CABIOS

Vol.9. no.6. 1993
Pages 745 – 756

Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation

Craig D.Livingstone and Geoffrey J.Barton[1]

Abstract

An algorithm is described for the systematic characterization of the physico-chemical properties seen at each position in a multiple protein sequence alignment. The new algorithm allows questions important in the design of mutagenesis experiments to be quickly answered since positions in the alignment that show unusual or interesting residue substitution patterns may be rapidly identified. The strategy is based on a flexible set-based description of amino acid properties, which is used to define the conservation between any group of amino acids. Sequences in the alignment are gathered into subgroups on the basis of sequence similarity, functional, evolutionary or other criteria. All pairs of subgroups are then compared to highlight positions that confer the unique features of each subgroup. The algorithm is encoded in the computer program AMAS (Analysis of Multiply Aligned Sequences) which provides a textual summary of the analysis and an annotated (boxed, shaded and/or coloured) multiple sequence alignment. The algorithm is illustrated by application to an alignment of 67 SH2 domains where patterns of conserved hydrophobic residues that constitute the protein core are highlighted. The analysis of charge conservation across annexin domains identifies the locations at which conserved charges change sign. The algorithm simplifies the analysis of multiple sequence data by condensing the mass of information present, and thus allows the rapid identification of substitutions of structural and functional importance.

Introduction

A protein that exhibits key biological functions will commonly have homologues sequenced from many different tissues and organisms. Accurate multiple sequence alignment of such a protein family can highlight the residues of common functional and structural importance. The location of identities and conservative substitutions may be used to guide the design of site-directed mutagenesis experiments whilst the identification of subtle patterns of residue conservation can yield improvements in the accuracy of secondary and tertiary structure predictions (Crawford, *et al.*, 1987; Zvelebil *et al.*, 1987; Benner and Gerloff, 1990; Barton *et al.*, 1991; Russell *et al.*, 1992). Such analyses of multiple sequence alignments have traditionally been performed by eye. However, for large alignments, only the most obvious patterns of residue conservation can be easily identified by this method. When many long sequences are to be scrutinized, the task becomes unmanageable, and the risk of missing interesting residue substitutions is great.

A number of computer programs have been developed to aid the interpretation of multiple sequence alignments. The programs PRETTY and PRETTYPLOT from the GCG package (Devereux *et al.*, 1984) derive consensus amino acid sequences and box the largest group of similar residues at each position of an alignment. ALSCRIPT (Barton, 1993) allows shading, boxing and colouring to be applied to an alignment. Colour is also exploited by the SOMAP program (Parry-Smith and Attwood, 1991), which colours residues according to which user-defined set they belong (e.g. hydrophobic, charged). The amino acid variation at a position in an alignment is reduced to a single figure of 'variability' by Kabat (1976), 'entropy' or 'variation' by Sander and Schneider (1991), 'information' by Smith and Smith (1990) and 'evolutionary divergence' by Brouillet *et al.*, (1992). In contrast, the novel set-based approach described by Taylor (1986), defines the minimal set of physico-chemical properties that represent any group of amino acids. This principle has been developed by Zvelebil *et al.* (1987) so that the minimal set of amino acids could be encoded as a single 'conservation number' at each position in the alignment. Although very effective at highlighting the overall similarity at each position in an alignment, none of these methods deal with the problem of quantifying similarities between subfamilies within a larger multiple sequence alignment.

It is frequently desirable to subdivide a protein family on the basis of function, origin, sequence similarity or other criteria. Indeed, most multiple alignment methods (e.g. Barton, 1990; Barton and Sternberg, 1987; Feng and Doolittle, 1987; Higgins and Sharp, 1989) first compare all sequences pairwise, then automatically cluster the sequences into subfamilies on the basis of sequence similarity. Such cluster analysis can readily identify the gross similarities between sequences but does not pinpoint the residue positions that are responsible for the clustering pattern. It may also be difficult to rationalize the clusters identified by overall sequence similarity with those implied by functional similarity since functional differences may reside in a few key residues. Although all previous methods for characterizing residue conservation (e.g. Kabat, 1976; Devereux

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

[1]*To whom correspondence should be addressed*

© Oxford University Press

et al., 1984; Taylor, 1986; Smith and Smith, 1990; Parry-Smith and Attwood, 1991; Sander and Schneider, 1991; Brouillet *et al.*, 1992) provide a clear overview of conservation across an alignment, they do not allow the automatic identification of residue positions specific to subgroups of sequences within the alignment.

In this paper we describe an algorithm for the systematic identification of residue conservation within aligned protein sequences. The algorithm operates in a hierarchical manner, by first characterizing conservation on a residue-by-residue basis within predefined subfamilies, then between all pairs of subfamilies. This hierarchical approach highlights positions that may be responsible for conferring the specific structural and functional properties of the subfamilies.

Systems and methods

The hierarchical conservation analysis algorithm is implemented in the computer program AMAS (Analysis of Multiply Aligned Sequences) written in ANSI-C. AMAS can generate commands for the ALSCRIPT program (Barton, 1993), which will automatically shade, box and colour a multiple alignment according to the identified conservation patterns. AMAS and ALSCRIPT have been used successfully on a number of Unix platforms. If the graphical display options are required, then a Postscript printer or interpreter is required.

Algorithm

Quantification of amino acid residue conservation

We have extended the work of Zvelebil *et al.* (1987) to give a general method for quantifying residue conservation. Our approach differs in detail to that described by Zvelebil *et al.*, so for the sake of completeness and to avoid possible confusion we here describe the protocol used to quantify and compare residue conservation.

Figure 1(a) illustrates a Venn diagram (for details see Taylor, 1986) which is contained within a boundary that symbolizes the universal set of 20 common amino acids (ϵ). The amino acids that possess the dominant properties—hydrophobic, polar and small (<60 Å^3)—are defined by their set boundaries. Subsets contain amino acids with the properties aliphatic (branched sidechain non-polar), aromatic, charged, positive, negative and tiny (<35 Å^3). Shaded areas define sets of properties possessed by none of the common amino acids. The Venn diagram may be simply encoded as the property table or index shown in Figure 1(b), where the rows define properties and the columns refer to each amino acid.

Cysteine occurs at two different positions in the Venn diagram. When participating in a disulphide bridge (C_{S-S}), cysteine exhibits the properties 'hydrophobic' and 'small'. In addition to these properties, the reduced form (C_{S-H}) shows polar character and fits the criteria for membership of the 'tiny' set.

When analysing proteins that do not have disulphides, an index which represents the properties of reduced cysteine is used (see SH2 domain analysis). In proteins where disulphide bonding is known to occur, or where the oxidation state of the cysteines is uncertain, an index representing cysteine in the oxidized form is generally more useful (as in Figure 1b).

