Volume 304, number 1, 15–20 FEBS 11086 June 1992

Conservation analysis and structure prediction of the SH2 family of phosphotyrosine binding domains

Robert B. Russell, Jason Breed and Geoffrey J. Barton

University of Oxford, Laboratory of Molecular Biophysics, The Rex Richards Building, South Parks Road, Oxford OX1 3QU, UK

Received 17 March 1992; revised version received 13 April 1992

Src homology 2 (SH2) regions are short (approximately 100 amino acids), non-catalytic domains conserved among a wide variety of proteins involved in cytoplasmic signaling induced by growth factors. It is thought that SH2 domains play an important role in the intracellular response to growth factor stimulation by binding to phosphotyrosine containing proteins. In this paper we apply the techniques of multiple sequence alignment, secondary structure prediction and conservation analysis to 67 SH2 domain amino acid sequences. This combined approach predicts seven core secondary structure regions with the pattern β-α-β-β-β-β-α, identifies those residues most likely to be buried in the hydrophobic core of the native SH2 domain, and highlights patterns of conservation indicative of secondary structural elements. Residues likely to be involved in phosphotyrosine binding are shown and orientations of the predicted secondary structures suggested which could enable such residues to cooperate in phosphate binding. We propose a consensus pattern that encapsulates the principal conserved features of the SH2 domains. Comparison of the proposed SH2 domain of *akt* to this pattern shows only 12/40 matches, suggesting that this domain may not exhibit SH2-like properties.

SH2 domain; Structure prediction; Conservation analysis; Alignment; Tyrosine kinase

1. INTRODUCTION

Several growth factors stimulate cell proliferation by binding to receptors with protein tyrosine kinase activities. Binding induces activation of the receptor and leads to autophosphorylation of the receptor. The induction of receptor activation is associated with tyrosine kinase activity and leads to phosphorylation of various cytoplasmic substrates, many of which are thought to be involved in intracellular signal transduction.

Recently, a conserved family of domains has been shown to mediate the association of cytoplasmic substrates and specific autophosphorylation sites on the receptors. These *src* homology 2 (SH2) domains have been found in a wide variety proteins involved in intracellular signal transduction (see Koch et al. [1], or Heldin [2] for reviews). SH2 domains are short (approximately 100 amino acids), non-catalytic regions, which are thought to facilitate recognition by binding to phosphotyrosine containing proteins.

No tertiary structure information on these SH2 domains is available at present. Further, since they show no obvious sequence similarity to any domains of known three-dimensional structure, structural inferences may not be made through homology modelling techniques.

Several sequence alignments of SH2 domains have been proposed [1,3–5]. Koch et al. [1] describe five conserved regions separated by four variable segments. The most highly conserved region is that corresponding to the FLVRES sequence found in *src*, which has been implicated in phosphotyrosine binding [1,6,7]. Other residues thought to be involved in phosphotyrosine binding include an arginine near the N-terminal and a histidine near the C-terminal end [1,6].

In this paper we extend these studies of residue conservation by applying the techniques of multiple sequence alignment, secondary structure prediction and conservation analysis to 67 SH2 domain sequences known to bind phosphotyrosine containing proteins. These methods have been previously used to identify accurately the secondary structure and core residues in the annexin domain family [8]. In the present study we highlight previously observed conservation patterns in the SH2 domains, identify residues likely to be buried in the hydrophobic core of the domain, predict the most likely regions of secondary structure (α-helix and β-sheet) and suggest residues likely to be involved in phosphate binding in the light of general studies of phosphate–protein interactions [9]. Our analyses suggest regions of the SH2 domains that are likely to be involved in phosphotyrosine binding and provide tertiary structural interpretations of recent site-directed mutagenesis experiments [6].

Correspondence address: G.J. Barton, University of Oxford, Laboratory of Molecular Biophysics, The Rex Richards Building, South Parks Road, Oxford OX1 3QU, UK. Fax: (44) (865) 510 454.

2. MATERIALS AND METHODS

SH2 domain sequences were obtained from the PIR [10] version 30 databank by scanning with the SH2 domain of human src using the rigorous Smith and Waterman [11] algorithm. Obvious mismatches were removed, and sequences were only included in our analysis if they (or a close relative) had been shown previously to bind phosphotyrosine containing proteins. A total of 60 sequences were extracted from the databank. Seven additional domains (Bovine p85 α and β N- and C-terminal SH2 domains [12]; human phosphotyrosine phosphatase (PTP) 1C N- and C-terminal SH2 domains [3]; avian tensin SH2 domain [5]) were obtained from the literature.

