler, P. B. (1991). Crystallographic analysis of the interaction of the glucocorticoid receptor with DNA [see comments]. *Nature*, **352**, 497–505.

Marengere, L. E. & Pawson, T. (1992). Identification of

residues in GTPase-activating protein Src homology 2 domains that control binding to tyrosine phosphorylated growth factor receptors and p62. *J. Biol. Chem.* **267**, 22779–22786.

Mayer, B. J., Hamaguchi, M. & Hanafusa, H. (1988). A novel viral oncogene with structural similarity to phospholipase. *Nature*, **332**, 272–275.

Mayer, B. J., Jackson, P. K., Van, Etten. R. A. & Baltimore, D. (1992). Point mutations in the abl SH2 domain coordinately impair phosphotyrosine binding in vitro and transforming activity in vivo. *Mol. Cell. Biol.* **12**, 609–618.

Musacchio, A., Noble, M., Pauptit, R., Wierenga, R. & Saraste, M. (1992). Crystal structure of a Src-homology 3 (SH3) domain. *Nature*, **359**, 851–855.

Musacchio, A., Saraste, M. & Wilmanns, M. (1994). High-resolution crystal structure of tyrosine kinase SH3 domains complexed with proline-rich peptides. *Nature Struct. Biol.* **1**, 546–551.

Newcomer, M. E., Pappas, R. S. & Ong, D. E. (1993). X-ray crystallographic identification of a protein-binding site for both all-*trans*- and 9-*cis*-retinoic acid. *Proc. Natl Acad. Sci. USA*, **90**, 9223–9227.

Pawson, T. (1994). SH2 and SH3 domains in signal transduction. *Advan. Cancer Res.* **64**, 87–110.

Pearson, WR, L. D. (1988). Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.

Rastinejad, F., Perlmann, T., Evans, R. M. & Sigler, P. B. (1995). Structural determinants of nuclear receptor assembly on DNA direct repeats. *Nature*, **375**, 203–211.

Ren, R., Mayer, B. J., Cicchetti, P. & Baltimore, D. (1993). Identification of a ten-amino acid proline-rich SH3 binding site. *Science*, **259**, 1157–1161.

Rusconi, S. & Yamamoto, K. R. (1987). Functional dissection of the hormone and DNA binding activities of the glucocorticoid receptor. *EMBO J.* **6**, 1309–1315.

Sadowski, I., Stone, J. C. & Pawson, T. (1986). A noncatalytic domain conserved among cytoplasmic protein-tyrosine kinases modifies the kinase function and transforming activity of Fujinami sarcoma virus P130gag-fps. *Mol. Cell. Biol.* **6**, 4396–4408.

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

Schreiber, G. & Fersht, A. R. (1995). Energetics of protein-protein interactions: analysis of the barnase-barstar interface by single mutations and double mutant cycles. *J. Mol. Biol.* **248**, 478–486.

Schwabe, J. W., Chapman, L., Finch, J. T. & Rhodes, D. (1993). The crystal structure of the estrogen receptor DNA-binding domain bound to DNA: how receptors discriminate between their response elements. *Cell*, **75**, 567–578.

Songyang, Z., Shoelson, S. E., Chaudhuri, M., Gish, G., Pawson, T., Haser, W. G., King, F., Roberts, T., Ratnofsky, S., Lechleider, R. J., Neel, B. J., Birge, R. B., Fajardo, J. E., Chou, M. M., Hanafusa, H., Schaffhausen, B. & Cantley, L. C. (1993). SH2 domains recognize specific phosphopeptide sequences. *Cell*, **72**, 767–778.

Sternberg, M. J. & Cohen, F. E. (1982). Interferon: a tertiary structure predicted from amino acid sequences. *Phil. Trans. Roy. Soc. ser. B*, **299**, 125–127.

Thomas, J. R. (1993). The case for DNA mediated allosteric regulation of the glucocorticoid receptor. Doctoral dissertation, University of California, San Francisco.

Waksman, G., Kominos, D., Robertson, S. C., Pant, N., Baltimore, D., Birge, R. B., Cowburn, D., Hanafusa, H., Mayer, B. J., Overduin, M., Resh, M. D., Rios, C. B., Silverman, L. & Kuriyan, J. (1992). Crystal structure of the phosphotyrosine recognition domain SH2 of v-src complexed with tyrosine-phosphorylated peptides. *Nature*, **358**, 646–653.

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 structures of the complexed and peptide-free forms. *Cell*, **72**, 779–790.

Wells, J. A. (1994). Structural and functional basis for hormone binding and receptor oligomerization. *Curr. Opin. Cell Biol.* **6**, 163–173.

Wu, X., Knudsen, B., Feller, S. M., Zheng, J., Sali, A., Cowburn, D., Hanafusa, H. & Kuriyan, J. (1995). Structural basis for the specific interaction of lysine-containing proline-rich peptides with the N-terminal SH3 domain of c-Crk. *Structure*, **3**, 215–226.

Yamamoto, K. R. (1985). Steroid receptor regulated transcription of specific genes and gene networks. *Annu. Rev. Genet.* **19**, 209–252.

Yoakim, M., Hou, W., Songyang, Z., Liu, Y., Cantley, L. & Schaffhausen, B. (1994). Genetic analysis of a phosphatidylinositol 3-kinase SH2 domain reveals determinants of specificity. *Mol. Cell. Biol.* **14**, 5929–5938.

Zvelebil, M. J. & Sternberg, M. J. (1988). Analysis and prediction of the location of catalytic residues in enzymes. *Protein Eng.* **2**, 127–138.

Zvelebil, M. J., Barton, G. J., Taylor, W. R. & Sternberg, M. J. (1987). Prediction of protein secondary structure and active sites using the alignment of homologous sequences. *J. Mol. Biol.* **195**, 957–961.

Edited by B. Honig

(Received 19 September 1995; accepted 15 December 1995)