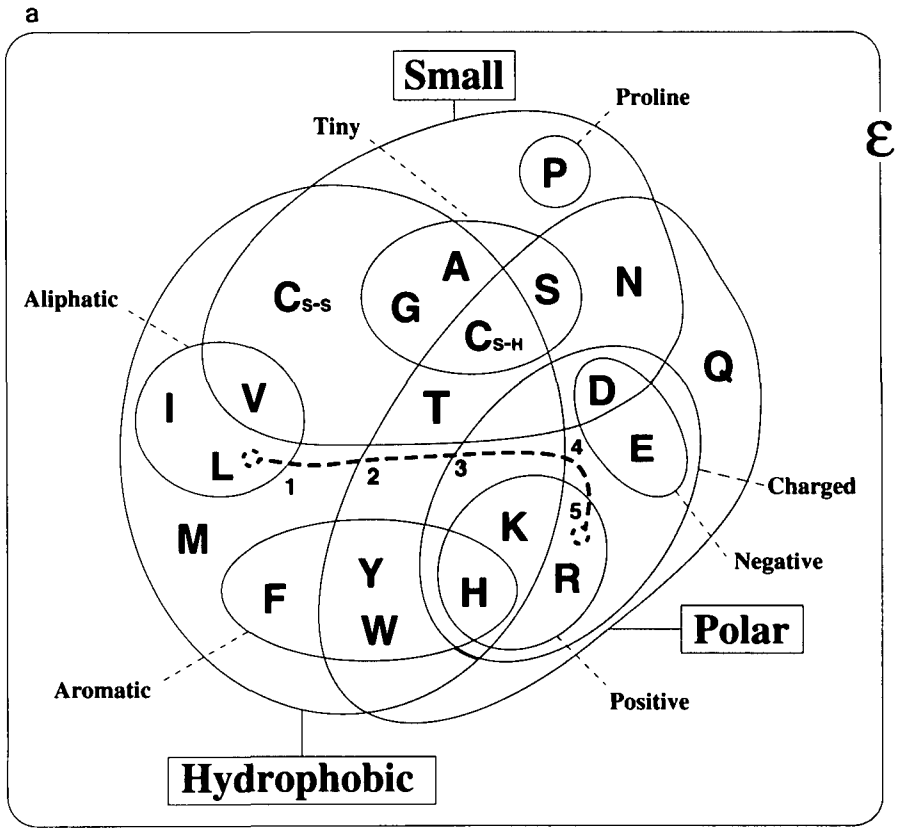
The illustrated Venn diagram (Figure 1a) assigns multiple properties to each amino acid; thus lysine has the property hydrophobic by virtue of its long sidechain as well as the properties polar, positive and charged. Alternative property tables may also be defined. For example, the amino acids might simply be grouped into non-intersecting sets labelled, hydrophobic, charged and neutral.

Figure 2 illustrates the stages involved in the calculation of conservation numbers for a simplified property index (Figure 2a and b). All of the amino acids are assigned to the universal set (ϵ), which in this simple example contains only the charged subset, which in turn is broken down into subsets containing positively and negatively charged amino acids. This property index allows the positions of conserved charges to be identified, together with positions where a conserved charge changes polarity between different groups of sequences within an alignment.

The amino acids occurring at each position in the multiple alignment are recorded (Figure 2d), then tested for the presence of each of the three properties (Figure 2b). This is represented by the columns of entries for each amino acid (Figure 2e). For example, at aligned position 11, the first column in Figure 2(e) represents the properties of arginine, the second column the properties of tryptophan and so on. Filled circles show the amino acid is a member of a property set, empty circles indicate non-membership.

Each property is considered in turn by examining the rows of entries in Figure 2(e). If all of the amino acids at a position possess the property, then the position shows positive conservation; all entries on that property's row in Figure 2(e) will be filled circles and a filled circle appears in Figure 2(f). If all amino acids at a position lack the property, then the position shows negative conservation; all entries on the row in Figure 2(e) will be empty circles and an empty circle is seen in Figure 2(f). If the possession of a property varies in the set of amino acids being considered, filled and empty circles appear in the equivalent row in Figure 2(e), the property is labelled as unconserved and a shaded circle is shown in Figure 2(f).

Two methods are used to quantify conservation at an alignment position using the information stored in Figure 2(f). Method 1 is similar to that of Zvelebil *et al.* (1987) and regards as conserved any property that is either positively or negatively conserved. The number properties obeying this rule (number of filled or empty circles for a position in Figure 2f) is summed



b

	ILVCAGMFYWHKREQDNSTPBZXΔ	
1	●●●●●●●●●●●○○○○○○○●○○○●●	Hydrophobic
2	○○○○○○○○●●●●●●●●●●●○●●●●	Polar
3	○○●●●●○○○○○○○○○●●●●●○○●●	Small
4	○○○○○○○○○○○○○○○○○○○●○○●●	Proline
5	○○○○●●○○○○○○○○○○○○●○○○○●●	Tiny
6	●●●○○○○○○○○○○○○○○○○○○○●●	Aliphatic
7	○○○○○○○●●●●○○○○○○○○○○○●●	Aromatic
8	○○○○○○○○○○●●●○○○○○○○○○○●●	Positive
9	○○○○○○○○○○○○○●○●○○○○○○○●●	Negative
10	○○○○○○○○○○●●●●○●○○○○○○○●●	Charged

Fig. 1. Physico-chemical properties of the amino acids. (**a**) The 20 common amino acids are shown in terms of 10 physico-chemical properties (Taylor, 1986; Zvelebil *et al.*, 1987). Grey-filled areas define sets of properties possessed by none of the common amino acids. The hydrophobic, polar and small sets dominate the figure. The remaining sets define subsidiary groups. The dotted line joining L to R shows the minimum number of five set boundaries which must be crossed in order to change an L to an R in this 10 property diagram (see text). (**b**) An amino acid property index derived from the Venn diagram in (a) (after Zvelebil *et al.* (1987), treating Cys as C_{S-S}). The columns represent the amino acids while rows represent properties. Filled circles show when an amino acid possesses a property. Δ represents gap which, in this index, is regarded as having all properties.

