Protein Sequence Alignments: A Strategy for the Hierarchical Analysis of Residue Conservation

Craig D. Livingstone and Geoffrey J. Barton†

Laboratory of Molecular Biophysics
University of Oxford
Rex Richards Building
South Parks Road
Oxford OX1 3QU
U.K.

†Author to whom correspondence should be addressed.
(e-mail: geoff@biop.ox.ac.uk)
Keywords: multiple sequence alignment, amino acid sequence,annexins, SH2 domains, protein structure prediction.
Published in *Computer Applications in the Biosciences*
Volume 9, Pages 745-756

Contents

1 Abstract

An algorithm is described for the systematic characterisation of the physicochemical 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 sub-groups on the basis of sequence similarity, functional similarity, evolutionary, or other criteria. All pairs of sub-groups are then compared to highlight positions that confer the unique features of each sub-group. 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.

2 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 [Zvelebil *et al.*, 1987, Barton *et al.*, 1991, Russell *et al.*, 1992, Crawford *et al.*, 1987, Benner & Gerloff, 1990]. 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 scrutinised, 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 [Devereux *et al.*, 1984] package 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 & 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 [Kabat, 1976], "entropy" or "variation" by Sander & Schneider [Sander & Schneider, 1991] "information" by Smith & Smith [Smith & Smith, 1990] and "evolutionary divergence" by Brouillet *et al.* [Brouillet *et al.*, 1992]. In contrast, the novel set-based approach described by Taylor [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.* [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 sub-families within a larger multiple sequence alignment.

It is frequently desirable to sub-divide a protein family on the basis of function, origin, sequence similarity or other criteria. Indeed, most multiple alignment methods (e.g. [Barton, 1990, Barton & Sternberg, 1987, Feng & Doolittle, 1987, Higgins & Sharp, 1989]) first compare all sequences pairwise, then automatically cluster the sequences into sub-families 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 rationalise 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 characterising residue conservation (e.g. [Devereux *et al.*, 1984, Parry-Smith & Attwood, 1991,

Taylor, 1986], [Kabat, 1976, Sander & Schneider, 1991, Smith & Smith, 1990, 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 sub-groups 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 characterising conservation on a residue by residue basis within pre-defined sub-families, then between all pairs of sub-families. This hierarchical approach highlights positions that may be responsible for conferring the specific structural and functional properties of the sub-families.

3 System 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.

4 Algorithm

4.1 Quantification of Amino Acid Residue Conservation

We have extended the work of Zvelebil *et al.* [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 1a[1] illustrates a Venn diagram (for details see [Taylor, 1986]) which

[1] figure1.ps

is contained within a boundary that symbolises the universal set of 20 common amino acids (ϵ). The amino acids that possess the dominant properties, hydrophobic, polar and small ($< 60\text{Å}^3$), are defined by their set boundaries. Subsets contain amino acids with the properties aliphatic (branched sidechain non-polar), aromatic, charged, positive, negative and tiny ($< 35\text{Å}^3$). 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 1b[2], 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 oxidised form is generally more useful (as in Figure 1b[3]).

The illustrated Venn diagram (Figure 1a[4]) assigns multiple properties to each amino acid; thus, Lysine has the property hydrophobic by virtue of its long side chain 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[5] illustrates the stages involved in the calculation of conservation numbers for a simplified property index (Figure 2a & b[6]). All of the amino acids are assigned to the universal set (ϵ), which in this simple example, only contains 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

[2]figure1.ps
[3]figure1.ps
[4]figure1.ps
[5]figure2.ps
[6]figure2.ps

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[7]), then tested for the presence of each of the three properties (Figure 2b[8]). This is represented by the columns of entries for each amino acid (Figure 2e[9]). For example, at aligned position 11, the first column in Figure 2e[10] 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 2e[11]. 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 2e[12] will be filled circles and a filled circle appears in Figure 2f[13]. If all amino acids at a position lack the property, then the position shows *negative conservation*; all entries on the row in Figure 2e[14] will be empty circles and an empty circle is seen in Figure 2f[15]. 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 2e[16], the property is labelled as unconserved and a shaded circle is shown in Figure 2f[17].

