J. Mol. Biol. (1993) **229**, 295–305

Predicted Secondary Structure for the Src Homology 3 Domain

Steven A. Benner, Mark A. Cohen and Dietlind Gerloff

Laboratory for Organic Chemistry
E.T.H. Zurich, CH-8092 Switzerland

(Received 12 October 1992; accepted 28 October 1992)

A *de novo* secondary structure prediction has been prepared for Src homology domain 3, in advance of any crystallographic information concerning any member of this interesting protein family. The prediction can be compared with a crystal structure that will be published in *Nature* on October 29, 1992. The prediction is based on analysis of a multiple alignment of homologous proteins. The patterns of variation and conservation of amino acids across the alignment allow the determination of surface and internal positions, which then allow the assignment of secondary structure. The prediction is quite different both in method and, in this case, result from predictions based on propensities (e.g. Garnier–Osgurthorpe–Robson) of particular amino acids to appear in particular types of secondary structure.

Keywords: protein structure prediction; SH3 domain

Methods based on an analysis of aligned homologous protein sequences have recently been used to make several predictions that, in light of subsequently determined crystal structures, proved to be remarkably accurate (Crawford *et al.*, 1987; Benner, 1989; Bazan, 1990; Benner & Gerloff, 1991; Benner, 1992). It is not yet known whether these methods represent a "major breakthrough" in efforts leading to tools for predicting protein conformation (Lesk & Boswell, 1992), or whether structure prediction still remains "more a matter for soothsayers than scientists" (Hunt & Purton, 1992). The best way to find out, however, is to continue to apply the methods to make predictions (Benner *et al.*, 1992*b*). To be useful, these predictions must be published before crystallographic data are available. This ensures that knowledge of the structure cannot bias the prediction, the predictions (both correct and incorrect) are visible, and the method is placed "at risk". The only obstacle is one of co-ordination. A prediction published years in advance of a crystal structure is uninteresting. A prediction submitted even days after a crystal structure appears is useless.

We were fortunate therefore that A. Musacchio and colleagues contacted us a few weeks ago to challenge us to predict the conformation of the Src homology domain 3 (SH3) (Musacchio *et al.*, 1992*a*) using the method developed in Zurich (Benner, 1989). These workers had just solved the crystal structure of a member of this protein family, and the manuscript describing the structure had just been accepted by *Nature*. Although there was not enough time to process a manuscript describing a prediction before the crystallographic work appeared (on October 29, 1992), the editors of *Nature* graciously agreed to publish a scientific correspondence (Benner *et al.*, 1992*a*) noting that a predicted secondary structure of the protein had been prepared and published elsewhere.

This note reports the predicted secondary structure of the SH3 domain. It was prepared with the sequences of a set of homologous SH3 domains as the only input. Thus, this paper offers another opportunity to compare a prediction made *de novo* with a crystal structure.

An outline of the philosophy underlying the prediction method developed at the E.T.H. in Zurich is available (Benner, 1989). Several worked examples have shown how it can be applied in modeling secondary, supersecondary, and tertiary structures (Benner & Gerloff, 1991; Benner *et al.*, 1992*b*). The method requires as input a set of aligned homologous sequences, with the evolutionary relationship clearly specified between these sequences. The alignments and phylogenetic trees used for the prediction reported here were generated by the DARWIN package (Gonnet & Benner, 1991) and modified slightly by hand. The degree of divergence between protein pairs is measured in Point Accepted Mutations (PAM), where 1 PAM represents one accepted mutation per 100 residues (Dayhoff *et al.*, 1978).

The method is based on heuristics that identify surface and interior positions (Benner *et al.*, 1993). These heuristics extract structural information from patterns of conservation and variation within an