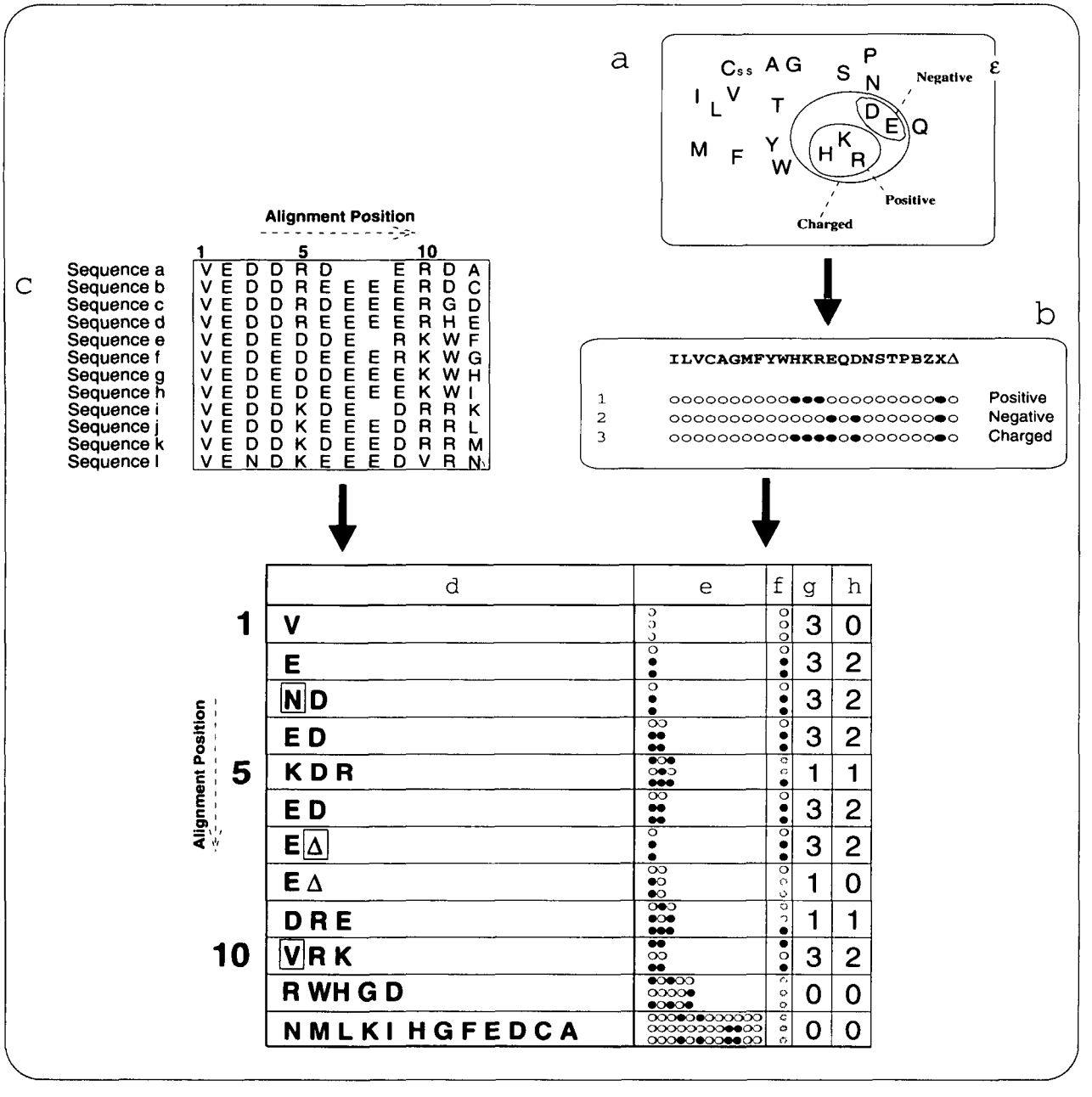


Fig. 2. Calculation of conservation numbers. The Venn diagram showing the relationship between the amino acids on the basis of charge (**a**) is converted to a property index (**b**), which is used to analyse the conservation of charged residues in the sequence alignment (**c**). The amino acids present at each sequence position are recorded (**d**) and tested for each of the properties in the index (**e**). Columns of filled (presence of a property) and empty (lack of a property) circles record the properties of each amino acid in the same vertical order as in the property index. The presence of properties is summed (**e**), filled circles show positive conservation of a property in the group of amino acids, shaded circles show where properties are present in some but not all of the amino acids, and empty circles show negatively conserved properties. A conservation score is arrived at by summing either the number of positively and negatively conserved properties (**g**—method 1) or the number of positively conserved properties alone (**h**—method 2) (see text).

to give the conservation number (Figure 2g). In contrast, method 2 only counts properties that are positively conserved (filled circles in Figure 2f) and gives the conservation numbers shown in Figure 2(h).

The method 1 conservation value is a function of the number of set boundaries P that must be crossed to visit all the amino acids at a position. If a property index contains N properties then the conservation number (C_n) is $N - P$. For example, the dotted line in Figure 1(a) joins Leu and Arg and crosses five set boundaries, thus for this property matrix, $C_n(L,R) = 10 - 5 = 5$. The maximum possible value for the conservation number calculated by method 1 is given by the number

of properties in the property index (3 for Figure 2b; 10 for Figure 1b).

Conservation by method 2 is calculated by counting the number of sets common to all amino acids at a position. Leu and Arg in Figure 1(a) share no properties; by method 2, their conservation number is 0. Asp and Glu in Figure 2(a) are both members of the sets charged and positive; their conservation number by method 2 is 2. The maximum value for the conservation value calculated by method 2 is the maximum number of properties possessed by a single amino acid in the property index.

Treatment of gaps and unusual residues

Insertions and deletions (gaps—Δ) are usually tolerated only in surface loop regions. Accordingly, gaps are normally given all properties in the property matrix so that aligned positions that contain a gap are assigned a low conservation value.

The set-based conservation analysis described here is independent of the number of sequences analysed. For example, a position in an alignment of 100 sequences that contains 99 alanines and one lysine will give the same conservation value as a position in an alignment of two sequences that has one alanine and one lysine. The advantage of this approach is that the tolerance of particular physico-chemical properties at a position indicates the likely environment of the amino acids in the common fold of the protein family. This reasoning suggests that a position that conserves valine in 99 sequences, but also shows aspartate is unlikely to be performing a common structural or functional role. However, it may sometimes be suspected that one or more of the sequences contain errors, or that there are errors in the alignment. It is then desirable to relax the strict conservation rules. Accordingly, a predetermined number of gaps or residues that represent $<N\%$ of the total at a position may be ignored when calculating conservation values. For example, alignment position 3 in Figure 2 is predominantly Asp. This position would not be recorded as conserved using the charge index due to the presence of a single Asn (1/12 or 8.3% of the sequences in the alignment). If a 10% threshold for unusual residues is set, then this Asn would be ignored when calculating the conservation value (similarly, Val at position 10). Positions where unusual residues have been ignored are reported only as conserved, never as identical even if the other residues present are identical (Figure 2, position 3). It is the ability to quantify the conservation of amino acids that gives the set-based approach its major advantage over averaging a single property scale, caution must therefore be exercised when deciding to ignore gaps and unusual residues.

