J. theor. Biol. (1986) **119**, 205–218

The Classification of Amino Acid Conservation

William Ramsay Taylor

Laboratory of Molecular Biology, Dept of Crystallography, Birkbeck College, London WC1E 7HX, U.K.

(*Received* 20 *September* 1985, *and in revised form* 1 *November* 1985)

A classification of amino acid type is described which is based on a synthesis of physico-chemical and mutation data. This is organised in the form of a Venn diagram from which sub-sets are derived that include groups of amino acids likely to be conserved for similar structural reasons. These sets are used to describe conservation in aligned sequences by allocating to each position the smallest set that contains all the residue types brought together by alignment. This minimal set assignment provides a simple way of reducing the information contained in a sequence alignment to a form which can be analysed by computer yet remains readable.

1. Introduction

The high level of expertise current in nucleic acid research has led to the revelation of large numbers of protein sequences and the ability specifically to alter these in a controlled way. New sequences often exhibit a close homology with proteins which have had their structure determined crystallographically and using advanced computer graphic facilities it is possible theoretically to alter the amino acid side chains of the known structure to represent the sequence of unknown structure (e.g. Blundell *et al.*, 1983). The new techniques are used increasingly to design proteins with altered properties using the methodology of site-specific mutagenesis.

These activities require a good understanding of the basic principles of protein structure and, in particular, it is necessary to anticipate the structural effect of introducing a new amino acid into a known structure. This assessment is often based on the likelihood matrix of amino acid mutabilities derived by Dayhoff *et al.* (1972, 1978) or on the number of nucleotide base changes required to effect the substitution (Fitch, 1966). Such measures, however, ignore aspects of the substitution that are relevant to the local structural environment or known function of the residue and in building hypothetical structures and designing new mutants it is these details which are important.

In this paper I consider measures of amino acid relatedness in common use with the aim of extracting from them features which will best assist a protein engineer faced with the problem of making a mutation and assessing a sequence alignment. These features are represented as groupings of amino acids (sets) which is a form that retains a descriptive quality yet allows quantitative manipulations using the formalism of set logic.

0022-5193/86/060205+14 $03.00/0

2. Measures of Amino Acid Relatedness

(A) MUTATION DATA

For every pair of the 20 naturally occurring amino acids Dayhoff *et al.* (1972) have determined the probability (or odds) that the mutation will occur in either direction. This matrix was most clearly presented by Sander & Schulz (1979) (see also Schulz & Schirmer, 1979) in a form where the entries have been ordered to bring frequently exchanging amino acids together. Even in its ordered form it is still difficult fully to appreciate the information contained in Dayhoff's matrix. However, using the technique of multi-dimensional scaling, French & Robson (1983) reduced the matrix to a two dimensional plot, in which frequently exchanging amino acids are closest together (see Fig. 1(a)). A similar diagram (Fig. 1(b)) was produced by minimising the deviation from a 2-D structure in which the entries of Dayhoff's matrix represent the inverse of ideal target distances between pairs of amino acids (Taylor, 1981). Both these figures are roughly elliptical and projecting the amino acids onto the circumference of each ellipse produces an even simpler representation with no great loss of information. In this form the long axis of the ellipse corresponds to molecular volume while the short axis corresponds to hydrophobicity. These simplifications are presented in Figs 1(c) and 1(d). The cyclic order of amino acids obtained from this simplification is almost the same as that derived by Swanson (1984) (see Fig. 1(e)).

All the above representations of Dayhoff's matrix indicate that it can be largely accounted for by the effect of only two determining factors; hydrophobicity and size. This remarkable observation must obviously dominate any attempt to codify amino acid conservation.

(B) PHYSICAL DATA

Physico-chemical properties

The wide variety of physico-chemical properties manifest in the amino acid side-chains has been thoroughly considered by Sneath (1966). These have been adopted, or deduced independently, by others and used as a basis for considering relatedness between protein sequences. McLachlan (1972) summarizes these relationships by assigning a value to each transition between pairs of amino acids. These scores are presented graphically in Fig. 2 using the circle of amino acids derived from Dayhoff's matrix as a frame on which lines connect related amino acids. All high scoring transitions and most other connections on the graph are local, indicating agreement between chemistry and mutability. The less local connections mainly join hydrophobic residues, which, with the exception of proline, are relatively adjacent on the less idealised representations of Dayhoff's matrix (Fig. 1). The missing links are, perhaps, more revealing: there is only a weak link between Tyr and Trp which are strongly tied in Dayhoff's mutation matrix and there is no connection between the adjacent negatively charged residues and Gly and Pro.