The alignment of 67 SH2 domains shown in Fig. 1 was generated by following a dendrogram derived from pairwise alignment scores [13]. Poorly conserved regions within the alignment were then adjusted by eye to place a single gap between the highly conserved regions. In the following discussion, positions within the alignment are referred to by the name of the most common amino acid occurring and its location (e.g. arginine-39 refers to the highly conserved arginine within the human *src* FLVRES sequence).

Secondary structure predictions were obtained by combining the results of five predictive algorithms applied independently to each aligned sequence. Helix and strand predictions were performed by combining the methods of Lim [14], Chou and Fasman [15] and Robson [16], whilst the methods of Rose [17] and Wilmot and Thornton [18] were used to predict turns. The combination of methods was performed as described in Banga et al. [19]. Where ambiguities exist, all predicted structural types are shown.

3. RESULTS AND DISCUSSION

3.1. *Secondary structure prediction and structural implications of residue conservation*

In a multiple alignment, regions found to contain a large number of gaps, or which are varied in composition across the family of sequences can be predicted as loop or non-core secondary structure regions with a high degree of confidence [20–22]. On this basis we define seven regions of probable secondary structure between such loops (labelled A through G in Fig. 1). The prediction of turns by the methods of Rose [17] and Wilmot and Thornton [18] reinforces the assignment of regions A–G as secondary structures separated by loops.

In order to satisfy thermodynamic requirements, hydrophobic residues (for example, leucine, valine or methionine) are most often buried within the hydrophobic protein core (e.g. see [23] and refs. therein). Accordingly, a position exhibiting conservation of hydrophobic character across all family members implies it is important to the core of the structure. We therefore suggest that positions 3, 4, 5, 8, 22, 36, 37, 38, 49, 51, 53, 68, 71, 73, 84, 86, 93, 96, 99, 100, 102 and 103 are likely to be buried in the core of the native SH2 three-dimensional structure.

Although current secondary structure prediction techniques achieve only about 60% accuracy when applied to a single sequence, prediction accuracy can be improved by applying several complementary techniques to an aligned family of sequences. In addition, conservation of hydrophobic residues across the family of proteins can be used to reinforce the results of secondary structure prediction where characteristic patterns of conservation for α-helix and β-strand are observed. The result of applying these principles to the SH2 domain alignment are summarised in Table I.