Hierarchical conservation analysis

The procedures described in the previous section are a straightforward extension of the principles described by Zvelebil *et al.* (1987) and Taylor (1986). Here we extend the set-based method to identify conserved features of sequence subgroups within larger protein sequence alignments.

The starting point for hierarchical conservation analysis is the identification of two or more subsets of sequences within a multiple sequence alignment. The subsets may be defined by grouping on the basis of overall sequence similarity, by functional similarity, origin or other criteria. Given such groupings, the aim is to highlight which residue positions define the unique properties of each group.

Figures 3 and 4 illustrate the result of applying hierarchical conservation analysis to a nine residue fragment of a 26 sequence multiple alignment using the 10 property index shown in Figure 1. The dendrogram shown at the left of Figure 3 shows the overall similarity between the sequences (i.e. not just the nine residues) and clearly splits the sequences into three subgroups labelled A, B and C.

Conservation numbers are calculated for each alignment position in each subgroup and a conservation threshold is set. This reference point is used to put each position within a sub-group into one of three classes: (i) identical positions; (ii) conserved positions, where the conservation number is greater than or equal to the threshold; and (iii) unconserved, where the conservation number is less than the threshold. The choice of threshold depends upon the particular conservation index being used. For the index shown in Figure 1, a threshold of between 6 and 8 normally gives the most informative results. In Figure 3, the different classifications using a threshold of 8 are illustrated by shading and font changes. For example, in subgroup A, identities are shown in white on dark grey at positions 2 and 4, conserved positions are in black on light grey (positions 6–9), and unconserved positions are illustrated in italics on a white background (positions 3 and 5). At position 1, the identity in all sequences is marked by white on black lettering, whilst at position 10 chancery script lettering is used to highlight the lack of conservation within all sub-groups.

Having classified the conservation within each subgroup, all pairs of subfamilies are compared and conservation numbers calculated for each position in the pairs. In the calculation of conservation for a pair of subfamilies, the residues from the pair are considered as members of a single group. C_n is then calculated, as described above, for the composite group according to which method was chosen. The change in conservation value that occurs when each pair of subfamilies is brought together reflects the similarities or differences in physico-chemical properties seen in each subgroup at that position. For example, at position 7 of subfamilies A and B the conservation values in A, B and A + B are 9, showing that the properties are conserved within each family, and across both families at this position. This is, therefore, a location that exhibits common physico-chemical properties between A and B, yet these properties are not conserved within group C. Accordingly, this may indicate a tertiary structural feature shared between A and B, but not C.