Measurement of properties cannot determine a common scale without reference to protein sequences and structures. The idea of such a scaling can be appreciated

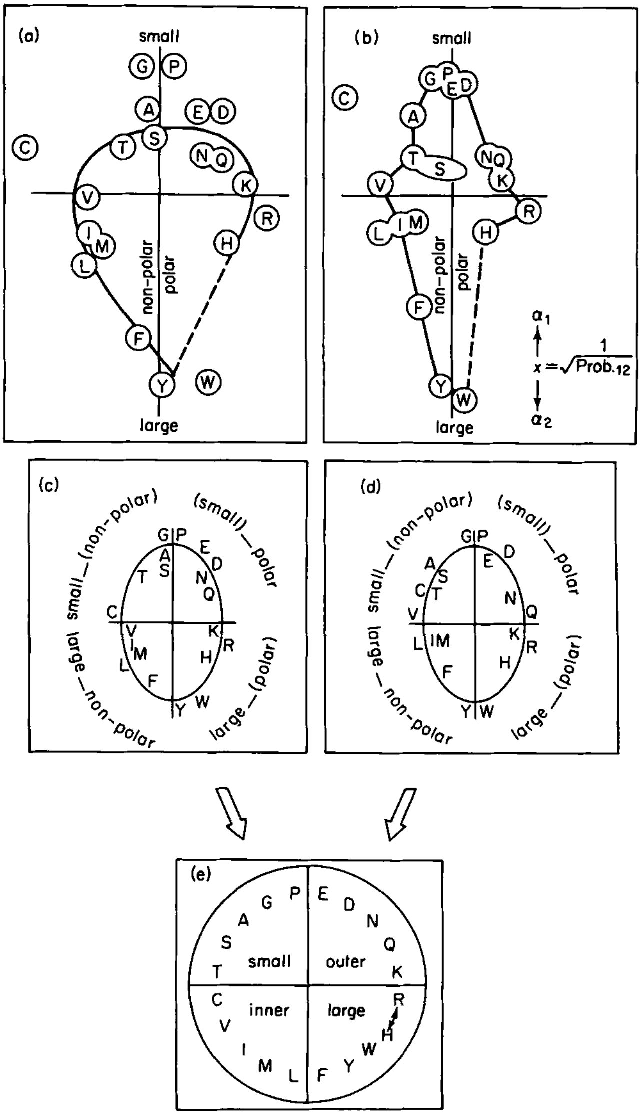


FIG. 1. Representations of Dayhoff's mutation odds matrix. (a) Projection of the matrix by multidimensional scaling (adapted from Robson & French (1983)). Amino acids which are close together exchange frequently. (b) The equivalent diagram to (a) produced by pseudo-energy minimization to a point at which each amino acid lies in a position which gives the minimum sum of squares over the distance equation shown. (c) and (d) Idealizations of (a) and (b) respectively. The properties associated with each quadrant are indicated with the property of lesser importance bracketed. (e) A further idealization of the two plots which are constrained to a circle. Ambiguities in the cyclic order have been reconciled by consideration of both original plots. The resulting order agrees with that obtained by Swanson (1984) except for the exchange of Arg and His as indicated by an arrow. (This probably arises from Swanson's use of the Dayhoff (1978) revised matrix). Swanson's nomenclature for the quadrant characteristics is also indicated. The one letter and three letter codes for the amino acids are as follows: Gly G, Ala A, Val V, Leu L, Ile I, Ser S, Thr T, Asp D, Glu E, Asn N, Gln Q, Lys K, His H, Arg R, Phe F, Tyr Y, Trp W, Cys C, Met M, Pro P, asx b, glx z.