0022–2836/93/020295–11 $08.00/0

```
     1        10        20        30        40        50        60        70        78
     |        |         |         |         |         |         |         |         |
a - GGVTTFVALYDYESRTET_____DLSFKKGERLQIVNNT_____EGDWWLAHSLT____TGQTG_YIPSNYVAPSDS
e - GGVTTFVALYDYESRTET_____DLSFKKGERLQIVNNT_____EGDWWLAHSLT____TGQTG_YIPSNYVAPSDS
b - GGVTTFVALYDYESWTET_____DLSFKKGERLQIVNNT_____EGDWWLAHSLT____TGQTG_YIPSNYVAPSDS
c - GGVTTFVALYDYESRTET_____DLSFKKGERLQIVNNT_____EGDWWLAHSLS____TGQTG_YIPSNYVAPSDS
f - GGVTTFVALYDYESRTET_____DLSFKKGERLQIVNNTRKVDVREGDWWLAHSLS____TGQTG_YIPSNYVAPSDS
d - GGVTTFVALYDYESRTET_____DLSFKKGERLQIVNNT_____EGDWWLARSLS____SGQTG_YIPSNYVAPSDS
g - GGVTTFVALYDYESRTET_____DLSFRKGERLQIVNNT_____EGDWWLARSLS____SGQTG_YIPSNYVAPSDS
h - gggTVFVALYDYEARTTD_____DLSFKKGERFQIINNT_____EGDWWEARSIA____TGKTG_YIPSNYVAPADS
i - gGVTVFVALYDYEARTTD_____DLSFKKGERFQIINNT_____EGDWWEARSIA____TGKTG_YIPSNYVAPADS
k - gGVTVFVALYDYEARTTE_____DLSFRKGERFQIINNT_____EGDWWEARSIA____TGKTG_YIPSNYVAPADS
j - gGVTIFVALYDYEARTTE_____DLSFKKGERFQIINNT_____EGDWWEARSIA____TGKNG_YIPSNYVAPADS
l - tGVTLFVALYDYEARTED_____DLSFQKGEKFQILNSS_____EGDWWEARSLT____TGGTG_YIPSNYVAPVDS
m - tGVTLFVALYDYEARTED_____DLSFHKGEKFQILNSS_____EGDWWEARSLT____TGETG_YIPSNYVAPVDS
n - tGVTIFVALYDYEARTGD_____DLTFTKGEKFHILNNT_____EYDWWEARSLS____SGHRG_YVPSNYVAPVDS
o - tGVTLFIALYDYEARTED_____DLTFTKGEKFHILNNT_____EGDWWEARSLS____SGKTG_CIPSNYVAPVDS
p - pGVTIFVALYDYEARISE_____DLSFKKGERLQIINTA_____DGDWWYARSLI____TNSEG_YIPSTYVAPekd
t - eevrfvVALFDYAAVNDR_____DLQVLKGEKLQVLRST_____GDWWLARSLV____TGREG_YVPSNFVAPVET
B - vlkrVVVSLYDYKSRDES_____DLSFMKGDRMEVIDDT_____ESDWWRVVNLT____TRQEG_LIPLNFVAeers
q - sedIIVVALYDYEAIHHE_____DLSFQKGDQMVVLEES_____GEWWKARSLA____TRKEG_YIPSNYVARVDS
r - sedTIVVALYDYEAIHRE_____DLSFQKGDQMVVLEEA_____GEWWKARSLA____TKKEG_YIPSNYVARVNS
s - eqgdIVVALYPYDGIHPD_____DLSFKKGEKMKVLEEH_____GEWWKAKSLL____TKKEG_FIPSNYVAKLNT
u - lqdNLVIALHSYEPSHDG_____DLGFEKGEQLRILEQS_____GEWWKAQSLT____TGQEG_FIPFNFVAKANS
v - lqdNLVIALHSYEPSHDG_____DLGFEKGEQLRILEQS_____GEWWKAQS_T____TGQEG_FIPFNFVAKANS
A - lqdKLVVALYDYEPTHDG_____DLGLKQGEKLRVLEES_____GEWWRAQSLT____TGQEG_LIPHNFVAMVNS
z - ddpqLFVALYDFQAGGEN_____QLSLKKGEQVRILSYNKS_____GEWCEAHS_____SGNVG_WVPSNYVTPLNS
w - npdnLFVALYDFVASGDN_____TLSITKGEKLRVLGYNHN_____GEWCEAQT_____KNGQG_WVPSNYITPVNS
x - npdnLFVALYDFVASGDN_____TLSITKGEKLRVLGYNHN_____GEWCEAQT_____KNGQG_WVPSNYITPVNS
y - npdnLFVALYDFVASGDN_____TLSITKGEKLRVLGYNHN_____GEWCEAQT_____KNGQG_WVPSNYITPVNS
H -maeevVVVAKFDYVAQQEQ_____ELDIKKNERLWLLDDSKS_____WWRVRN_S____MNKTG_FVPSNYVERKNS
#O - tgkelvlALYDYQ_____EKSPREVTMKKGDILTLLNST_____NKDWWK____VEVNDR__QG_FVPAAYVKKLDp
P - tgkecvvALYDYT_____EKSPREVSMKKGDVLTLLNSN_____NKDWWK____VEVNDR__QG_FVPAAYIKKIDA
M - vetkfvqALFDFN_____PQESGELAFKRGDVITLINKD_____DPNWWEG____QLNNR__RG_IFPSNYVcpyns
R - pgpeqarALYDFA_____AENPDELTFNEGAVVTVINKS_____NPDWWEG____ELNGQ__RG_VFPASYVelipr
Q - pakpqvkALYDYD_____AQTGDELTFKEGDTIIVHQKD_____PAGWWEG____ELNGK__RG_WVPANYVqdi
S - kyfgtakARYDFC_____ARDRSELSLKEGDIIKILNKKG_____QQGWWRG____EIYGR__VG_WFPANYVEEdys
L -    meavAEHDFQ_____AGSPDELSFKRGNTLKVLNK_DE____DPHWYKA____ELDGN__EG_FIPSNYIrmtec
U - rpigivvAAYDFNYPIKKDSSSQLLSVQQGETIYILNK_NS____S_GWWDGLVIDDSNGKVNRG_WFPQNFgrplrd
T -      vqALYPFSSSNDE_____ELNFEKGDVMDVIEKPEN____DPEWWK___crking__mvg_lvpknyvtvmqn
V -      vrALFDYDPNRDDGLPSRGLPFKHGDILHVTNASDD_____EWWQarrvlgdnedeqig_ivpskrrwerkm
N -      vrALFDFKGN_DDG____DLPFKKGDILKIRDKPEE_____QWWN___aedmdg__krg_mipvpyvekcrp
I - dlnmpayvkFNYMAERED_____ELSLIKGTKVI_____VMEKCSDGWWRG____SY_N_GQVG_WFPSNYVvteegd
C - mpqrtVKALYDYKAKRSD_____ELSFCRGALIH_____NVSKEPGGWWKG____DY_G_TRIQQYFPSNYVEDIst
D - TFKCAVKALFDYKAQRED_____ELTFTKSAIIQ_____NVEKQDGGWWRG____DY_G_GKKQLWFPSNYVEEMin
E - TFKCAVKALFDYKAQRED_____ELTFTKSAIIQ_____NVEKQEGGWWRG____DY_G_GKKQLWFPSNYVEEMvs
F - TFKCAVKALFDYKAQRED_____ELTFIKSAIIQ_____NVEKQEGGWWRG____DY_G_GKKQLWFPSNYVEEMvn
J - kenpwatAEYDYDAAEDN_____ELTFVENDKII_____NIEFVDDDWWLG____ELKD_GSKGL_FPSNYVSlgn
K - elgitaiALYDYQAAGDD_____EISFDPDDIIT_____NIEMIDDGWWRG____VCK__GRYGL_FPANYVElqr
G - algisavALYDYQGEGSD_____ELSFDPDDVIT_____DIEMVDEGWWRG____RCH__GHFGL_FPANYVKlle
                                                          *
                                                 xxxxxxxxxxx
```

Figure 1. The master alignment. Except where noted, all the numbering in this paper refers to this alignment. All alignments were generated by the DARWIN package (Gonnet & Benner, 1991). The multiple alignment is generated using all the shown sequences, subsets of sequences used to generate sub-alignments may differ in detail, especially with respect to placement of gaps. Lower case letters are positions that do not significantly align to the longest sequence. An asterisk (*) indicates a position conserved across the whole alignment. Deletions are indicated by an underscore (_). The sequences are a, ASV v-SRC; b, RSV v-SRC; c, H c-SRC-1; d, Xl c-SRC-1; e, C c-SRC; f, M n-SRC; g, Xl c-SRC-2; h, ASV v-YES; i, C c-YES; j, H c-YES-1; k, Xl c-YES; l, Xl c-FYN; m, H c-FYN; n, M c-FGR; o, H c-FGR; p, Ha STK; q, H HCK; r, M HCK; s, H LYN; t, M BLK; u, M LSK-T; v, H LCK; w, FSV v-ABL; x, H c-ABL; y, M c-ABL; z, Dm ABL-1; A, C c-TKL; B, Dm SRC-1; C, H PLC; D, R PLC-II; E, B PLC-II; F, H PLC1; G, H HS1; H, H NCK/1; I, H NCK/2; J, Y ABP1; K, C P80/85; L, Ce sem-5/1; M, Ce sem-5/2; N, ASV GAGCRK; O, C Spc-a; P, Dm Spc-a; Q, Ac MIL-B; R, Ac MIL-C; S, H VAV; T, H NCK/3; U, Y CDC25; V, Dm DLG (for the key, refer to Musacchio *et al.*, 1992*a*). The sequence for which the crystal structure has been solved is O, and is indicated by a #. In the light of the structure prediction, some of the gaps have to be rearranged, and this is particularly important in the region indicated by the 11 x at the bottom of the Figure. The deletion in the sequences C to K must be shifted 6 positions to the right, else it occurs in the centre of a helix (see the text).