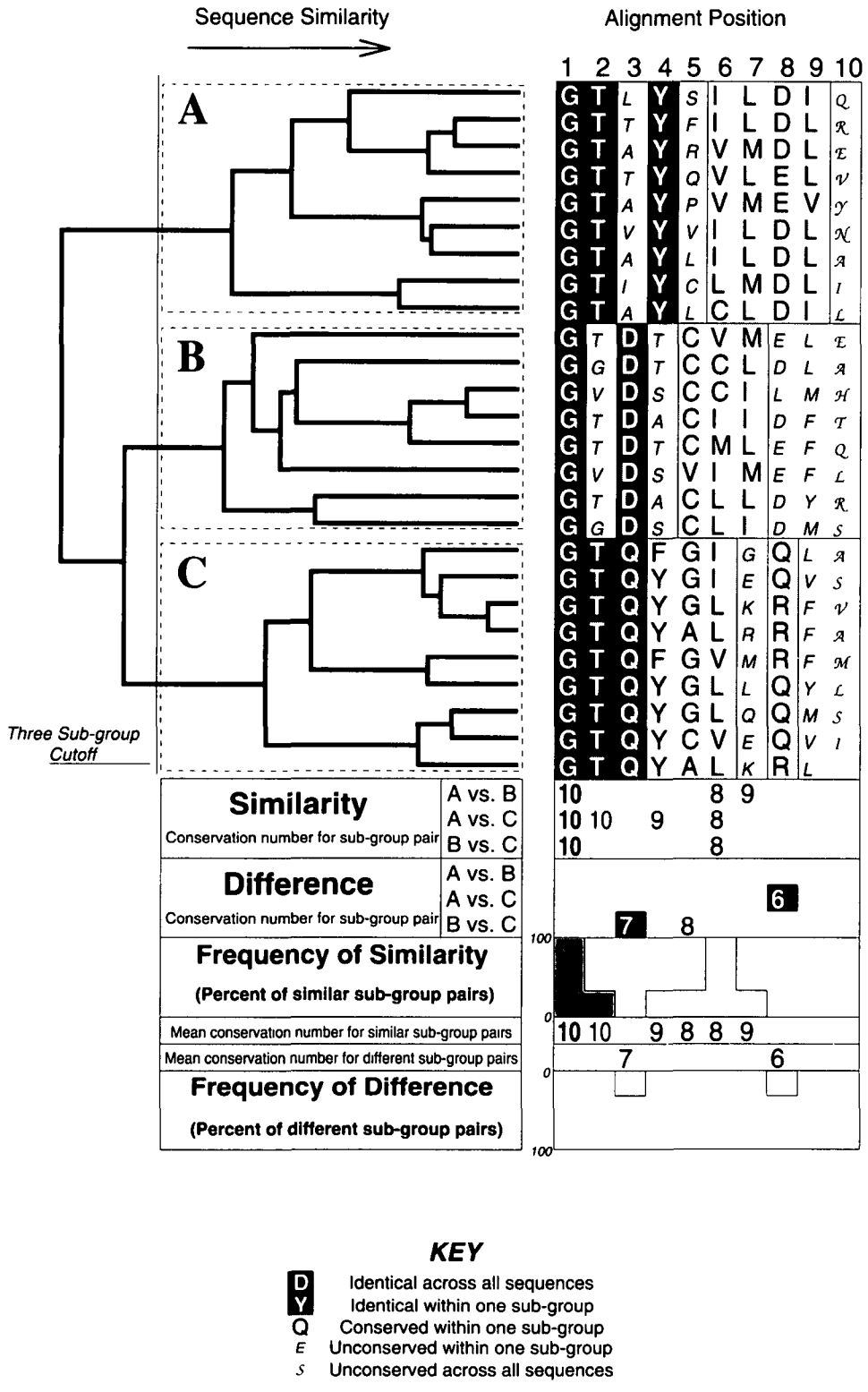


Fig. 3. Hierarchical conservation analysis. A 10 residue fragment of a multiple sequence alignment of 26 sequences is shown to the right of the figure. The relationship between the sequences in the whole alignment is represented by the dendrogram to the left, which shows three sub-groups: A, B and C. Each position of the groups in the multiple sequence alignment has been analysed for residue conservation using the property index in Figure 1(b). The conservation threshold was set to 8. Information about the conservation pattern is given at the foot of the alignment in numerical and graphical form. The representation of the alignment and the conservation patterns to the right of the figure were imported directly from the graphical output of the programm AMAS.

```
Conservation Threshold:  8 (Maximum: 10)
Number of gaps ignored:  0
Percentage of unusual residues ignored:  0%

Sub-group A, sequences:  1 to 9
Sub-group B, sequences:  10 to 17
Sub-group C, sequences:  18 to 26

❶ Identity between all sub-groups
( 1) ALL [G]

  Identity between sub-group pairs
( 2) A,C [T]

  Identity within one sub-group
( 3) B [D]  C [Q]
( 4) A [Y]

❷ Conservation between all sub-groups
                                                          %     %     %
( 6) ALL ( 8, 8, 8) A-[ILVC] B-[ILVCM] C-[ILV]  Cons: Hydrophobic  Diff: Small    A-[ 44] B-[ 37] C-[ 22]
                                                                   Aliphatic  A-[ 88] B-[ 62] C-[100]

  Conservation between sub-group pairs
                                                          %     %
( 4) A,C ( 9) A-[Y] C-[FY]      Cons: Hydrophobic  Diff: Polar      A-[100] C-[ 77]
                                      Aromatic

( 7) A,B ( 9) A-[LM] B-[ILM]    Cons: Hydrophobic  Diff: Aliphatic  A-[ 66] B-[ 75]

  Difference between sub-group pairs
                                                                    %     %
( 3) B,C ( 7) B-[D] C-[Q]       Cons: Polar        Diff: Small      B-[100] C-[ 0]
                                                        Negative    B-[100] C-[ 0]
                                                        Charged     B-[100] C-[ 0]

( 5) B,C ( 8) B-[VC] C-[CAG]    Cons: Hydrophobic  Diff: Tiny       B-[ 0] C-[ 88]
                                      Small             Aliphatic   B-[ 12] C-[ 0]

( 8) A,C ( 6) A-[ED] C-[RQ]     Cons: Polar        Diff: Small      A-[ 77] C-[ 0]
                                                        Positive    A-[ 0] C-[ 44]
                                                        Negative    A-[100] C-[ 0]
                                                        Charged     A-[100] C-[ 44]

  Conserved within one sub-group
( 9) A ( 9) A-[ILV]             Cons: Hydrophobic
                                      Aliphatic

❸ Unconserved within one sub-group
( 2) B ( 7) B-[TGV]
( 3) A ( 6) A-[LTAVI]
( 4) B ( 7) B-[TSA]
( 5) A ( 1) A-[SFRQPVL]
( 7) C ( 2) C-[GEKRMLQ]
( 8) B ( 4) B-[EDL]
( 9) B ( 7) B-[LMFY]            C ( 6) C-[LVFYM]

  Unconserved across all sub-groups
( 10) A ( 1) A-[QREVYNAIL]  B ( 1) B-[EAHTQLRS]  C ( 0) C-[ASVMLIΔ]
```

Fig. 4. Text representation of sequence conservation. With reference to Figure 3. The text representation of the analysis gives a more detailed description of the conservation of physico-chemical properties at each alignment position. Each record identifies the sequence position to which it refers (rounded brackets), the sub group(s) involved in the pattern being reported, the pair conservation number(s) of those groups where non-identities are reported (rounded brackets), the residues present in each group (square brackets) and the properties which are conserved by them and which differ between them. Differences in properties between subgroups are reported; the percentage of residues in each subgroup that have a property is shown in square brackets.

In contrast, at position 8 of subgroups A and C, in order to 'visit' all members of the combined set of amino acids from A + C (DEQR) a minimum of four set borders must be crossed, giving a value of C_n as 10 — 4 =6. The conservation values for A, C and A + C are, therefore 9, 8 and 6 respectively. Thus, although properties are conserved within each subgroup

at this position, the properties that are conserved differ between the subgroups. This type of conservation pattern might highlight a position in the protein structure that defines the specificity for a substrate. For example, the switch from a predominantly negative to positive charge between groups A and C may signal increased binding for a negatively charged moiety for the group C sequences when compared to group A.

General rules for linking such substitution patterns to changes in three-dimensional structure or function are as yet unknown. However, changes in conservation of charge, hydrophobicity or amino acid size are likely to be of importance in all protein families.

The result of the pairwise comparison of subfamilies is summarized below the alignment in Figure 3. The conservation values for the pairs of subgroups are either displayed as similarities of differences according to the rules shown in Table I. The similarity and difference sections are also summarized as histograms.