Two methods are used to quantify conservation at an alignment position using the information stored in Figure 2f[18]. Method 1 is similar to that of Zvelebil *et al.* [Zvelebil *et al.*, 1987] and regards as conserved any property which is either positively or negatively conserved. The number properties obeying this rule (number of filled or empty circles for a position in Figure

[7] figure2.ps
[8] figure2.ps
[9] figure2.ps
[10] figure2.ps
[11] figure2.ps
[12] figure2.ps
[13] figure2.ps
[14] figure2.ps
[15] figure2.ps
[16] figure2.ps
[17] figure2.ps
[18] figure2.ps

2f[19]) is summed to give the conservation number (Figure 2g[20]). In contrast, Method 2 only counts properties which are positively conserved (filled circles in Figure 2f[21]) and gives the conservation numbers shown in Figure 2h[22].

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 1a[23] joins Leu and Arg and crosses 5 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[24]; 10 for Figure 1b[25]).

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 1a[26] share no properties; by Method 2, their conservation number is 0. Asp and Glu in Figure 2a[27] 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 the the maximum number of properties possessed by a single amino acid in the property index.

4.2 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

[19] figure2.ps
[20] figure2.ps
[21] figure2.ps
[22] figure2.ps
[23] figure1.ps
[24] figure2.ps
[25] figure1.ps
[26] figure1.ps
[27] figure2.ps

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 are represent less than $N\%$ of the total at a position may be ignored when calculating conservation values. For example, alignment position 3 in Figure 2[28] is predominantly Asp.

This position would not be recorded as conserved using the charge index due to the presence of a single Asn (1 out of 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[29], position 3). It is the ability to quantify the conservation of amino acids which 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.

4.3 Hierarchical conservation analysis

The procedures described in the previous section are a straghtforward extension of the principles described by Zvelebil *et al.* [Zvelebil *et al.*, 1987] and Taylor [Taylor, 1986]. Here we extend the set based method to identify conserved features of sequence sub-groups within larger protein sequence alignments.

The starting point for hierarchical conservation analysis is the identification of two or more sub-sets 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

[28] figure2.ps
[29] figure2.ps

unique properties of each group.

Figure 3[30] and 4[31] 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[32]. The dendrogram shown at the left of Figure 3[33] shows the overall similarity between the sequences (i.e. not just the 9 residues) and clearly splits the sequences into three sub-groups labelled **A**, **B** and **C**.

Conservation numbers are calculated for each alignment position in each sub-group and a conservation threshold is set. This reference point is used to put each position within a sub-group into one of three classes: (1) Identical positions; (2) conserved positions, where the conservation number is greater than or equal to the threshold; and (3) 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[34], a threshold of between 6 and 8 normally gives the most informative results.

In Figure 3[35], the different classifications using a threshold of 8 are illustrated by shading and font changes. For example, in sub-group **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 sub-group, all pairs of sub-families are compared and conservation numbers calculated for each position in the pairs. In the calculation of conservation for a pair of sub-families, the residues from the pair are considered as members of a single group. C_n is then calcualted, as described above, for the composite group according to which method was chosen. The change in conservation value that occurs when each pair of sub-families is brought together reflects the similarities or differences in physico-chemical properties seen in each sub-group at that position. For

[30] figure3.ps
[31] figure4.ps
[32] figure1.ps
[33] figure3.ps
[34] figure1.ps
[35] figure3.ps

example, at position 7 of sub-families **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**.

In contrast, at position 8 of sub-groups **A** and **C**, in order to "visit" all members of the combined set of amino acids from **A** + **C** (**DEQR**) a minimum of 4 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 sub-group at this position, the properties that are conserved differ between the sub-groups. 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 -ve to +ve charge between groups **A** and **C** may signal increased binding for a -ve 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 sub-families is summarised below the alignment in Figure 3[36]. The conservation values for the pairs of sub-groups are either displayed as similarities or differences according to the rules shown in Table I. The similarity and difference sections are also summarised as histograms.