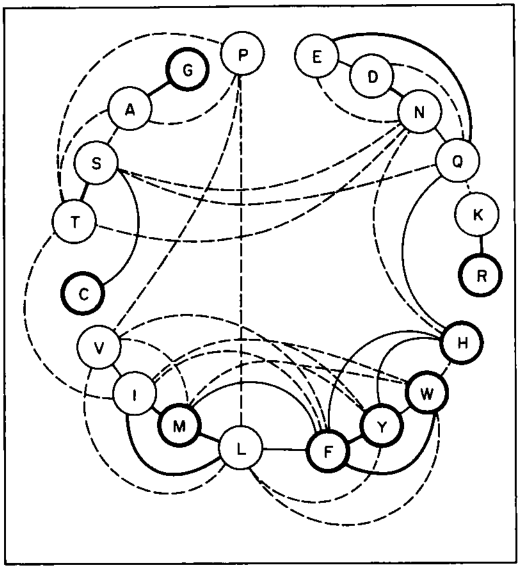


FIG. 2. Chemical relatedness of the amino acids as quantified by McLachlan (1972) displayed on the circle of residues derived from Dayhoff's matrix (Fig. 1(e)). The degrees of expected conservation are indicated as follows (with McLachlan's score in brackets). Strong conservation of type = heavy circle (6); Moderate conservation of type = lighter circle (5); Strong conservation of properties = heavy line (3); Moderate conservation of properties = finer line (2); Weak conservation of properties = broken line (1). Most connections are local (i.e. chemistry agrees with mutation data) with the exception of long links from hydrophobic residues to Pro and Ser & Thr to Asn & Gln and the absence of links between Gly & Pro and Glu & Asp.

from the relative lengths of the ellipses derived from Dayhoff's matrix (see Figs. 1(c) and 1(d)) in which the longer axis is associated with change of size indicating that this, on average, is dominant over hydrophobicity. Grantham (1974) scaled three properties to McLachlan's (1971) matrix of amino acid substitution frequencies. These were volume, polarity and the fraction of carbon in the side-chain (composition). He found the most dominant property was again volume followed by polarity.

Secondary structure propensities

From statistical analysis of the sequences of proteins of known structure, propensities to adopt a secondary structure have been determined for each amino acid (e.g. Chou & Fasman, 1974; Garnier *et al.*, 1979). These preferences have been used to account for clusters of amino acids which are unexpected on a physico-chemical basis. A clear example is the close association of G, P, D and E (see Fig. 1). These associate because of a propensity to lie in sharply turning regions on the surface of the protein. Gly, because of the flexibility it imparts to the local chain; Pro because of the built in turn configuration created by its back-bonding side chain; and Asp and Glu because of a requirement to expose their charges to solvent. Robson & French (1983) indicate other instances including the close association of

Glu and Ala both of which favour α-helical structure, and the tight association of L, I, M, V and to a lesser extent, F and Y which tend towards β-structure.

3. Venn Diagram of Amino Acid Sets

The idea of using a Venn diagram to represent the different relationships among the amino acids was adopted from Dickerson & Geis (1969) and extended to incorporate some of the observations discussed above. The overall layout of the diagram (Fig. 3(a)) was based on the 2-D arrangement derived from Dayhoff's mutation matrix. Amino acids were then displaced (by as little as possible) from this arrangement to form groups of residues related by common physio-chemical properties.

(A) SIZE AND HYDROPHOBICITY

The major sets group the amino acids by size and hydrophobicity: both properties which were seen to dominate the structure of Dayhoff's Matrix. Two overlapping sets were used to describe hydrophobicity. One was defined as all amino acids which have a polar group in their side-chain and is referred to as *polar.* The second group is less well defined and contains the amino acids which were considered to be hydrophobic. This set contains some amino acids which have polar side-chains. These consequently lie in the intersecting region of the two sets which can be considered the set of aminò•acids which are ambivalent to water. The inclusion of Lys in this set is justified by its long aliphatic side-chain which has been observed (Cohen *et al.*, 1982) to extend from a buried location and expose the terminal charge to solvent.

The location of Pro in Fig. 1 conflicts with its hydrophobic character. It was, thus, left unclassified by the two sets *hydrophobic* and *polar.* Similarly, Gly is often considered to lack a side-chain and be consequently unclassifiable in hydrophobic terms (e.g. Rees & Sternberg 1984). However, as Gly is often found buried in the interior of proteins it was classified as hydrophobic.

The volume of the side-chain was considered to be sufficiently important to justify classification by two sets. The larger of these, called *small,* contains the nine smallest amino acids by side-chain volume (Klapper, 1971) each less than 60 Å^3. A subset of this, called *tiny,* includes the four acids with less than three (non-H) side-chain atoms all of which are smaller than 35 Å^3.