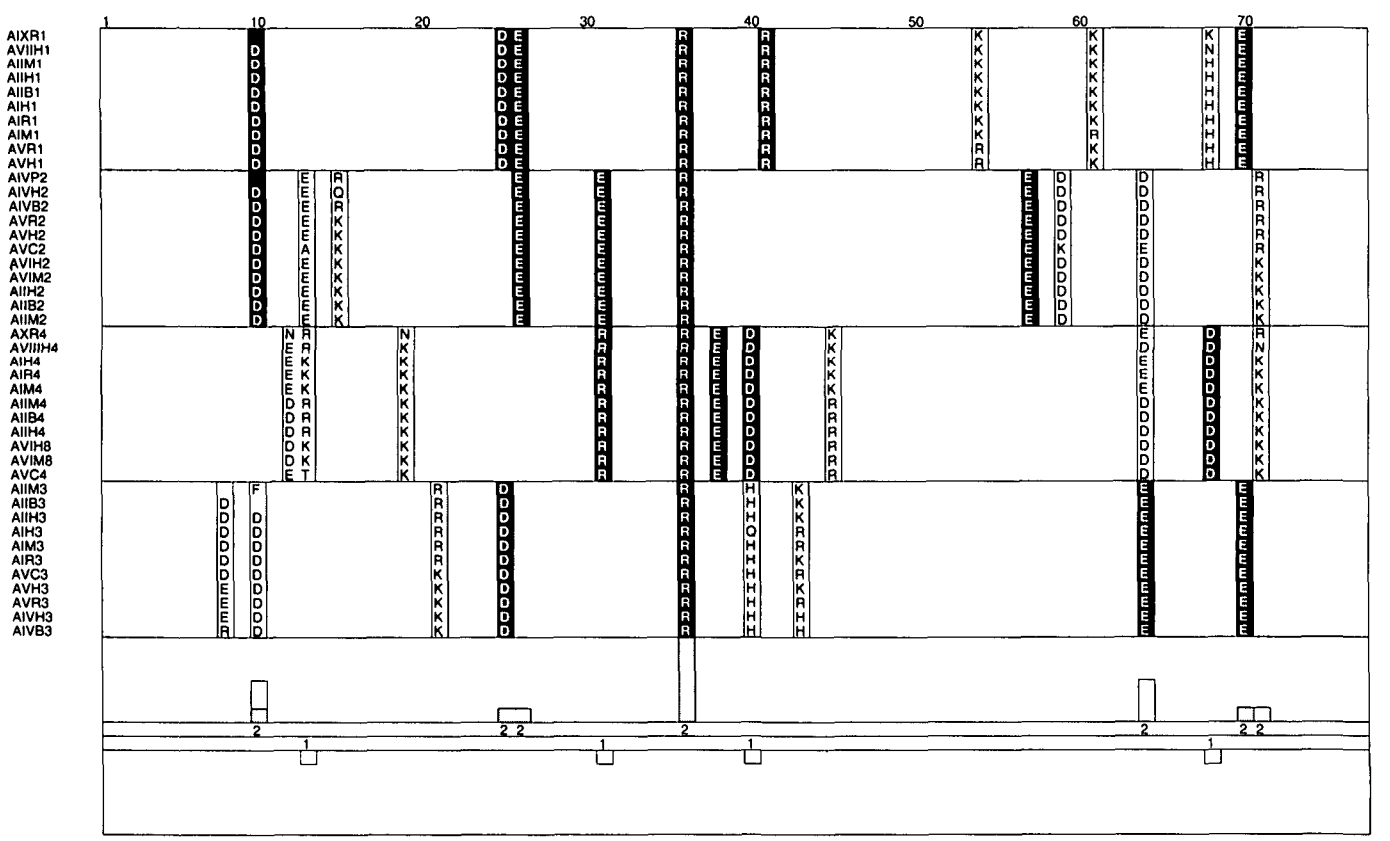
The hierarchical clustering approach addresses the problem of how to weight the information content of each sequence in

Table I. Pair comparison of conserved sequence subgroups

Sub-groups compared			Display C_{A+B} as similarity/difference
A	B	A + B	
$C_A \geq T$	$C_B \geq T$	$C_{A+B} \geq \min C_A, C_B$	similarity
$C_A \geq T$	$C_B \geq T$	$C_{A+B} < \min C_A, C_B$	difference but conserved
$C_A \geq T$	$C_B \geq T$	$C_{A+B} < T$	difference and unconserved
$C_A < T$	$C_B \geq T$	–	–
$C_A \geq T$	$C_B < T$	–	–
$C_A < T$	$C_B < T$	–	–

Conservation values are calculated for the sub-groups A and B, and for the sub-groups combined A + B. A conservation threshold T is set, similarities or differences are reported according to the logical operations shown.

a



AMAS conservation plot of the sequence alignment newest
(Stringency is set to 66% with 3 property matrix /home/orac/cdt/brv/ch pt)

b

Conservation Threshold: 2 (Maximum: 3)
Number of gaps ignored: 2
Percentage of unusual residues ignored: 12%

Sub-group A, sequences: 1 to 10
Sub-group B, sequences: 23 to 33
Sub-group C, sequences: 45 to 55
Sub-group D, sequences: 67 to 77

❶

Identity between all sub-groups

(36) ALL [R]

Identity between sub-group pairs

(10) A,B [D]
(25) A,D [D]
(26) A,B [E]
(70) A,D [E]

Identity within one sub-group

(31) B [E] C [R]
(38) C [E]
(40) C [D]
(41) A [R]
(57) B [E]
(64) D [E]
(68) C [D]

❷

Conservation between all sub-groups

NONE

Conservation between sub-group pairs

(10) A,D (2) A-[D] D-[D] Cons: Negative Charged Diff
B,D (2) B-[D] D-[D] Cons Negative Charged Diff.
(64) B,C (2) B-[D] C-[ED] Cons. Negative Charged Diff:
B,D (2) B-[D] D-[E] Cons: Negative Charged Diff
C,D (2) C-[ED] D-[E] Cons: Negative Charged Diff:
(71) B,C (2) B-[KR] C-[K] Cons Positive Charged Diff:

Difference between sub-group pairs

				%	%
(13) B,C (1) B-[E] C-[KR]	Cons Charged	Diff	Positive	B-[0]	C-[90]
			Negative	B-[90]	C-[0]
(31) B,C (1) B-[E] C-[R]	Cons: Charged	Diff:	Positive	B-[0]	C-[100]
			Negative	B-[100]	C-[0]
(40) C,D (1) C-[D] D-[H]	Cons: Charged	Diff:	Positive	C-[0]	D-[90]
			Negative	C-[100]	D-[0]
(68) A,C (1) A-[H] C-[D]	Cons Charged	Diff	Positive	A-[90]	C-[0]
			Negative	A-[0]	C-[100]

Conserved within one sub-group

(8) D (2) D-[ED] Cons: Negative Charged
(12) C (2) C-[ED] Cons: Negative Charged
(15) B (2) B-[KR] Cons Positive Charged
(19) C (2) C-[K] Cons: Positive Charged
(21) D (2) D-[KR] Cons. Positive Charged
(43) D (2) D-[HKR] Cons: Positive Charged
(45) C (2) C-[KR] Cons Positive Charged
(54) A (2) A-[KR] Cons. Positive Charged
(59) B (2) B-[D] Cons Negative Charged
(61) A (2) A-[K] Cons Positive Charged

Fig. 5. Charge conservation in 40 annexin repeats. (**a**) The pattern of conserved charge in 40 annexin repeats determined using the charge property index described in Figure 2. Only positive property conservation is considered at a conservation threshold of 2, this means that a subgroup position must conserve both charge and polarity to be reported. Conserved positions alone are reported in order to highlight the pattern of charged residues; the residues at unconserved positions have been masked out. Two gaps, and residues constituting $<10\%$ of a subgroup position have been screened from the conservation calculation. Identities and conserved positions are identified according to the shading protocol given in Figure 3. A charge difference is clearly seen in the histogram at position 31, reflecting the switch between a conserved E (negative) in repeat 2 and a conserved R (positive) in repeat 4. (**b**) Text output accompanying the analysis in Figure 6(a). The record format used is identical to that used in Figure 4.

an alignment. At the simplest level, each sequence would be treated equally but this relies on the sequences being equally diverse throughout the alignment. The use of clustering to derive conservation patterns ensures equal weight is given to different groups of proteins irrespective of the number of examples of each type. Inevitably, this process involves the loss of informa-

tion about the minor sequence variation which is responsible for subtle differences in character similar proteins in a subgroup. This loss is balanced by the ability to detect the more substantial changes in conservation which determine the differences in properties between the separate subgroups.