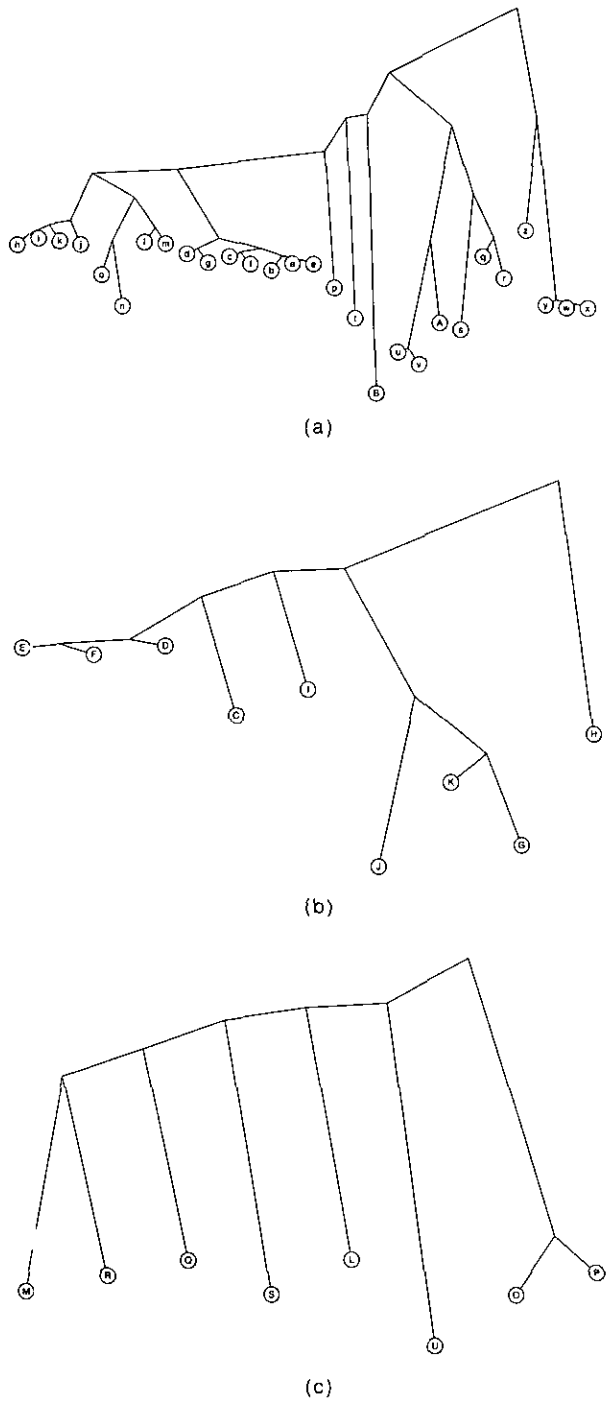


Figure 2. Phylogenetic trees for the 3 sub-families (a) SH3 A, (b) SH3 B and (c) SH3 C1. The trees are generated by the DARWIN package as the most probable phylogenetic tree by a least-squares fitting of the PAM distance data. The key is the same as in Fig. 1.

aligned set of homologous sequences; the strength of the assignments depends on the type and distribution of amino acids. The alignment is then "parsed", divided into short segments whose secondary structure can be analyzed independently. Heuristics assign parses from the occurrence of deletions, patterns of conservation and variation in "structure disrupting" residues (e.g. Pro, Gly, Asn, Asp and Ser), strings of these residues within single sequences, and patterns in surface assignments (Cohen *et al.*, 1983). Secondary structural assignments are then made from patterns of surface and interior assignments; for example, amphiphilic surface helices are assigned when 3·6 residue periodicity of surface and interior assignments is observed. The predicted secondary structural units are then assembled using distance constraints implied by active site assignments, covariation analysis, and adaptive variation.

These methods were used without substantial modification in the prediction presented here. Because the master alignment (Fig. 1) for the SH3 family represents a substantial amount of divergent evolution, it is better described as three (and possibly four) subfamilies. The alignment containing all four of the subfamilies is poor; indeed homology between the subfamilies is not indisputably established by the sequence data alone. As poor alignments are a principal cause of error in *de novo* predictions using this method, secondary structure predictions were made separately for each subfamily. Comparing these yields a consensus secondary structure for the overall SH3 family. Some differences in the details of the predicted secondary structure is expected under these circumstances. Further, the protein whose crystal structure has been solved (Musacchio *et al.*, 1992*b*) is not in the subfamily that has the largest amount of data and where the prediction is the most reliable. This should be kept in mind when comparing the prediction here with the crystal structure reported by Musacchio *et al.* (1992*b*).

The SH3 A subfamily

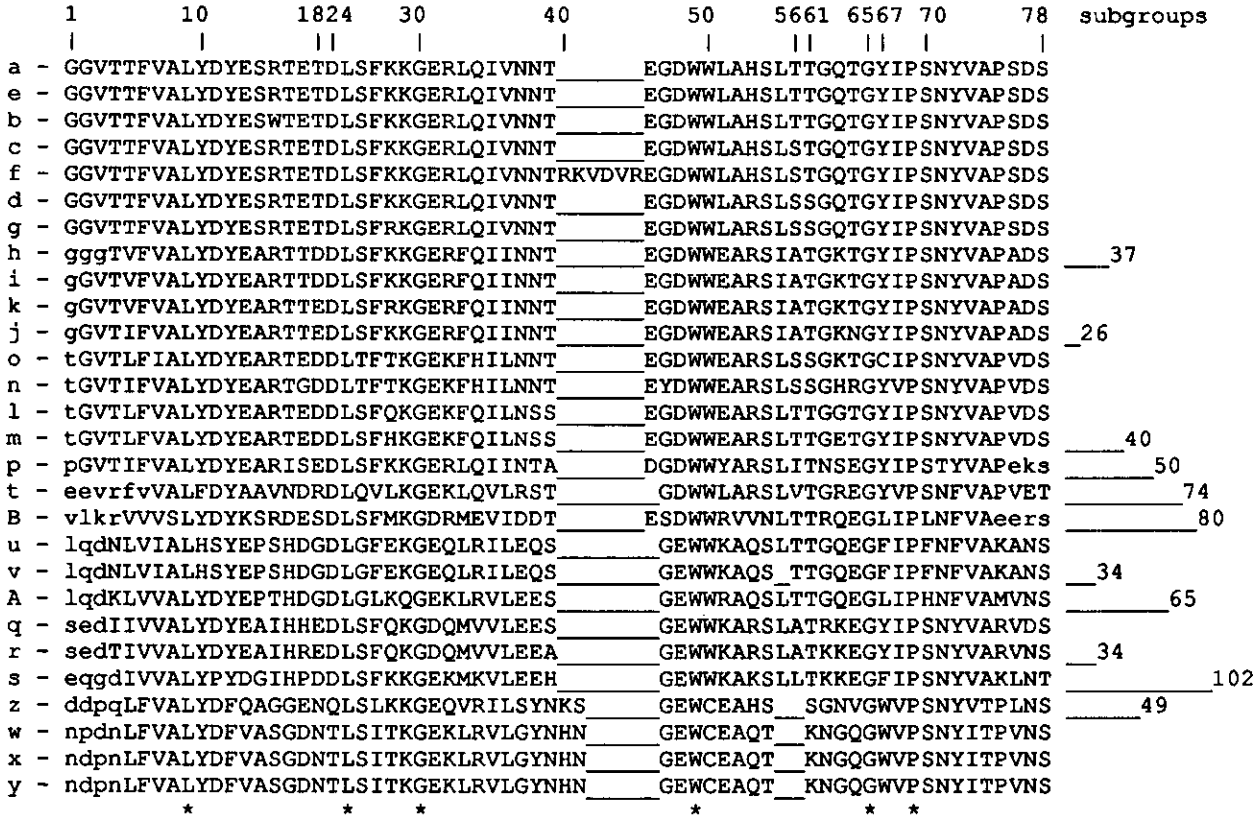
The SH3 A subfamily contains 28 sequences (Fig. 3(a)) arranged on a tree (Fig. 2(a)) with an overall PAM width of 102. The subfamily is divided into three segments by two primary parses (segment I, 05–39; segment II, 47–54; and segment III, 61–78). These segments can be subdivided by confirmed secondary parses as follows.

Segment I (05–39): Weaker parses occur at positions 14 (dipeptide parse PS, confirmed by a deletion in subfamily C1), 18–24 (tripeptide parse GDD, confirmed by a deletion in subfamily C2), 30 (all positions conserved (APC) G, confirmed by a conserved G in other subfamilies), and 37–38 (dipeptide parse NN, confirmed in subfamily C1).