The relationship of Cys to the sets defined above is rather ambiguous. Although its side-chain has only two atoms, the sulphur atom is relatively large, placing it on the *Tiny-Small* borderline. Its classification is further complicated by the occurrence of the sulphur in two oxidation states. The reduced form contains a polarizable S-H bond which suggests a similarity to Ser (O-H) and with which it is associated in McLachlan's table (Fig. 2). On formation of a disulphide bond, however, this property is lost, placing the residue more firmly in the hydrophobic camp. The associated loss of conformational freedom is difficult to assess but may be associated with an effective increase in volume as the linked residue cannot accommodate as easily to structural fluctuations. Poor packing in the hydrophobic core of the

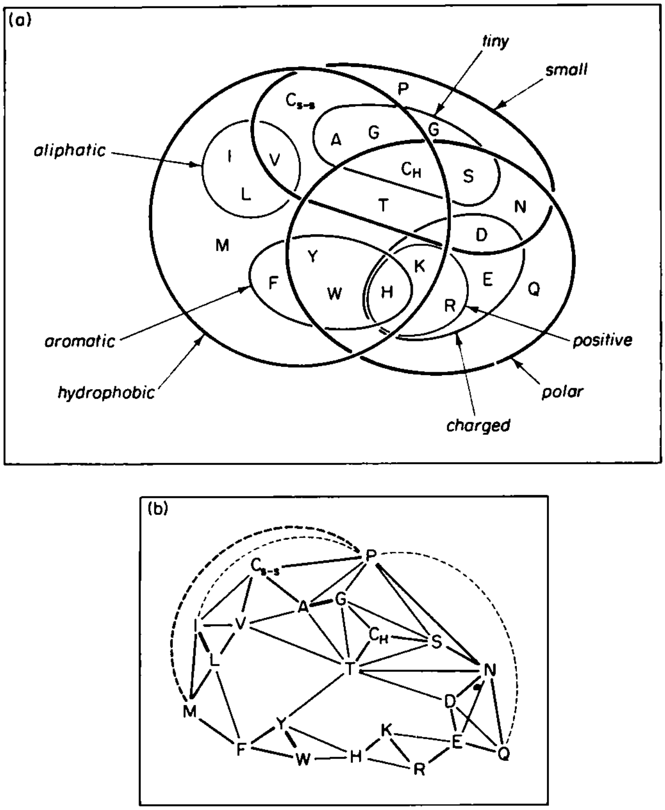


FIG. 3. (a) The Venn diagram shows the relationship of the 20 naturally occurring amino acids to a selection of physio-chemical properties which are important in the determination of protein tertiary structure. The diagram is dominated by properties relating to size and hydrophobicity. The amino acids are divided into two major sets, one containing all amino acids which contain a polar group (*polar*) and a set which exhibit a hydrophobic effect (*hydrophobic*). A third major set, *small*, is defined by size and contains the nine smallest amino acids. Within this is an inner set of smaller residues, *tiny*, which have at most two side-chain atoms. The location of Cys is ambiguous as the reduced form (CH) has similar properties to serine, while the oxidized form (Css) may be more equivalent to Val. Other sets include full-charge (referred to as *charged*) which contains the subset *positive* (negative is defined by implication) and *aromatic* and *aliphatic*. The latter set is not as general as the name implies and contains only residues containing a branched aliphatic side-chain. Because of its unique backbone properties, proline was excluded from the main body of the diagram. An equivalent exclusive position is suggested for Gly by a small G. (b) An alternate representation of the relationships. A network of amino acids is formed by connecting those which differ by no more than two properties in the Venn diagram. Pairs which share the same subset are connected by a heavy bar, those with only one different property by a bold line and those which differ by two properties by a fine line. To improve clarity a few of the latter connections are omitted. Many of these only connect to one of an unresolved pair (e.g. I and L), and it can be assumed that the connection is made to both. Because of its unique position, Pro is able to make quite long range connections some of which are indicated by broken lines.

immunoglobulins near the intra-molecular disulphide link is suggestive of this (Cohen *et al.*, 1981, Taylor, 1981). Considering all these aspects, two locations are suggested for Cys on the Venn diagram. However, a location in the sub-set containing Thr is also possible.

(B) OTHER SETS

The remaining set allocations are based on obvious physico-chemical properties. These include *aromatic* (ring containing side-chains) and *aliphatic*. The latter set, however, includes only amino acids with branched aliphatic side-chains and largely reflects the frequency with which this type of residue is found in β-pleated sheet structure.

The set of *charged* amino acids contains only those which are normally (or often) fully ionized. The subset *positive* is included, with *negative* defined by implication.

For simplicity, as few sets as possible were introduced, yet even these produce almost complete segregation of the amino acids with only Y–W, I–L and A–G grouped in the same sub-sets. Additional sets can easily be imagined but these often only create further distinction between residues which are already segregated. However, an important property not well represented is hydrogen-bonding ability. To distinguish the sets of hydrogen-bond donors and acceptors on the Venn diagram (Fig. 4) would greatly reduce clarity they are thus indicated separately in Figs 4(c) and 4(d).