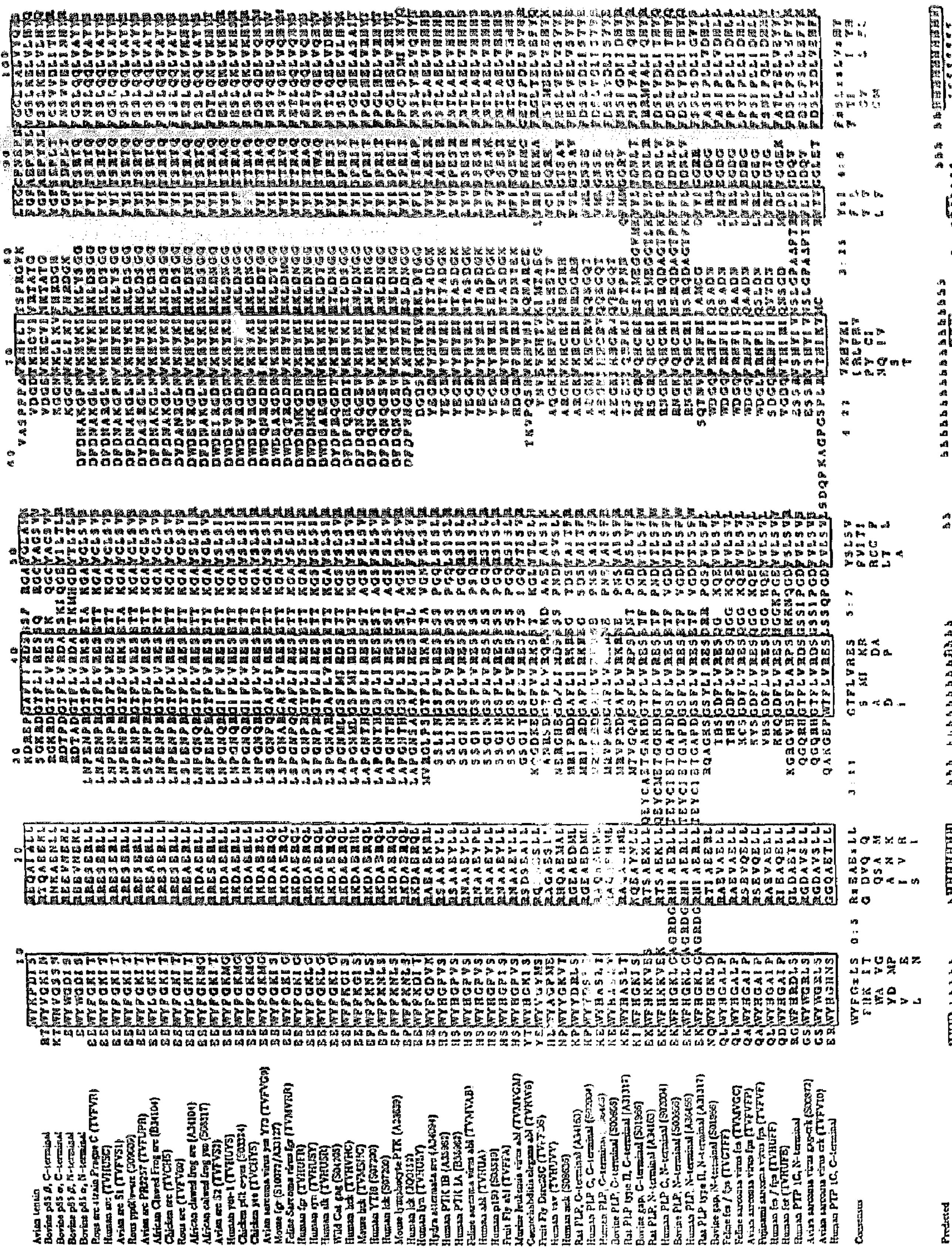
Between positions 96 and 103, region G exhibits a striking pattern of hydrophobic conservation at i (= 96), $i + 3$, $i + 4$ and $i + 7$; were this to be an α-helix, these residues would all lie on one side. The pattern is often seen for a helix which packs against the core of the protein and has been observed to interact with a single hydrophobic residue, perhaps on a sheet [24].

In a strand or extended structure, the protein backbone adopts a conformation where the sidechains of sequential residues point in opposite directions. Conservation patterns involving alternating hydrophobic and hydrophilic residues are therefore often indicative of surface β-strands. Within those regions predicted as β-strands in our alignment, region D (hydrophobic positions 51, 53; hydrophilic positions 50, 52) and region E (hydrophobic positions 71, 73; hydrophilic positions 70, 72) exhibit this pattern across most of the 67 SH2 domains considered.

Short conserved stretches of hydrophobic residues are often indicative of buried β-strands. Regions A (positions 3, 4, 5), C (positions 36, 37, 38) and F (positions 84, 85, 86) show this pattern. Within region A this observation disagrees with the results of the combined secondary structure prediction. However, the tolerance of helix breaking residues such as glycine at position 6 and proline at positions 7 and 9 are suggestive of a G_β β-bulge [25], and hence support the prediction of β-strand within region A. In region C, there is a conserved hydrophobic at position 37 and a small residue conserved at position 35. A small sidechain at position $i \pm 2$ from the central hydrophobic residue in a strand often accommodates the packing of the hydrophobic face of an α-helix against the sheet [24].

→

Fig. 1. Multiple alignment of 67 SH2 domains. Sequences are identified by their name and NBRF-PIR databank code (in parentheses) where available. The last digit in the number above the alignment shows alignment position. A 15 residue segment in avian tensin (Δ, position 67), and a 16 residue segment in avian sarcoma virus *crk* and *gag-crk* (Δ, position 80) were removed for clarity. Boxed amino acids correspond to regions A–G as defined in the text. The numbers between highly conserved regions show the smallest and largest lengths present within the family. The consensus sequence shows residues (top to bottom) which occur at a given position in order of their frequency. Consensus residues are only reported for positions within regions A–G, and only those amino acids which occur more than once are reported. If more than six amino acids occur more than once at a position, the position is considered highly variable and denoted by *x* in the consensus line. Predicted secondary structure is described as helix (h or H), extended (β-strand) (e or E) and turn (t or T). Capital boxed regions highlight the most strongly predicted secondary structures. The summary secondary structure is as defined in Table I.