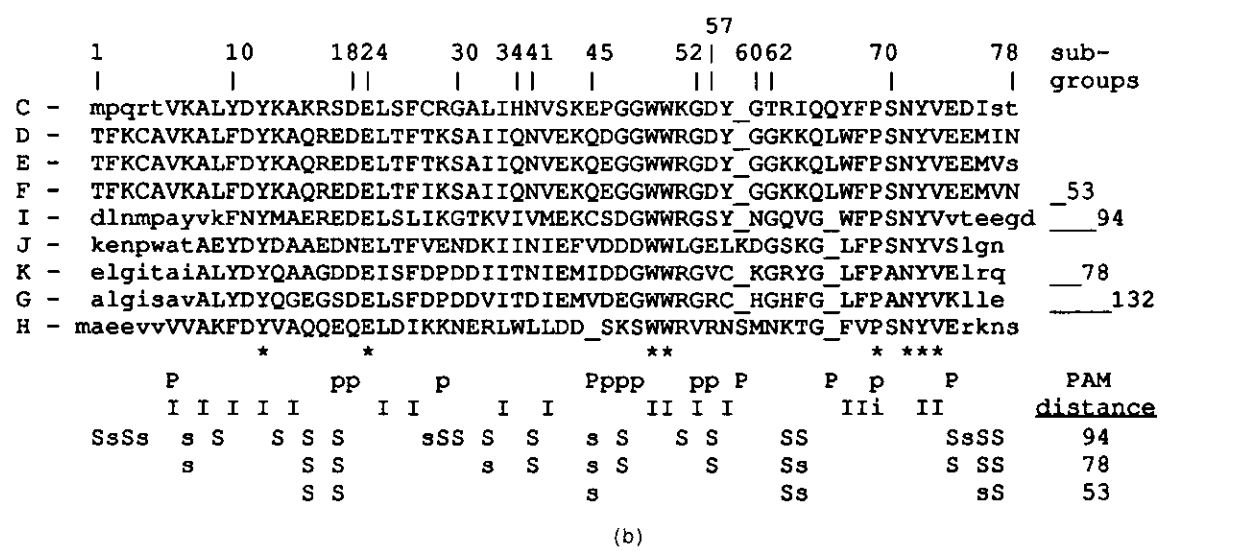
Segment II (47–54): A weaker parse occurs at positions 47–48 (dipeptide parse GD, confirmed in subfamilies C1 and C2).

Segment III (61–78): Weaker parses occur at positions 65 (APC G, confirmed in subfamily C), 69 (APC P, APC across the entire alignment, confirmed by tripeptide parse PSN), and 75 (tetrapeptide parse PSDS, no confirmation).

These secondary parses divide the alignment for subfamily SH3 A into the following seven segments: segment 1, 05–17; segment 2, 25–30; segment 3, 30–37; segment 4, 49–54; segment 5, 61–65; segment


1 10 1824 30 40 50 5661 6567 70 78 subgroups
a - GGVTTFVALYDYESRTETDLSFKKGERLQIVNNT_____EGDWWLAHSLTTGQTGYIPSNYVAPSDS
e - GGVTTFVALYDYESRTETDLSFKKGERLQIVNNT_____EGDWWLAHSLTTGQTGYIPSNYVAPSDS
b - GGVTTFVALYDYESWTETDLSFKKGERLQIVNNT_____EGDWWLAHSLTTGQTGYIPSNYVAPSDS
c - GGVTTFVALYDYESRTETDLSFKKGERLQIVNNT_____EGDWWLAHSLSTGQTGYIPSNYVAPSDS
f - GGVTTFVALYDYESRTETDLSFKKGERLQIVNNTRKVDVREGDWWLAHSLSTGQTGYIPSNYVAPSDS
d - GGVTTFVALYDYESRTETDLSFKKGERLQIVNNT_____EGDWWLARSLSSGQTGYIPSNYVAPSDS
g - GGVTTFVALYDYESRTETDLSFRKGERLQIVNNT_____EGDWWLARSLSSGQTGYIPSNYVAPSDS
h - gggTVFVALYDYEARTTDDLSFKKGERFQIINNT_____EGDWWEARSIATGKTGYIPSNYVAPADS ___37
i - gGVTVFVALYDYEARTTDDLSFKKGERFQIINNT_____EGDWWEARSIATGKTGYIPSNYVAPADS
k - gGVTVFVALYDYEARTTEDLSFRKGERFQIINNT_____EGDWWEARSIATGKTGYIPSNYVAPADS
j - gGVTIFVALYDYEARTTEDLSFKKGERFQIINNT_____EGDWWEARSIATGKNGYIPSNYVAPADS _26
o - tGVTLFIALYDYEARTEDDLTFTKGEKFHILNNT_____EGDWWEARSLSSGKTGCIPSNYVAPVDS
n - tGVTIFVALYDYEARTGDDLTFTKGEKFHILNNT_____EYDWWEARSLSSGHRGYVPSNYVAPVDS
l - tGVTLFVALYDYEARTEDDLSFQKGEKFQILNSS_____EGDWWEARSLTTGGTGYIPSNYVAPVDS
m - tGVTLFVALYDYEARTEDDLSFHKGEKFQILNSS_____EGDWWEARSLTTGETGYIPSNYVAPVDS _____40
p - pGVTIFVALYDYEARISEDLSFKKGERLQIINTA_____DGDWWYARSLITNSEGYIPSTYVAPeks _____50
t - eevrfvVALFDYAAVNDRDLQVLKGEKLQVLRST_____GDWWLARSLVTGREGYVPSNFVAPVET _____74
B - vlkrVVVSLYDYKSRDESDLSFMKGDRMEVIDDT_____ESDWWRVVNLTTRQEGLIPLNFVAeers _____80
u - lqdNLVIALHSYEPSHDGDLGFEKGEQLRILEQS_____GEWWKAQSLTTGQEGFIPFNFVAKANS
v - lqdNLVIALHSYEPSHDGDLGFEKGEQLRILEQS_____GEWWKAQS_TTGQEGFIPFNFVAKANS __34
A - lqdKLVVALYDYEPTHDGDLGLKQGEKLRVLEES_____GEWWRAQSLTTGQEGLIPHNFVAMVNS _____65
q - sedIIVVALYDYEAIHHEDLSFQKGDQMVVLEES_____GEWWKARSLATRKEGYIPSNYVARVDS
r - sedTIVVALYDYEAIHREDLSFQKGDQMVVLEEA_____GEWWKARSLATKKEGYIPSNYVARVNS __34
s - eqgdIVVALYPYDGIHPDDLSFKKGEKMKVLEEH_____GEWWKAKSLLTKKEGFIPSNYVAKLNT _____102
z - ddpqLFVALYDFQAGGENQLSLKKGEQVRILSYNKS_____GEWCEAHS__SGNVGWVPSNYVTPLNS _____49
w - npdnLFVALYDFVASGDNTLSITKGEKLRVLGYNHN_____GEWCEAQT__KNGQGWVPSNYITPVNS
x - ndpnLFVALYDFVASGDNTLSITKGEKLRVLGYNHN_____GEWCEAQT__KNGQGWVPSNYITPVNS
y - ndpnLFVALYDFVASGDNTLSITKGEKLRVLGYNHN_____GEWCEAQT__KNGQGWVPSNYITPVNS
PAM distance
SSSS 102
SSSS 80
SSSS 74
SSSS 65
SSSS 50
sSSS 49
sSSS 40
sSSS 37
34
sS 26