(C) NETWORK REPRESENTATION

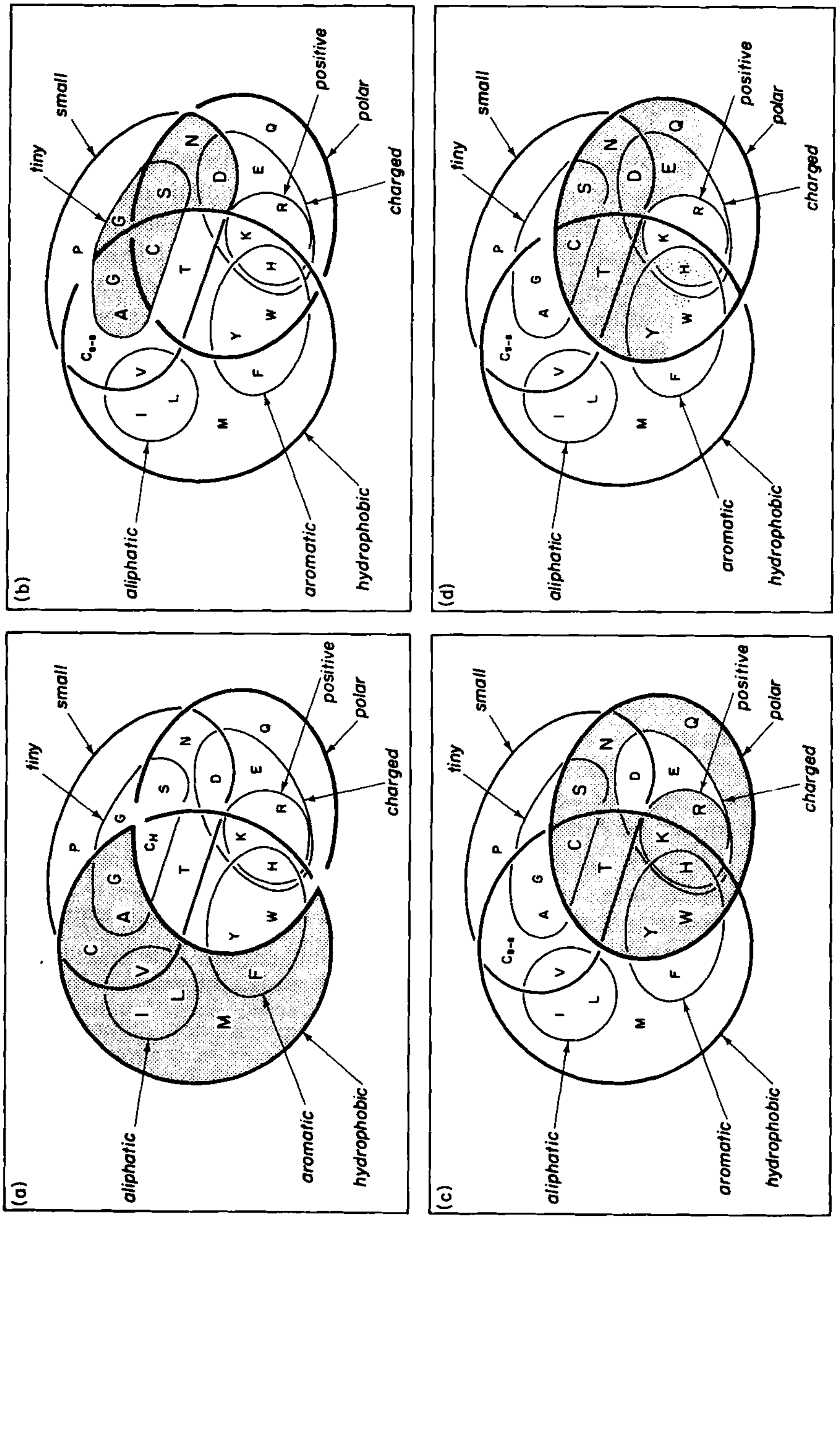
A useful representation of the Venn diagram can be made by removing the set boundaries and connecting adjacent residues to form a network (Fig. 3(b)) in which no connected pair differ by more than two properties. In this form the structure is more easily compared to other representations and the circular arrangement of amino acids corresponding to Fig. 1 is readily apparent.