Only region B (positions 15–22) is lacking in a clear conservation pattern indicative of a specific type of secondary structure.

3.2. *Mode of phosphotyrosine stablisation*

Within protein structures, phosphate containing compounds or phosphate groups are usually stabilised in the following ways [9]: (i) by positively charged sidechains of arginine or less frequently of lysine or histidine; (ii) by a network of hydrogen bonds involving serine and threonine side chains and main chain NH atoms of small amino acids, such as glycine; or (iii) by the positive (N-terminal) end of a helix dipole [26].

Positively charged residues are present in all of the SH2 domain sequences. The only utterly conserved residue is arginine-39 (Fig. 1), which has been shown recently to be critical for phosphate binding [6]. Arginine-15 is conserved in all but three sequences, and histidine-70 is conserved in all but four sequences. If these residues cooperate in phosphate binding, they must be close in the folded native SH2 domain structure. It is important to note that not all members of the family exhibit conservation at all of these positions. The presence or absence of these residues may account for differing affinities of these domains towards substrates. An interesting observation is that bovine GTPase activating protein (gap) C-terminal SH2 domain has a lysine at position 15 and an arginine at position 70 whilst the N-terminal domain has arginine and histidine at these positions. This may account for the lower affinity of the C-terminal domain towards activated receptors [27].

In addition to the residues at positions 15, 39 and 70, there are positively charged residues conserved within some subfamilies. The most striking example is the near total conservation of arginine at positions 20, 74 and 89 within *src*-like SH2 domains. Accordingly, substitution of arginine-20, -74 or -89 with lysine, or a non-charged amino acid may disrupt the specific phosphotyrosine binding properties of the *src* sub-family of SH2 domains.

Although there is insufficient information available to positively identify residues that may form a hydrogen bond network, it is interesting to note that position 34 is glycine, and position 52 is a serine or threonine in all but two sequences.

3.3. *Site-directed mutagenesis*

Site-directed mutagenesis (SDM) is a powerful method for inferring the structural importance of individual residues or small regions within proteins. However, it is important to understand whether an introduced mutation affects activity by altering or removing a functional residue, or rather by causing gross conformational changes to the native three-dimensional structure. Moreover, when the precise function of a particular structure (or family of structures) is poorly understood, it is difficult to interpret macroscopic results (such as transformation studies) in structural terms.

Our findings suggest that only those point mutations found within the conserved regions A through G may be interpreted structurally, since there is substantial variation outside of these regions. Furthermore, our analysis suggests that deletion mutation experiments should be restricted to the loops linking regions A through G. Mutations that delete the predicted secondary structure regions are likely to severely disrupt the native SH2 domain conformation.

Recently, Mayer et al. [6] performed substitution mutations in region C (Fig. 1) of *abl*. They assayed mutants for specific phosphotyrosine binding, and applied 1D-NMR to show that the mutant structures exhibited a similar conformation to the native. Mayer et al. found that mutation of arginine-39 to lysine (R39K) and mutants S41C and S43C had no detectable binding to

Table I

Summary of predicted secondary structure regions within the SH2 family

Region	Length	Predicted conformation	Comments
A	3–9	Strand	Weak helix prediction. Glycine at 6 suggests turn, or $G_i\beta$-bulge. Hydrophobics at 3, 4, 5 possibly form a buried β-strand.
B	15–22	Helix	Weak prediction. No clear hydrophobic pattern.
C	34–41	Strand	Conserved hydrophobics at 36, 37, 38 suggest a buried strand. Polar residues at 39–41 suggest a strand with both sides exposed.
D	49–53	Strand	Conserved hydrophobics at 49, 51, 53 and polar residues at 50, 52 suggest surface strand.
E	68–73	Strand	Conserved hydrophobics at 68, 71 and 73 and polar residues at 70 and 72 suggest half-buried strand.
F	84–86	Strand	Hydrophobic character at 84 and 86 suggests buried strand.
G	93–103	Helix	Conserved hydrophobics at 96, 99, 100, 103 suggest side packing against the protein core.

phosphotyrosine agarose, whilst mutants V38L, E40Q and E42Q bound to phosphotyrosine agarose with the same affinity as the native protein. Mutations R39K, S41C and S43C lie at j (= 39), $j + 2$ and $j + 4$ whilst (V38L, E40Q and E42Q) lie at $j - 1$, $j + 1$ and $j + 3$. These findings are easily interpreted if the conformation of this region is a β-strand as we predict since the two sets of residues would point out on opposite sides of the strand, with only one side of the strand stabilising phosphate.