(a)


1 10 1824 30 3441 45 52 57 6062 70 78 sub-groups
C - mpqrtVKALYDYKAKRSDELSFCRGALIHNVSKEPGGWWKGDY_GTRIQQYFPSNYVEDIst
D - TFKCAVKALFDYKAQREDELTFTKSAIIQNVEKQDGGWWRGDY_GGKKQLWFPSNYVEEMIN
E - TFKCAVKALFDYKAQREDELTFTKSAIIQNVEKQEGGWWRGDY_GGKKQLWFPSNYVEEMVs
F - TFKCAVKALFDYKAQREDELTFIKSAIIQNVEKQEGGWWRGDY_GGKKQLWFPSNYVEEMVN _53
I - dlnmpayvkFNYMAEREDELSLIKGTKVIVMEKCSDGWWRGSY_NGQVG_WFPSNYVvteegd ___94
J - kenpwatAEYDYDAAEDNELTFVENDKIINIEFVDDDWWLGELKDGSKG_LFPSNYVSlgn
K - elgitaiALYDYQAAGDDEISFDPDDIITNIEMIDDGWWRGVC_KGRYG_LFPANYVElrq __78
G - algisavALYDYQGEGSDELSFDPDDVITDIEMVDEGWWRGRC_HGHFG_LFPANYVKlle ____132
H - maeevvVVAKFDYVAQQEQELDIKKNERLWLLDD_SKSWWRVRNSMNKTG_FVPSNYVErkns
PAM distance
SsSs 94
78
53

(b)

Fig. 3.

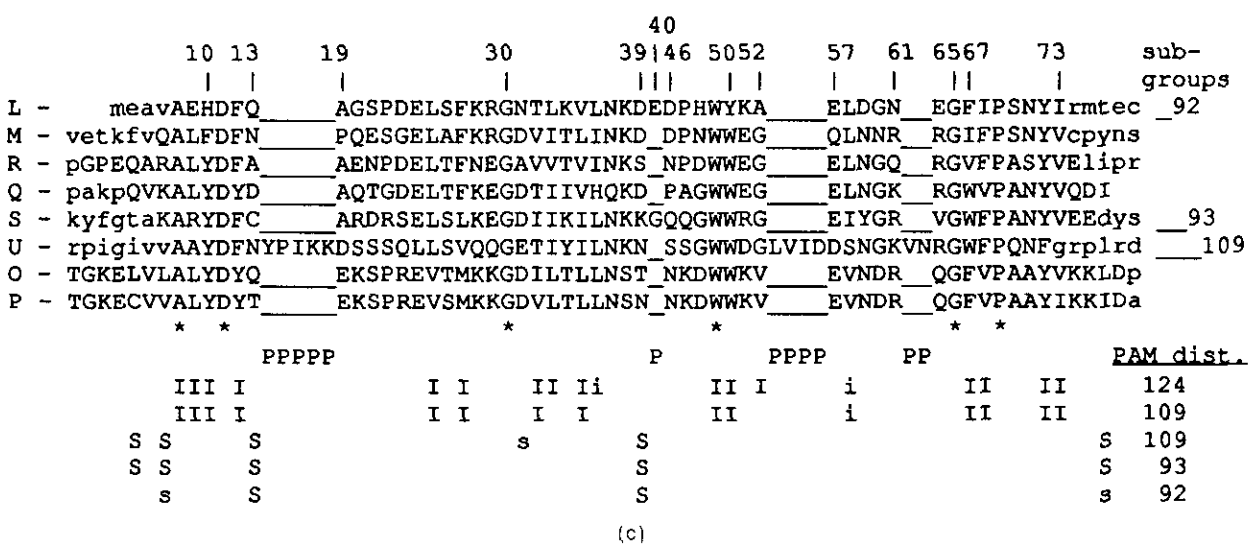


Figure 3. Multiple alignments of the sub-families (a) SH3 A, (b) SH3 B and (c), SH3 C1. The numbering refers to the alignment in Fig. 1, which results in some discontinuities where the other subfamilies have inserts. In the lines below the alignment, parses are indicated either as P, strong parse, or p, weak parse. Surface assignments are given as either S, strong assignments, or s, weak assignments; and interior assignments are given as I, strong assignments or i, weak assignments. The PAM distance for which the surface assignments are made is indicated in the right-hand column. The assignments are made by examining the variability of the amino acids at each position within subgroups defined at different PAM distances as described elsewhere (Benner & Gerloff, 1991). The subgroups and the PAM distance are indicated on the right of the alignment. The subgroups can be delineated by extension of the lines indicated at the different PAM distances on the right of the alignment through the alignment. The number of subgroups always increases as the PAM distance decreases, until at very low PAM distance each sequence is in its own individual subgroup (not shown). The key is the same as in Fig. 1.

6, 67–69; and segment 7, 70–78. The SH3 A subfamily is then divided into clusters of subgroups at different maximum PAM widths (Fig. 3(a)). These are used to assign surface and interior positions, and then secondary structure (Table 1). Unfortunately, the overall divergence within this subfamily is rather small (only 102 PAM units). This means that heuristics assigning interior positions are on the whole less reliable, although they identify a higher fraction of the interior positions. Conversely, the surface heuristics are stronger, although they identify a smaller fraction of the surface positions. This subfamily has a well-branched tree, and a variety of surface predictions can be made at different PAM distances, each one confirming and supporting the others.

Segment 1: positions 5 to 17

At PAM 102 a string of interior assignments are made at positions 5 to 10. Interior assignments are made at positions 12 and 14 and surface assignments at positions 13, 15 (strong) and 11 (weak). At PAM 65, surface assignments are made at positions 11 13, 15 and 17. Six internal amino acid residues between positions 5 and 10 is too short a segment to form an internal helix. This segment is assigned as a β-strand (strong assignment). In general, short internal stretches of sequences are assigned as β-strands (Cohen *et al.*, 1982), this assignment holds unless there is good evidence to the contrary.

The principal problem in assigning secondary structure to this segment is to decide whether one β should be assigned to the first part (for example, positions 5 to 10) and a second assigned to the second part (for example, 12 to 15), or whether the two segments might represent a single secondary structural unit. The PS dipeptide parse indicates a break in secondary structure with approximately 75% accuracy, so the two-segment assignment was chosen.

Segment 2: positions 25 to 30

Interior assignments are made at positions 25 and 27 (PAM 102), and a surface assignment at position 28 at PAM 80, weak surface assignments based on less reliable heuristics can be made at positions 24 and 26. If this segment were extended past the weak parse at position 30, its length (13 positions) would suggest a single α-helix. No 3·6 residue pattern of periodicity is observed across the entire 13 positions, however. The two segments are therefore considered separately. In the first segment, the alternating pattern of periodicity from positions 24 to 28 (polar-inside-surface-inside-surface) gives a β assignment of moderate strength.