The most significant deviation from the form of Dayhoff's matrix is the separation between the negatively charged amino acids and Pro and Gly. This feature corresponds more to the physico-chemical relationships defined by McLachlan (Fig. 2). However, leaving Pro (and perhaps Gly) unclassified with respect to hydrophobicity allows connections to be made with both hydrophilic (S, N, Q) and hydrophobic residues (F, M, L, I, V and A). Some of these longer connections are indicated in Fig. 3(b) by broken lines.

It is also possible to formally reduce the Venn diagram (or network) to a tree structure of the type described by Jiménez-Montaño (1984). However, unless the corresponding tree contained multiple amino acid entries information in the Venn diagram would be lost.

4. Minimal Set Assignment

The Venn diagram (Fig. 3(a)) represents a compromise between mutation data and chemistry. A similar compromise might have been achieved in a less graphical way by calculating a distance matrix of the type described by Grantham (1974). Such tables of relatedness are useful for considering amino acid changes without regard for the local environment in the protein structure in which the change occurs (e.g. exposure to solvent, secondary structure, etc.) and, consequently, can be applied uniformly along any pair of sequences to produce an overall measure of relatedness


(a)
(b)
(c)
(d)
tiny
small
positive
polar
charged
aliphatic
aromatic
hydrophobic
P
G
A
C
C_H
C_{S-S}
S
N
D
T
E
Q
K
R
H
V
I
L
M
F
Y
W

between them. With greater knowledge of protein structures, however, it is increasingly common that substitutions are considered in a sequence of known structure. This requires a new approach to assessing the substitution odds for a given structural environment. The use of tables of relatedness, moreover, cannot be easily adapted to tackle this problem as every environment would require its own table which must be previously calculated from substitutions occurring in that environment. This would be a very useful set of tables to have but for present work a more flexible approach is required.

With the vast increase in known protein sequences it is now common to have many related sequences to compare and not simply a pair. This type of problem can, of course, be tackled using a distance measure for all or part of a sequence and compiling a table of homology for each pair of sequences. The result, however, obscures the qualitative description of the conservation and workers currently interested in this type of problem are often nucleic-acid biochemists who have little time for numerical abstractions.

With these problems in mind, a method was devised to use the sets defined by the Venn diagram (Fig. 3(b)) to describe sequence alignments. This was, simply, to find the smallest set or sub-set which includes all the amino acid types brought together by alignment. The resulting set will be referred to as the minimal set. Applying this operation over the whole length of the aligned sequences produces a qualitative description of the conservation at every point. Furthermore, the number of amino acids which constitute the assigned set give a measure of the degree of conservation at that point ranging from one (for absolute conservation of type) to twenty (the set of all amino acids).

Despite the few sets included in the Venn diagram, the number of possible subsets is very large and includes many which have little physical meaning. For example; it is difficult to imagine a structural environment or function which would promote the conservation of the set formed by the union of *aliphatic* and *positive* (V, I, L, H, K, R). A list of almost seventy sets and subsets has, thus, been compiled which might be maintained by structural selection pressures. These naturally consist of the union and intersection of sets which overlap in the Venn diagram (graphical examples of two of these are shown in Figs 4(a) and (b)). One set which is an exception to this was created to reflect the conservation of Gly, Pro and the negatively