3.4. *Constraints on the native SH2 domain fold*

Although we are unable to predict the detailed three-dimensional structure of an SH2 domain, the secondary structure prediction, conservation patterns and site-directed mutagenesis data provide clues to the nature of the common fold and phosphotyrosine interaction site of the SH2 domain family.

The conservation at positions 15, 39 (arginine) and 70 (histidine) in nearly all SH2 domains suggests their involvement in phosphotyrosine binding, and that they will be close together in the native fold. Furthermore, the native structure should place positions exhibiting conserved hydrophobic character within the core of the protein.

Several topologies may satisfy these constraints. If we assume that the assignment of secondary structure is broadly correct within regions A–G, then one plausible topology would be a single five-stranded β-sheet with α-helices (regions B and G) packing against each face. However, this arrangement seems unlikely since only two helices are available to bury all the conserved hydrophobics on the single β-sheet. Other plausible topologies include orthogonal and parallel double sheet arrangements, which might better account for the conserved amphipathy of the putative strands.

In any topology, the proximity requirement of positions 15, 39 and 70 should be satisfied. To this end, arginine-39 and histidine-70 may lie on the same face of a β-sheet, or on different sheets facing each other. In either arrangement, arginine-15, predicted to lie at the N-terminal end of a helix could lie near these residues and further stabilise the phosphate group by virtue of both the positively charged side chain and the helix dipole.

3.5. *Putative SH2-like domain of* akt

Bellacosa et al. [4] have reported that a retroviral oncogene, *akt*, encodes a serine-threonine kinase which includes an SH2-like region. Fig. 2 illustrates the alignment of this sequence with the *src* SH2 domain, for which it has a pairwise identity of 20%. The possibility that this region of *akt* may share a similar function to the SH2 domains cannot be ruled out by our analysis. However, the SH2-like domain of *akt* shares only 12 of the 40 consensus residues shown in Fig. 1. *akt* lacks four of the most common SH2 consensus residues (leucine-22; glycine-34; arginine-39; histidine-70). In particular, the change of the highly conserved arginine-39 to lysine is unlikely to be tolerated in light of the SDM results of Mayer et al. [6]. *akt* also lacks two patterns of hydrophobic residue conservation in regions D, and G, and the insertion placed within region A implies further structural dissimilarity to this family of domains.

Our initial search of the PIR databank revealed several other proteins which show sequence similarity to SH2 domains. In particular, a C-terminal region of Tacaribe virus L protein RNA polymerase [28] shares 24/

```
Akt          GWLH    KRGEYIKTW      RPRYFLLKNDGT  FIGYKERPQD
Human src    EWYFGKITRRESERLLLNAENPRGTFLVRESETTKGAYCLSVSD
Match          •  •   •  •              ••            •
Consensus    WYFGxLSRxEAExLL          GTFLVRES        YSLSV
               FHK ITG DVQ Q            S MI KR       FVITI
                WA VG  QSA M            D   DA        RCC F
                YD MP  A N K            A   P         LT  L
                 V  E  I V R            I               A
                 L  N  S  I

Region       AAAAAAABBBBBBBB          CCCCCCCC        DDDDD

Akt          VDQ RESPLNNFSV  AQCQ LMKTER   PRPNTFIIR
Human src    FDNAKGLNVKHYKI RKLDS GGFYITSRTQFNSLQQLVAYYSK
Match                •  •  •     •              •
Consensus            VKHYKI      YxI      FxSLxxLVxHY
                     IRLFRV      F L      Y TI   I YH
                     PY CI       V V        GV   L FL
                     NQ IV       L P        CM
                      S            S
                      T            A

Region                EEEE        FFF      GGGGGGGGGG
```

Fig. 2. Sequence alignment of the SH2-like region of *akt* with human *src*. The alignment was determined by aligning *akt* multiply with all 67 SH2 domains. Dots indicate positions where the *akt* sequence matches the consensus shown in Fig. 1.

40 consensus residues including the sequence TFVLRD within region C. The confirmed members of the SH2 domain family share between 31 and 40 of the consensus residues, suggesting that sequences with fewer than 31 matches may not exhibit SH2-like properties.