Implementation

Text representation

AMAS accepts command line arguments and provides a detailed textual breakdown of the conservation within a multiple alignment. Figure 4 illustrates the AMAS textual analysis that corresponds to the alignment shown in Figure 3. Only those positions that display conservation of the properties in the chosen property index are described. The presentation of the text results is hierarchical. Identities are described first (1), followed by positions showing conservation of physico-chemical properties (2), and unconserved positions listed last (3). Each entry contains a record of the alignment position (rounded brackets to the left), of the subgroup(s) to which it refers and a list of the residues in each subgroup cited (square brackets). In addition for positions that do not show identities, the properties conserved at the position, and those that differ are reported. With reference to Figure 4:

- *Identities.* Section 1 lists those sequence positions that are identical across the whole alignment, between pairs of subgroups and within one subgroup. Information is not repeated lower down the hierarchy if it has already been presented, e.g. the Gly at position 1 in the alignment is not also reported as two pairs of identical subgroups or as three identical individual subgroups.
- *Conservation of properties.* Conservation of physico-chemical properties between subgroups (following the same redundancy rules as for identities) is reported in section 2. The four categories of conserved positions are: (1) all subgroups conserve similar properties; (2) pairs of conserved subgroups share properties; (3) pairs of conserved subgroups have dissimilar properties; and (4) individual subgroups are conserved. The properties that are positively conserved between pairs of subgroups are listed, as are those properties that cause differences between subgroups. For each of a pair of different subgroups, the percentage of residues that display the differing properties is shown in square brackets.
- *Unconserved.* There are two divisions, the first for single unconserved subgroups and the second for entirely unconserved alignment positions.

Graphical display

The optional graphical representation of results mimics a hand analysis of the alignment using coloured marker pens. In Figure 3 the alignment is shown divided into three subfamilies. Within the subfamilies, at each alignment position, the amino acids are appropriately highlighted. Conserved subgroups, subgroups showing identity and positions that show identity across the whole alignment are labelled. Figures 5 and 6 illustrate the graphical representation applied to the annexin and SH2 domains.

Three highlighting methods have been explored. Monochrome methods allow grey shading (Figure 5 and 6) or the use of different fonts (not shown) to highlight the differences in conservation. Grey shading is preferable for publication, whilst unshaded alignments are useful as working copies for hand annotation. Colour may be specified as an alternative to shading to provide additional visual impact.

Discussion

The strategy described in this paper is extremely flexible: it allows different physico-chemical properties to be examined independently, or in concert. In addition, an alignment may be dissected into any combination of subgroups and their relative conservation analysed. As with any analytical procedure, the strategy is most effective when one has a clear idea of what one is looking for. For example: 'What makes subgroup A different from B and C?', or 'Which residues in subgroup D should I change to make D more like A?' If no clear questions have been defined, then the general property index (Figure 1b) is a useful starting point to highlight patterns of residue conservation. This is illustrated in Figure 6 for an alignment of 67 SH2 domains (Russell *et al.*, 1992). Since SH2 domains are cytoplasmic, Cys was assigned the properties of the free amino acid (C_{S-H}) in this analysis (Figure 1b). The alignment is divided into eight subgroups on the basis of overall sequence similarity. Subgroups 1−7 (numbering from the top) share >20% sequence identity, whilst sequences not fitting into one of these subgroups are collected in subgroup 8. The overall conservation of physico-chemical properties is highlighted by the histogram at the base of the alignment. The upper histogram indicates the normalized frequency of similarities between pairs of subgroups, whilst the lower plot shows the frequency of pair differences. Dark shading of the histogram indicates the frequency of pairs of subgroups that show sequence identity. A hand analysis of an alignment similar to that shown in Figure 6 correctly identified the location of the core secondary structures, and phosphotyrosine-binding residues (Russell *et al.*, 1992; Barton and Russell, 1993). Since completion of that study, the three-dimensional structures of three SH2 domains have been determined by the techniques of X-ray crystallography and NMR. The secondary structures of these are illustrated at the base of Figure 6 (Booker *et al.*, 1992; Overduin *et al.*, 1992; Waksman *et al.*, 1992). The conservation histograms clearly correspond to the regions of secondary structure, and are helpful in identifying patterns characteristic of α-helix and β-strand. For example, at positions 15 and 97, CXXCCXXC patterns