The hierarchical clustering approach addresses the problem of how to weight the information content of each sequence in 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 information about the minor sequence variation which is responsible for subtle differences in character between sim-

[36]figure3.ps

ilar proteins in a sub-group. This loss is balanced by the ability to detect the more substantial changes in conservation which determine the differences in properties between the separate sub-groups.

5 Implementation

5.1 Text Representation

AMAS accepts command line arguments and provides a detailed textual breakdown of the conservation within a multiple alignment. Figure 4[37] illustrates the AMAS textual analysis that corresponds to the alignment shown in Figure 3[38]. 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 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 sub-groups(s) to which it refers and a list of the residues in each sub-group cited (square brackets). In addition, for positions that do not show identities, the properties conserved at the position, and those which differ are reported. With reference to Figure 4[39]:

- *Identities*: Section 1 lists those sequence positions that are identical across the whole alignment, between pairs of sub-groups and within one sub-group. 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 sub-groups or as three identical individual sub-groups.

- *Conservation of Properties*: Conservation of physico-chemical properties between sub-groups (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 sub-groups share similar properties; (3) pairs of conserved

[37] figure4.ps
[38] figure3.ps
[39] figure4.ps

sub-groups have dissimilar properties; and (4) individual sub-groups are conserved. The properties that are positively conserved between pairs of sub-groups are listed, as are those properties that cause differences between subgroups. For each of a pair of different sub-groups, the percentage of residues that display the differing properties is shown in square brackets.

- *Unconserved*: There are two divisions, the first for single unconserved sub-groups and the second for entirely unconserved alignment positions.

5.2 Graphical Display

The optional graphical representation of results mimics a hand analysis of the alignment using coloured marker pens. In Figure 3[40] the alignment is shown divided into three sub-families. Within the sub-families, at each alignment position, the amino acids are appropriately highlighted. Conserved sub-groups, sub-groups showing identity and positions that show identity across the whole alignment are labelled. Figures 5a[41] and 6[42] illustrate the graphical representation applied to the annexin and SH2-domains.

Three highlighting methods have been explored. Monochrome methods allow grey shading (Figures 5[43] and 6[44]) 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.

6 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 sub-groups and their relative conservation analysed. As with any analytical procedure,

[40] figure3.ps
[41] figure5a.ps
[42] figure6.ps
[43] figure5a.ps
[44] figure6.ps

the strategy is most effective when one has a clear idea of what one is looking for. For example: " What makes sub-group A different from B and C? " , or " Which residues in sub-group D should I change to make D more like A? ". If no clear questions have been defined, then the general property index (Figure 1b[45]) is a useful starting point to highlight patterns of residue conservation. This is illustrated in Figure 6[46] 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[47]). The alignment is divided into eight sub-groups on the basis of overall sequence similarity. Sub-groups 1-7 (numbering from the top) share more than 20% sequence identity, whilst sequences not fitting into one of these sub-groups are collected in sub-group 8. The overall conservation of physico-chemical properties is highlighted by the histogram at the base of the alignment. The upper histogram indicates the normalised 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 sub-groups that show sequence identity. A hand analysis of an alignment similar to that shown in Figure 6[48] correctly identified the location of the core secondary structures, and phosphotyrosine-binding residues [Russell *et al.*, 1992, Barton & 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[49] ([Waksman *et al.*, 1992, Overduin *et al.*, 1992, Booker *et al.*, 1992]). The conservation histograms clearly correspond to the regions of secondary structure, and are helpful in identifying patterns characteristic of $\alpha-$ helix and $\beta-$ strand. For example, at positions 15 and 97 CXXCCXXC patterns (where C=Conserved) characteristic of $\alpha-$ 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 approximately 80 amino acids. In-

[45] figure1.ps
[46] figure6.ps
[47] figure1.ps
[48] figure6.ps
[49] figure6.ps

spection 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[50] illustrates the application of hierarchical conservation analysis to a subset of these annexin repeats. Only conserved charges are shown (Figure 5a[51]), 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 5b[52]. 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 sub-groupings to be investigated rapidly, and provides shading and boxing that is structurally relevant.