FIG. 4. (a) and (b) indicate how two subsets are derived from the sets indicated on the Venn diagram (Fig. 3(a)). (a) is formed by the amino acids that are *hydrophobic* but not *polar* (or in set nomenclature: *hydrophobic* $\wedge$ $\sim$*polar*, where $\sim$ indicates negation and $\wedge$ indicates set intersection). In the alternative nomenclature defined in the legend to Fig. 5 this becomes *hydrophobic* non-*polar* but as it is a commonly assigned set it is given the "trivial" name of *very-hydrophobic*. (b) illustrates a more complex derivative sub-set formed by *tiny* $\cup$ (*small* $\wedge$ (*polar* $\wedge$ $\sim$*hydrophobic*)), where $\cup$ indicates set union. In Fig. 5 *polar* $\wedge$ $\sim$ *hydrophobic* is given the trivial name of *hydrophylic* and union is indicated by . or., producing the more comprehensible name of *tiny*. or. *small_hydrophylic*. (c) and (d) indicates the two sets of hydrogen-bond donors (c) and acceptors (d). Their definitions were taken from Baker & Hubbard (1984). It should be noted, however, that hydrogen bonds involving Cys are rare and that Met might possibly receive a hydrogen-bond. Also, the classification of His is complicated by its frequent change of ionisation state as only the unprotonated state can receive a hydrogen bond. For clarity, and as there is still some uncertainty in these sets, they were not included in the main Venn diagram despite their obvious structural importance.

```
29 'POSITIVE'                                       R K H
30 'CHARGED'                                        D E R K H
   'CHARGED_non-H'                                  D E R K
   'Negative' (CHARGED_non-POSITIVE)                D E
   'Hydrophylic_non-POSITIVE'                   b z S N D E Q
   'Hydrophylic' (POLAR_non-HYDROPHOBIC)        b z S N D E Q R
   'CHARGED.or.Hydrophylic'                     b z S N D E Q R K H
   'CHARGED.or.Hydrophylic.or.P'                b z S N D E Q R K H P
   'POLAR_non-AROMATIC.or.CHARGED.or.P'         b z P T S N D E Q R K H
40 'POLAR'                                      b z T S N D E Q R K H W Y
   'POLAR.or.P'                                 b z P T S N D E Q R K H W Y
   'POLAR_non-AROMATIC.or.CHARGED'              b z T S N D E Q R K H
   'POLAR_non-AROMATIC_non-POSITIVE.or.P'       b z P T S N D E Q
   'POLAR_non-AROMATIC_non-POSITIVE'            b z T S N D E Q
   'SMALL_POLAR.or.P'                             b P T S N D
   'SMALL_POLAR'                                  b T S N D
   'SMALL_Hydrophylic'                            b S N D
50 'TINY'                                           A G S
   'TINY.or.SMALL_POLAR'                          b A G T S N D
   'TINY.or.SMALL_POLAR.or.P'                     b P A G T S N D
   'TINY.or.Negative_Hydrophylic.or.T'          b z A G T S N D E Q
   'TINY.or.Negative_Hydrophylic.or.T.or.P'     b z P A G T S N D E Q
  ✓'TINY.or.POLAR_non-AROMATIC'                 b z A G T S N D E Q R K
   'TINY.or.POLAR_non-AROMATIC.or.P'            b z P A G T S N D E Q R K
   'TINY.or.POLAR'                              b z A G T S N D E Q R K H W Y
   'SMALL_non-P.or.POLAR'                       b z V C A G T S N D E Q R K H W Y
60 'SMALL.or.POLAR'                             b z P V C A G T S N D E Q R K H W Y
   'SMALL_non-P.or.POLAR_non-AROMATIC'          b z V C A G T S N D E Q R K
   'SMALL.or.POLAR_non-AROMATIC'                b z P V C A G T S N D E Q R K
   'SMALL_non-P.or.Hydrophylic'                 b z V C A G T S N D E Q R
65 'SMALL'                                        b P V C A G T S N D
   'SMALL_non-P'                                  b V C A G T S N D
   'SMALL_HYDROPHOBIC.or.TINY'                      V C A G T S
   'SMALL_HYDROPHOBIC'                              V C A G T
70 'SMALL_non-POLAR_non-P'                          V C A G
   'SMALL_non-POLAR'                                V C A G P
   'SMALL_non-Hydrophylic'                          V C A G P T
   'ALIPHATIC.or.SMALL_non-Hydrophylic'             L I V C A G P T
   'ALIPHATIC.or.SMALL_non-POLAR'                   L I V C A G P
   'ALIPHATIC.or.SMALL_HYDROPHOBIC'                 L I V C A G
   'ALIPHATIC'                                      L I V
80 'ALIPHATIC.or.Large_non-POLAR'                   F M L I V
   'Very-hydrophobic' (HYDROPHOBIC_non-POLAR)       F M L I V C A G
   'Very-hydrophobic.or.P'                          P F M L I V C A G
   'Very-hydrophobic.or.T'                          F M L I V C A G T
   'Very-hydrophobic.or.T.or.P'                     P F M L I V C A G T
   'Very-hydrophobic.or.T.or.K'                     F M L I V C A G T K
   'Very-hydrophobic.or.SMALL_non-P.or.K'         b F M L I V C A G T K S N D
   'Very-hydrophobic.or.SMALL.or.K'               b F M L I V C A G T K S N D P
   'HYDROPHOBIC.or.SMALL'                         b H W Y F M L I V C A G T K S N D P
   'HYDROPHOBIC.or.SMALL_non-P'                   b H W Y F M L I V C A G T K S N D
90 'HYDROPHOBIC'                                    H W Y F M L I V C A G T K
   'HYDROPHOBIC.or.P'                               H W Y F M L I V C A G T K P
   'AROMATIC.or.Very-hydrophylic'                   H W Y F M L I V C A G
   'AROMATIC.or.ALIPHATIC.or.M'                     H W Y F M L I V
   'AROMATIC.or.M'                                  H W Y F M
95 'AROMATIC'                                       H W Y F
   'Large_non-Negative'                             R K H W Y F
   'Large_POLAR'                                  z Q E R K H W Y
   'Large_non-ALIPHATIC'                          z Q E R K H W Y F M
   'Large' (non-SMALL)                            z Q E R K H W Y F M L I
```