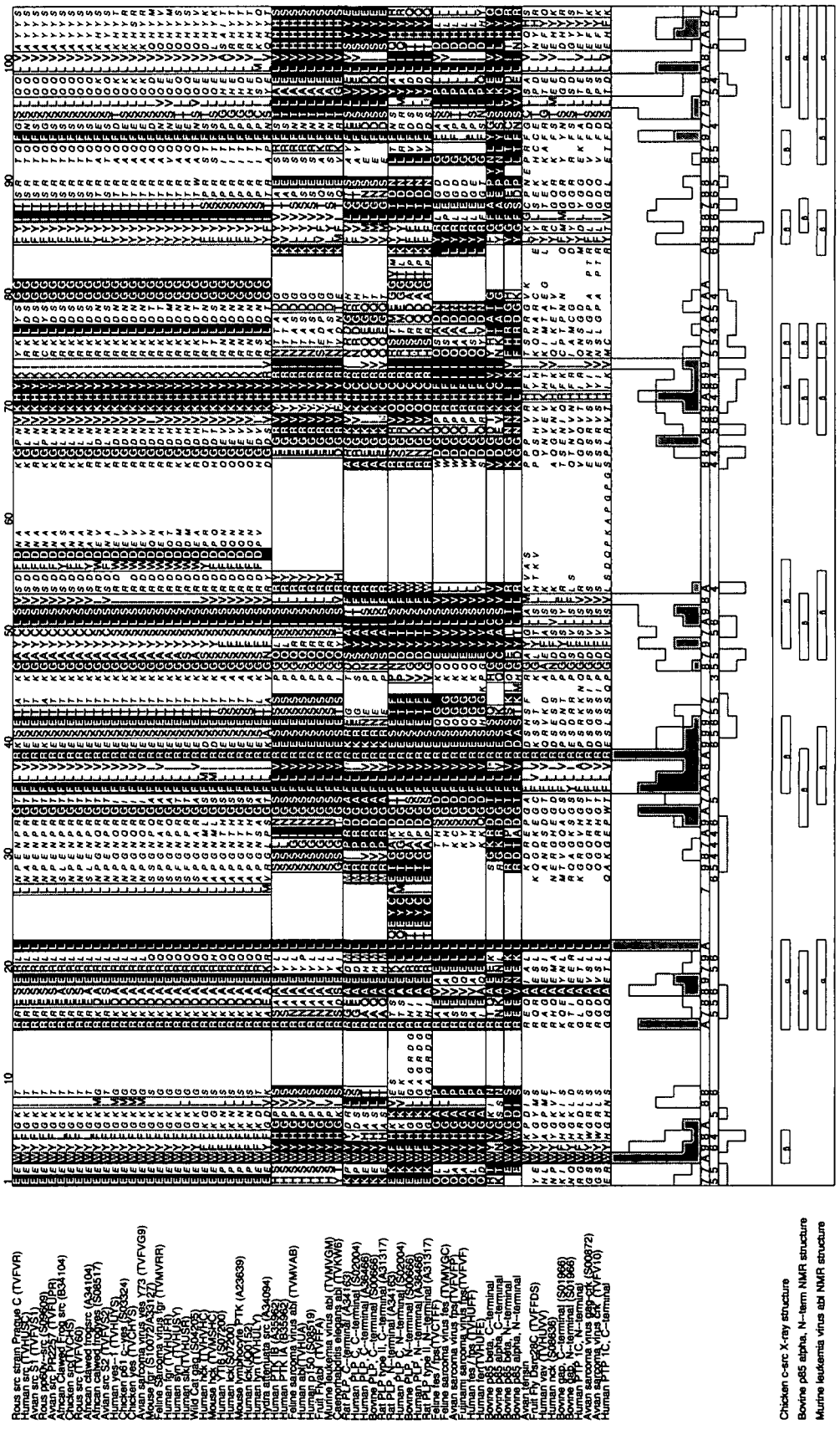


Fig. 6. Conservation analysis of 67 SH2 domains. An alignment of 67 SH2 domains analysed using the general property index (Figure 1b). A key to the shading strategy is given in Figure 3 (see text). The mean pair conservation number for conserved subgroup pairs at each position is reported below the histogram if it is equal to or exceeds the threshold of 7 for the plot. One gap per subgroup was ignored.

(where C = conserved) characteristic of α-helix are clearly visible.

The annexins are a family of proteins that bind phospholipid in a calcium-dependent manner. Annexins consist of a variable N-terminal sequence followed by four or eight repeats, each of ~80 amino acids. Inspection of a multiple sequence alignment of 40 repeats identified the unique features of each repeat family, and located patterns of residue substitution characteristic of the secondary structures (Barton *et al.*, 1991). Figure 5 illustrates the application of hierarchical conservation analysis to a subset of these annexin repeats. Only conserved charges are shown (Figure 5a), and the differences summary clearly locates the position of a change in charge sign (position 31). This charge swap corresponds to the site of an inter-repeat salt bridge (Barton *et al.*, 1991). Additional charge changes are also seen at positions 13, 31, 40 and 68 as listed in the textual summary shown in Figure 5(b). While all these features can be identified by hand inspection of the alignment, the process is laborious and error-prone. The strategy described in this paper reduces the scope for error, allows alternative subgroupings to be investigated rapidly, and provides shading and boxing that is structurally relevant.

AMAS and Alscript are available from the authors.

Acknowledgements

We thank Professor L.N.Johnson for her encouragement and support, and R.B.Russell for his critical reading of the manuscript. C.D.L. is supported by an MRC student award and is a member of Green College, Oxford. G.J.B. thanks the Royal Society for support.

References

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

Barton,G.J. (1993). Alscript: a tool to format multiple sequence alignments. *Protein Engng.*, **6**, 37–40.

Barton,G.J., Newman,R.H., Freemont,P.F. and Crumpton,M.J. (1991). Amino acid sequence analysis of the annexin super-gene family of proteins. *Eur. J. Biochem.*, **198**, 749–760.

Barton,G.J. and Russell,R.B. (1993) Protein structure prediction. *Nature*, **361**, 505–506.

Barton,G.J. and Sternberg,M.J.E. (1987) A strategy for the rapid multiple alignment of protein sequences: confidence levels from tertiary structure comparisons. *J. Mol. Biol.*, **198**, 327–337.

Benner,S. and Gerloff,D. (1990) 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. *Adv. Enzyme Regul.*, **31**, 121–181.

Booker,G.W., Breeze,A.L., Downing,A.K., Panayotou,G., Gout,I., Waterfield,M.D. and Campbell,I.D. (1992) Structure of an sh2 domain of the p85alpha subunit of phosphatidylinositol-3-oh kinase. *Nature*, **358**, 684–687.

Brouillet,S., Risler,J. and Slonimski,P. (1992) Evolutionary divergence plots of homologous proteins. *Biochimie*, **74**, 571–580.

Crawford,I.P., Niermann,T. and Kirchner,K. (1987) Prediction of secondary structure by evolutionary comparison: Application to the alpha subunit of tryptophan synthase. *Proteins: Struct., Funct., Genet.*, **2**, 118–129.

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

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

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

Kabat,E.A. (1976) *Structural Concepts in Immunology and Immunochemistry*, 2nd edn. Holt, Rinehart and Winston, New York.

Overduin,M., Rios,C.B., Mayer,B.J., Baltimore, D. and Cowburn,D. (1992) Three dimensional solution structure of the src homology 2 domain of c-abl. *Cell*, **70**, 697–704.

Parry-Smith,D.J. and Attwood,T.K. (1991). Somap: a novel interactive approach to multiple protein sequence alignment. *Comput. Applic. Biosci.*, **7**, 233–235.

Russell,R.B., Breed,J. and Barton,G.J. (1992) Conservation analysis and secondary structure prediction of the sh2 family of phosphotyrosine binding domains. *FEBS Lett.*, **304**, 15–20.

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

Smith,R.F. and Smith,T.F. (1990) Automatic generation of primary sequence patterns from sets of related protein sequences. *Proc. Natl. Acad. Sci. USA*, **87**, 118–122.

Taylor,W.R. (1986) Classification of amino acid conservation. *J. Theor. Biol.*, **119**, 205–218.

Waksman,G., Kominos,D., Robertson,S., Pant,N., Baltimore,D., Birge,R.B., Cowburn,D., Hanafusa,H., Mayer,B.J., Overduin,M., Resh,M.D., Rios,C.B., and Kuriyan,J. (1992) Crystal structure of the phosphotyrosine recognition domain of sh2 of v-src complexed with tyrosine-phosphorylated peptides. *Nature*, **358**, 646–653.

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

Received on May 23, 1993; accepted on August 5, 1993

Circle No. 18 on Reader Enquiry Card