Segment 3: positions 30 to 37

At PAM 102, interior assignments are made at positions 33, 35 and 36, and surface assignments at positions 32 and 34. A β assignment is supported by the alternate periodicity from positions 32 to 35

Table 1

A summary of the secondary structure prediction for the three sub-families of the SH3-homology domain

	SH3 subfamily A				SH3 subfamily B				SH3 subfamily C			
Assignment	Minimum	Maximum	Preferred	Length	Minimum	Maximum	Preferred	Length	Minimum	Maximum	Preferred	Length
Coil/turn†	01–03	01–05	01–04	3	01–04	01–07	01–05	5	01–03	01–07	01–06	6
β1a	06–09	04–11	05–10	6	08–13	05–15	06–14	9	08–11	04–13	07–13	7
β1b	12–15	10–18	12–15	4	Combined with β-strand above				Combined with β-strand above			
Coil/turn	—	16–18	16–18	3	16–18	14–24	15–18	4	14–18	13–18	14–18	5
β2	24–28	24–28	24–28	5	25–28	24–28	24–28	5	24–28	24–28	24–28	5
Coil/turn	29	29	29	1	—	29–31	29	1	29	29–30	29–30	2
α1	33–38	30–39	32–39	8	32–43	29–44	30–44‡	9	31–37	30–39	31–38	8
Coil/turn	40–46	39–49	40–47	2§	45–46	44–49	45–47	3	40	38–48	39–48	5§
β3	50–53	47–54	48–53	6	50–58	47–58	48–58	6	49–50	48–52	49–52	4
Coil/turn	55–64	54–65	54–65	8§	59–65	59–66	59–66	7§	53–63	51–66	53–64	12
β4	67–68	65–69	67–69	3	67–68	66–69	67–68	2	67–68	64–69	65–68	3
Coil/turn	70	69–71	70	1	—	69–71	69	1	—	69–71	69–70	2
β5	72–74	71–74	71–74	4	72–73	69–74	70–73	4	72–73	70–74	71–73	3

Minimum and Maximum indicate the minimum and maximum extents of the secondary structural elements found in the different sequences of the alignment. Preferred indicates the region of the secondary structural element that is most clearly defined; this is the core region found in most sequences in the alignment. The Length column refers to the preferred length.

† This coil marks the beginning of the domain.

‡ Based on this assignment, the multiple alignment was adjusted; see the text.

§ Note discontinuous numbering in the alignment.

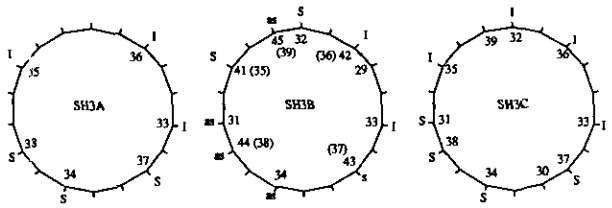


Figure 4. The central helical segment of the 3 subfamilies. I, interior assignment, S, surface assignment, s, weaker surface assignment, as, amphiphilic split. Note how the orientation of the helix in family SH3B has altered slightly with respect to the other two. The sequences of the SH3B family are highly divergent from those of the other families in this region. This may be a consequence or cause of this change in helix orientation.

(surface-interior-surface-interior). This pattern is broken by the interior assignment at position 36. A helical assignment is supported by the 3·6 residue periodicity from positions 33 to 36 (interior-surface-interior-interior). This pattern is broken by the surface assignment at position 32. However, surface residues are often found on the interior arc of a helix projection when they are at the end of a helix. The secondary structure assigned depends on assignments made at positions 31 and 37 to 39. All are weak surface at this PAM distance (single variable subgroup or polar but uncharged variable residues).

At PAM 80, positions 31 and 38 are assigned to the surface (2 variable subgroups). Position 37 has only a single variable subgroup containing a charged residue, for the strongest indications of a surface position two or more variable subgroups containing charged residues are required. Position 39 remains hydrogen-bonding variable (typically 70 % interior). At PAM 65, positions 31 and 37 both have a single variable subgroup. The surface prediction for position 38 is very strong with two variable subgroups. At PAM 37, position 31 remains assigned to the surface, while the surface assignment at position 37 is lost at PAM 50.

Given the two competing assignments, the helix is preferred. Decisive is the fact that position 38, but not position 39, is assigned to the surface by heuristics operating on subgroups with lower maximum PAM widths. Thus, the pattern of amphiphilicity extends from positions 33 to 39 (7 residues), the minimum for a 2-turn helix. Ignoring the surface assignment at position 32, the pattern can be extended back to position 30, where it is broken again at position 29 by another surface assignment within the interior arc. The assignment of a helix has moderate strength and is most reliable from positions 33 to 39 (Fig. 4).

Segment 4: positions 48 to 54

At PAM 102, interior assignments are made at positions 49, 50 and 52. Surface assignments are made at positions 53 and 54. The surface assignment at position 53 breaks the 3·6 residue periodicity of surface and interior positions. At PAM 80, surface assignments are made at positions 51, 53 and 54. The first two remain at PAM 65. The assignment of a β-strand (moderate to strong) to this segment is suggested by the alternating pattern in the assignments between positions 50 and 53 (interior-surface-interior-surface). The conservation at positions 48 and 49 suggests that the strand might extend to the left to include these positions.

Segment 5: positions 61 to 65

Four consecutive surface assignments at PAM 102 indicate a coil/turn.

Segment 6: positions 67 to 69

At PAM 102, three consecutive interior positions (67 to 69) are assigned. No surface assignments are made at PAM 80. Short segments of this nature are canonically assigned β structures, with moderate reliability (Cohen *et al.*, 1982), attention must be directed to the subsequent segment in a search for larger structural continuity in the event that the APC Pro at position 69 misassigns a parse.

Segment 7: positions 70 to 78

At PAM 102, interior assignments are made at positions 72 to 74. The first of three consecutive weak surface assignments is made at position 75. These are stronger at PAM 80. There is a possible surface assignment at position 71, which remains at PAM 65.

The alignment ceases to be significant for all proteins in this subfamily after position 74. This, together with the consecutive surface positions assigned starting at position 75, indicates the end of the domain. Three consecutive interior residues suggests an interior β-strand (moderate reliability). The presence of a parsing tripeptide (PSN) suggests that there is a break between the two assigned β-strands. This is supported by the lack of any 3·6 residue periodicity on connection of this segment with the preceding segment (positions 67 to 69).

The SH3 B subfamily

The SH3 B subfamily contains nine proteins (Fig. 3(b)) arranged on a tree (Fig. 2(b)) with an overall PAM width of 190. The multiple alignment is parsed by primary parsing units into four segments (segment I, 6–44; segment II, 46–58; segment III, 60–65; and segment IV, 67–73). These segments can be subdivided by secondary parses as follows.

Segment I (6–44): Weaker parses at positions 17–18 (dipeptide parse SD and DD, confirmed in other subfamilies) and 29 (tripeptide parse DPD, unconfirmed in other subfamilies).

Segment II (46–58): Weaker parse at position 46–48 (tripeptide parse PPG and DDG, confirmed in subfamily C1 and C2).

Segment III (60–65): Weaker parse at position 62–63 (dipeptide parse GS, confirmed in subfamily C).

Segment IV (67–73): Weaker parse at positions 69–70 (APC P, tripeptide parse PSN, confirmed by other subfamilies).

These secondary parses divide the alignment into the following working segments: segment 1, 06 to 17; segment 2, 24 to 28; segment 3, 31 to 44; segment 4, 49 to 58; segment 5, 62 to 65; segment 6, 67 to 69; and segment 7, 70 to 73.