Acknowledgements: We thank Professor L.N. Johnson for her encouragement and support. R.B.R. is a Commonwealth Scholar, and a member of Keble College, Oxford. J.B. is a member of St. John's College, Oxford. G.J.B. thanks the Royal Society for support. We are grateful to a referee for detailed and constructive comments on the manuscript.

REFERENCES

[1] Koch, C.A., Anderson, D., Moran, M.F., Ellis, C. and Pawson, T. (1991) Science 252, 668–673.
[2] Heldin, C. (1991) Trends Biochem. Sci., 16, 450–452.
[3] Shen, S., Bastien, L., Posner, B.I. and Chretien, P. (1991) Nature 352, 736–739.
[4] Bellacosa, A., Testa, J.R., Staal, S.P. and Tsichlis, P.N. (1991) Science 254, 274–277.
[5] Davis, S., Lu, M.L., Lo, S.H., Lin, S., Butler, J.A., Druker, B.J., Roberts, T.M., An, Q. and Chen, L.B. (1991) Science 252, 712–715.
[6] Mayer, B.J., Jackson, P.K., Van Etten, R.A. and Baltimore, D. (1992) Mol. Cell. Biol. 12, 609–618.
[7] Hidaka, M., Homma, Y. and Takenawa, T. (1991) Biochem. Biophys. Res. Commun. 180, 1490–1497.
[8] Barton, G.J., Freemont, P.F., Newman, R.H. and Crumpton, M.J. (1991) Eur. J. Biochem. 198, 749–760.
[9] Johnson, L.N. (1984) In: Inclusion Compounds: Vol. 3, Physical Properties and Application. J.L. Atwood, J.E.D. Davies and MacNicol, D.D. (Eds.) pp. 507–569, Academic Press, London.
[10] Barker, W.C., George, D.G. and Hunt, L.T. (1990) Methods Enzymol. 183, 31–49.
[11] Smith, T.F. and Waterman, M.S. (1981) J. Mol. Biol. 147, 195–197.
[12] Otsu, M., Hiles, I., Gout, I., Fry, M.J., Ruiz-Larrea, F., Panayotou, G., Thompson, A., Dhand, R., Hsuan, J., Totty, N., Smith, A.D., Morgan, S.J., Courtneidge, S.A., Parker, P.J. and Waterfield, M.D. (1991) Cell 65, 91–104.
[13] Barton, G.J. (1990) Methods Enzymol. 183, 403–428.
[14] Lim, V.I. (1974) Mol. Biol. 88, 873.
[15] Chou, P.Y. and Fasman, G.D. (1978) Adv. Enzymol. 47, 45–148.
[16] Garnier, J., Osguthorpe, D.J. and Robson, B. (1978) J. Mol. Biol. 120, 97–120.
[17] Rose, G.D. (1978) Nature 272, 586–590.
[18] Wilmot, A.C.M. and Thornton, J.M. (1988) J. Mol. Biol. 203, 221–232.
[19] Banga, J.P., Mahadevan, D., Barton, G.J., Sutton, B.J., Saldanha, J.W., Odell, E. and McGregor, A.M. (1990) FEBS Lett. 266, 133–141.
[20] Crawford, I.P., Niermann, T. and Kirschner, K. (1987) Proteins: Structure, Function and Genetics 1, 118–129.
[21] Benner, S.A. and Gerloff, D. (1990) Adv. Enz. Regul. 31, 121–181.
[22] Thornton, J.M., Flores, T.P., Jones, D.T. and Swindells, M.B. (1991) Nature 354, 105–106.
[23] Schulz, G.E. and Schirmer, R.H. (1979) Springer-Verlag, New York.
[24] Cohen, F.E., Sternberg, M.J.E. and Taylor, W.R. (1982) J. Mol. Biol. 156, 821–862.
[25] Richardson, J., Getzoff, E. and Richardson, D. (1978) Proc. Natl. Acad. Sci. 75, 2574–2578.
[26] Hol, W.G.J., van Duijnen, P.T. and Berendsen, H.J.C. (1978) Nature 273, 443–446.
[27] Anderson, D., Koch, C.A., Grey, L., Ellis, C., Moran, M.F. and Pawson, T. (1990) Science 250, 979–982.
[28] Iapalucci, S., Lopez, R., Rey, O., Lopez, N., Franze-Fernandez, M.T., Cohen, G.N., Lucero, M., Ochoa, A. and Zakin, M.M. (1989) Virology 170, 40–47.