AMAS and Alscript are available from the authors.

7 Acknowledgements

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

References

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

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

[50] figure5a.ps
[51] figure5a.ps
[52] figure5b.ps

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

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

[Barton & Sternberg, 1987] Barton, G. J. & 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 & Gerloff, 1990] Benner, S. & 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. Enz. Reg.* **31**, 121–181.

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

[Brouillet *et al.*, 1992] Brouillet, S., Risler, J. & Slonimski, P. (1992). Evolutionary divergence plots of homologous proteins. *Biochimie,* **74**, 571–580.

[Crawford *et al.*, 1987] Crawford, I. P., Niermann, T. & 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 *et al.*, 1984] Devereux, J., Haeberli, P. & Smithies, O. (1984). A comprehensive set of sequence analysis programs for the vax. *Nucl. Acid. Res.* **12**, 387–395.

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

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

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

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

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

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

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

[Smith & Smith, 1990] Smith, R. F. & 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, 1986] Taylor, W. R. (1986). Classification of amino acid conservation. *J. Theor. Biol.* **119**, 205–218.

[Waksman *et al.*, 1992] 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., L., S. & Kuriyan, J. (1992). Crystal structure of the phosphotyrosine recognition domain of sh2 of v-src complexed with tyrosine-phosphorylated peptides. *Nature (London),* **358**, 646–653.

[Zvelebil *et al.*, 1987] Zvelebil, M. J. J. M., Barton, G. J., Taylor, W. R. & 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.

8 Table I

Sub-groups compared			Display C_{A+B} as
A	**B**	**A + B**	Similarity/Difference
$C_A \geq T$	$C_B \geq T$	$C_{A+B} \geq \min C_A, C_B$	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
$C_A < T$	$C_B \geq T$	-	-
$C_A \geq T$	$C_B < T$	-	-
$C_A < T$	$C_B < T$	-	-

Legend to Table I

Pair comparison of conserved sequence sub-groups. 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.

9 Figure Legends

Figure 1[53] Physico-chemical Properties of the Amino Acids

(a) The 20 common amino acids are shown in terms of ten 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 a L to an R in this ten property diagram (see text).

(b) An amino acid property index derived from the Venn diagram in Figure 1a[54] (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.

Figure 2[55] 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).

---

[53] figure1.ps
[54] figure1.ps
[55] figure2.ps

Figure 3[56] 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 1b[57]. 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 program AMAS.

Figure 4[58] Text Representation of Sequence Conservation

With reference to Figure 3[59]. 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 sub-groups are reported; the percentage of residues in each sub-group that have a property is shown in square brackets.

Figure 5[60] Charge Conservation in 40 Annexin Repeats

(a)[61] The pattern of conserved charge in 40 annexin repeats determined using the charge property index described in Figure 2[62]. Only positive property conservation is considered at a conservation threshold of 2, this means

[56] figure3.ps
[57] figure1.ps
[58] figure4.ps
[59] figure3.ps
[60] figure5a.ps
[61] figure5a.ps
[62] figure2.ps

that a sub-group 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 less than 10% of a sub-group position have been screened from the conservation calculation. Identities and conserved positions are identified according to the shading protocol given in Figure 3[63]. 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)[64] Text output accompanying the analysis in Figure 5a[65]. The record format used is identical to that used in Figure 4[66].

[63] figure3.ps
[64] figure5b.ps
[65] figure5a.ps
[66] figure4.ps

Figure 6[67] Conservation Analysis of 67 SH2 Domains

An alignment of 67 SH2 domains analysed using the general property index (Figure 1b[68]). A key to the shading strategy is given in Figure 3[69] (see text). The mean pair conservation number for conserved sub-group 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 sub-group was ignored.

[67] figure6.ps
[68] figure1.ps
[69] figure3.ps