Surface and interior assignments are made after division of the SH3 B subfamily into clusters of subgroups at different maximum PAM widths as above. Unfortunately, the evolutionary tree is not balanced. Thus, at PAM 132, the subfamily fragments into two subgroups, but one of these contains only a single protein sequence. Heuristics based on a search for concurrent variation can be applied only on the cluster of subgroups at PAM 94. Analysis at lower PAM widths increases the reliability of the surface assignments, but not their number.

Segment 1: positions 6 to 17

At PAM 132, positions 6 to 14 contain a perfectly alternating set of interior assignments at the even-numbered positions. At PAM 94, the odd-numbered positions 7, 9, 13, 15 and 17, are assigned to the surface. Assignments at positions 7, 15 and 17 remain at PAM 53. This provides a very strong β assignment for the segment. The maximum extent is from position 5 to position 18. The minimum extent is from positions 6 to 14. This is a long strand for a small protein, and may be bent in the middle.

Segment 2: positions 24 to 28

Interior assignments at position 25 and 27 are made at PAM 132. At PAM 94, surfaces assignments are made at position 26 and 28. If the parse at position 30 is viewed as reliable (>90%), this segment is too short to be anything other than a β-strand, a coil, or a single turn of an α-helix. A pattern of alternating assignments at positions 25 to 28 provides a strong β assignment.

Segment 3: positions 31 to 44 (note discontinuity in alignment numbering)

Two strong interior assignments are made at positions 33 and 42 at PAM 132. At PAM 94, surface assignments are made at positions 29, 30, 32 and 41. Surface assignments at positions 32 and 41 are confirmed at PAM 78, where a surface assignment at position 29 is based on a single variable subgroup.

The nine positions in this segment can build a 2-turn α-helix, a rather long β-strand, or a rather long coil. The only two interior anchors for the segment are at positions 33 and 42. These are adjacent on a helical wheel. The surface arc of the wheel includes two amphiphilic splits and a hydrogen-bonding variable. The former are generally on the surface (approx. 80% at these PAM distances). Hydrogen-bonding variable positions are on the surface approximately 40% of the time in such alignments. As there is no reliable 2-residue periodicity, this pattern allows a moderately strong helix assignment to be made (Fig. 4).

Segment 4: positions 49 to 58 (note discontinuity in alignment numbering)

At PAM 132, interior assignments are made at positions 49, 50, 52 and (weakly) 58. At PAM 94, positions 51 and 57 are on the surface. The assignment of secondary structure is based on several independent analyses. The segment most probably extends from positions 49 to 58, a total of six positions, insufficient for a 2-turn helix. Only if the segment includes the PGG (and aligned sequences) at positions 46 to 48 is there sufficient length for a helix to be considered. Any 3·6 residue amphiphilicity is, however, destroyed by surface assignments at position 57 (strong) and position 46 (weak). Thus, a plausible helix cannot extend over more than positions 47 to 52, a total of six residues. Such short helices, when they exist, are essentially interchangeable with coils. It is unlikely, however, that this segment builds a coil-like structure, given the pair of conserved Trp residues (positions 49 and 50). While surface coils often have single hydrophobic anchors, dipeptide anchors are rare, and not usually so highly conserved. The shortness of the segment permit only a moderately strong β assignment.

Segment 5: positions 62 to 65

At PAM 132, no interior assignment is made. At PAM 94, two strong surface assignments (position 63 and 64) are flanked by potential parses. These designate this segment as a coil.

Segment 6: positions 67 to 69

At PAM 132, two strong interior assignments are made (positions 67 and 68) together with a weaker assignment (position 69, also a parse). No surface assignment is made. Position 70 has a single hydrogen-bonding variable subgroup; such positions generally lie inside. As with segment 4, a moderately strong β assignment is indicated. The question remains whether the Pro (position 69) indicates a parse, or whether it simply causes a kink in a β-strand, or is in the first turn of a helix that continues later. Thus, the assignment made for the segment following is relevant.

Segment 7: positions 70 to 73

At PAM 132, interior assignments are made at positions 72 and 73. Position 71 contains an APC N. A surface assignment is made at position 74 at PAM 94. The end of the domain is reliably assigned at position 73 by the fact that every position that

follows is assigned to the surface, and the alignment in this region is not significant. Coupling with the preceding segment does not yield a pattern of 3·6 residue amphiphilicity. Therefore, segments 6 and 7 are assigned as separate β-strands, possibly kinked at position 69.

The SH3 C subfamily

The SH3 C subfamily contains 11 sequences, one of which (chick spectrin α) is the sequence for which the crystal structure has been determined (Musacchio *et al.*, 1992*b*) (Fig. 3(c)) arranged on a tree (Fig. 2(c)) with an overall PAM width of 190. The subfamily can be divided into two smaller subfamilies as the alignment between the first eight and the last three sequences is not particularly good. The discussion below is based on an analysis of the multiple alignment containing only the first eight sequences (subfamily SH3 C1), with a PAM width of 124. Based on primary parses, the multiple alignment is divided into six segments (segment I, 7–13; segment II, 19–37; segment III, 39–46; segment IV, 48–52; segment V, 57–61; and segment VI, 64–72). Segments I, III and IV have no secondary parses. Segment II (19–37) has weaker parses at positions 22 (dipeptide parse, confirmed subfamily C2) and 30 (dipeptide parse, APC G, confirmed in subfamily A). Segment V (57–61) has a weaker parse at positions 59-60 (dipeptide parse, confirmed in subfamily B). Segment VI (64–72) has weaker parses at positions 65 (APC G, confirmed in subfamily A) and 69 (APC P, confirmed throughout).

The SH3 C subfamily then is divided into clusters of subgroups at different maximum PAM widths, and surface and interior positions assigned. Unfortunately, both the size of the alignment and the balance in the evolutionary tree are not optimal for making surface and interior assignments. Therefore, secondary structure predictions are in this subfamily weaker than in subfamilies A and B.

Segment 1: positions 7 to 13

Interior assignments are made at positions 8, 9, 10 and 12. Surface assignments are made at positions 7 and 13, allowing a weak-moderate β assignment.

Segment 2: positions 19 to 21

Weak surface assignments suggest a coil assignment.

Segment 3: positions 23 to 29

Interior assignments at positions 25 and 27, with hydrogen-bonding variable at position 26 and a surface at position 28 allow a moderately strong β assignment.

Segment 4: positions 31 to 37

Interior assignments are made at positions 32, 33, 35 and 36. Weak surface assignments are made at positions 31, 34 and 37 (1 variable subgroup each). An amphiphilic helix can be built in this region. Breaking this helix is position 30 (APC G on surface arc), and 39 (2 variable subgroups), a surface assignment on the interior arc of the helix wheel.

Segment 5: positions 39 to 46

Three positions are assigned to the surface and contain parsing elements, yielding an assignment as a coil/turn.

Segment 6: positions 48 to 52

Interior assignments are made at positions 49, 50 and 52, suggesting a weak β assignment.

Segment 7: positions 57 to 61

An interior assignment is possible only at position 58. Positions 57, 59, 60 and 61 are all on the surface, suggesting a coil/turn.

Segments 8 and 9: positions 64 to 68 and 70 to 73

These segments indicate the problems encountered when attempting to predict secondary structure based on surface and interior assignments derived from a small alignment with a poorly balanced tree. The fusion of the carboxy terminus to a continuing peptide makes the end of the segment ill defined. A helix assignment can therefore be made with the following assignments: 67 inside, breaks amphiphilicity; 68 inside, fits helical wheel; 69 APC, P is assumed to lie inside; 70 single hydrogen-bonding variable only to PAM 92, probably inside, at interface between surface and interior arc; 71 single variable surface down to PAM 77, presumably surface; 72 inside; 73 inside; 74 single variable surface; 75 surface still fits helical wheel; 76 two variable subgroups, one with a polar residue, weak surface breaks amphiphilicity. Alternatively, if the conserved Pro at position 69 is considered a parse, and the segment is truncated at position 73 where the significant alignment ends (followed by a string of surface assignments), two β-strands are assigned (positions 65 to 68 and 71 to 73). Only by analogy with the A and B subfamilies is the second assignment preferred.

Discussion

The secondary structure assignments, collected in Table 1, illustrate three general points. One subfamily (A) was not highly divergent. One (subfamily B) had a poorly balanced evolutionary tree. Two

subfamilies (B and C) had very few sequences. In each case, prediction was hampered, but not prevented, by a less-than-optimal situation. By applying the prediction method independently to the three subfamilies, the homology of the subfamilies themselves can be indisputably established. We have recently applied similar predictions to analyze homologies among pyridoxal-dependent enzymes (Benner *et al.*, 1992*b*).

For comparison, a prediction based on the "GOR" method was investigated (Garnier *et al.*, 1978). A number of GOR predictions were generated for different sequences using the University of Wisconsin GCG software (Devereux *et al.*, 1984). In all cases the results suggested a highly helical protein, although in several cases the predictions obtained on homologues did not correspond closely with each other. Thus, the GOR prediction contrasts sharply with the prediction made here, which is for a predominantly β protein, with only a single 2-turn helix.

Finally, the predicted secondary structures allowed us to revise the original master alignment to allow secondary structures to overlap better (Fig. 1). This is, we believe, the first time predicted structures have been used to adjust alignments of distantly related proteins. It remains to be seen, of course, whether this revision improves the quality of the multiple alignment. For this, the crystal structure reported by Musacchio *et al.* (1992*b*) must be analyzed together with as yet unavailable crystal structures for representatives of subfamilies A and B.

Another test of the prediction will come from an analysis of surface and interior assignments *vis a vis* side-chain surface accessibility parameters obtained from the crystal structure. Finally, and most obviously, the quality of the prediction can be assessed by comparing the predicted secondary structure with the actual secondary structure of the proteins in the different subfamilies.

Although it is possible at this point to assemble a tertiary structural model for the SH3 domain from the predicted secondary structural elements (Benner & Gerloff, 1991), this process requires more care, introspection and, unfortunately, time. As this time is simply not available to us if we hope to have this prediction appear before the crystal structure appears, no tertiary structural model is presented at this time.

Predictions such as these are extremely important to the development of methods for predicting protein conformation. We welcome additional challenges to make predictions using our method, especially if (1) a structure shortly will be solved, (2) no structure is already available for any obviously homologous protein, (3) sequences are sent by computer mail with literature citations that provide an overview of the chemistry and biology of the protein family, and (4) this material is sent enough in advance to allow co-ordination of the publication of the prediction and publication of the structure.

We are indebted to Dr Andrea Musacchio and Dr Toby Gibson for informing us of the upcoming publication of a structure of an SH3 domain and issuing the challenge to predict the structure before its publication. M.A.C. was supported by a Wellcome Trust Traveling Fellowship.

References

Bazan, J. F. (1990). Structural design and molecular evolution of a cytokine receptor superfamily. *Proc. Nat. Acad. Sci., U.S.A.* **87**, 6934–6938.

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

Benner, S. A. (1992). Predicting *de novo* the folded structure of proteins. *Curr. Opin. Struct. Biol.* **2**, 402.

Benner, S. A. & Gerloff, D. (1991). Patterns of divergence in homologous proteins as indicators of secondary and tertiary structure: a prediction of the structure of the catalytic domain of protein kinases. *Advan. Enzym. Regulat.* **31**, 121–181.

Benner, S. A., Cohen, M. A. & Gerloff, D. (1992*a*). Correct structure prediction? *Nature (London)*, **359**, 781.

Benner, S. A., Cohen, M. A., Gonnet, G. H., Berkowitz, D. B. & Johnsson, K. (1992*b*). Reading the palimpsest: contemporary biochemical data and the RNA world. In *The RNA World* (Gesteland, R. & Atkins, J., eds), Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY. In the press.

Benner, S. A., Badcoe, I., Cohen, M. A. & Gerloff, D. (1993). *De novo* prediction of folded structures: assigning interior and surface residues from patterns of variation and conservation in homologous protein sequences. Submitted to *Journal of Molecular Biology*.

Cohen, F. E., Sternberg, M. J. E. & Taylor, W. R. (1982). Analysis and prediction of the packing of α-helices against a β-sheet in the tertiary structure of globular proteins. *J. Mol. Biol.* **156**, 821–862.

Cohen, F. E., Abarbanel, R. M., Kuntz, I. D. & Fletterick, R. J. (1983). Secondary structure assignment for α/β proteins by a combinatorial approach. *Biochemistry*, **22**, 4894–4909.

Crawford, I. P., Niermann, T. & Kirschner, K. (1987). Prediction of secondary structure by evolutionary comparison: application to the α subunit of tryptophan synthase. *Proteins*, **2**, 118–129.

Dayhoff, M. O., Schwartz, R. M. & Orcutt, B. C. (1978). A model of evolutionary change in proteins. In *Atlas of Protein Sequence and Structure* (Dayhoff, M. O., ed.), vol. 5, suppl. 3, pp. 345–352, National Biomedical Research Foundation Washington, DC.

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

Garnier, J., Osguthorpe, D. J. & Robson, B. (1978). Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol.* **120**, 97–120.

Gonnet, G. H. & Benner, S. A. (1991). Computational Biochemistry Research at ETH, *Technical Report 154, Department Informatik*, E.T.H., Zurich.

Hunt, T. & Purton, M. (1992). 200 issues of TIBS. *Trends Biochem. Sci.* **17**, 273.

Lesk, A. M. & Boswell, D. A. (1992). Does protein structure determine amino acid sequence? *BioEssays*, **14**, 407.

Musacchio, A., Gibson, T., Veil-Pekka, L. & Saraste, M. (1992*a*). SH3, an abundant protein domain in search of a function. *FEBS Letters*, **307**, 55–61.

Musacchio, A., Nobel, M., Paupit, R., Wierenga, R. & Sataste, M. (1992*b*). Crystal structure of an Src-homology 3 (SH3) domain. *Nature (London)*, **359**, 851–855.

Editorial Footnote: This paper by Benner *et al.* on a secondary structure prediction of the Src-homology (SH3) domain has been published under unusual circumstances. The authors accepted a challenge made by the crystallography group that solved the structure to predict the structure in advance of its publication (Musacchio *et al.*, 1992). A summary of the prediction (Benner *et al.*, 1992) was published in the issue of *Nature* that contained the crystal structure. Benner *et al.* submitted a full paper to this journal prior to the publication of the crystal structure. The predictive paper was submitted and accepted with the proviso that it would not be materially altered after reviewing and in light of the knowledge of the crystal structure. For a comparison of the predicted and actual structures see Rost & Sander (1992).

References

Benner, S. A., Cohen, M. A. & Gerloff, D. (1992). Correct structure prediction? *Nature (London)*, **359**, 781.

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

Rost, B. & Sander, C. (1992). Jury returns on structure prediction. *Nature (London)*, **360**, 540.

Edited by A. R. Fersht