charged amino acids in the bend regions of protein structures. Other minor sets consisting of closely related pairs, including Ser and Thr, Phe and Tyr, and Arg and Lys, were also added. All these sets are defined in Fig. 5 where they are described using a nomenclature adapted from set logic.

A few example applications of the minimal set assignments to aligned sequence fragments are shown in Fig. 6. To give some impression of how set assignment varies with the number of aligned sequences and the overall homology of the sequences, a progression is shown from a few closely related sequences of a particular immunoglobulin domain to an extended alignment of the same domain (Fig. 6). In these it is clear that conservation is maintained mainly in the regions of secondary structure. This observation can be quantified by plotting the set size at each position in the alignment. In Fig. 7 this is done for different numbers of closely related sequences of the immunoglobulin κ-chain light-variable domain. Of these sequences the Bence-Jones protein REI has a known crystallographic structure (Epp *et al.*, 1975). The strands of β-structure found in this protein lie in two sheets which stack together like a sandwich (Cohen *et al.*, 1981). One side of each sheet is buried while the other is exposed to solvent. As the amino acid side-chains in a β-strand alternately point to either side of the sheet they are consequently alternately buried and exposed to solvent. This structure is reflected in the degree to which they are conserved in Fig. 7 where alternately conserved and mutable positions can be seen in the β-strand regions. The effect is most clearly seen in strands which do not lie on the edge of the β-sheet. In these the conserved positions are generally hydrophobic.

Plotting the degree of conservation with increasing numbers of aligned sequences revealed the interesting observation that many positions rapidly acquire a degree of conservation that is unchanged by the addition of further sequences to the alignment. Conservation is rapidly lost on aligning the first 20 sequences in order of decreasing homology to REI (see Fig. 6(b)) but the addition of another 50 sequences to the alignment causes little alteration of the conservation profile.

FIG. 5. Intersection and union of the sets defined in Fig. 6 can produce a vast number of amino acid combinations. Those which, on an intuitive basis, seem most relevant to protein structure are defined below. These are generally unions and intersections of adjacent sets and thus produce groups of amino acids which share similar properties. Many of the sets have effectively two entries, one with proline and one without. The nomenclature used was chosen to be more readable than standard set notation and uses . or. as the inclusive "or" to represent set union. Intersection of two sets is represented by the underline character "_". Negation is indicated by the prefix "non-" and applies only to the set to which it is attached. This produces recognisable phrases such as "non-POLAR" and "SMALL_HYDROPHOBIC". Occasionally a commonly used set is given a "trivial" name: for example, the set of CHARGED_non-POSITIVE is referred to as "Negative" and the set of HYDROPHOBIC_non-POLAR as "Very-hydrophobic". Some sets which have rather long formal names are simplified by reference to individual amino acids using the one letter code, for example, the set of all hydrophobic residues excluding polar-aromatics is simply called "Very-hydrophobic. or. T. or. K". The ambiguous residue codes asx (b) and glx (z) are included when the two residues they represent occur in the set. They do not, however, count when the number of members in the set is considered.

5. Conclusions

In the rapidly developing field of protein engineering it is important to have a measure of amino acid conservation that can be applied to several homologous sequences of which at least one has a known tertiary structure. General measures of amino acid conservation, such as Dayhoff's likelihood matrix are best suited to situations where there is no structural information about the sequences. Their use becomes limiting when applied to local regions of the protein sequence where, for structural reasons, the mutational freedom of a particular residue may be greatly restrained. With knowledge of the local structure, however, it is possible to analyse these restrictions and use them predictively. An example of loss of information by averaging, which is apparent in Dayhoff's matrix, is the resistance of cyst(e)ine to mutation. This, undoubtedly, arises from the evolutionary need to conserve disulphide bonds, but such a restraint does not apply in the reducing intracellular environment and to apply it to the comparison of the sequences of cytoplasmic proteins is, therefore, misleading.

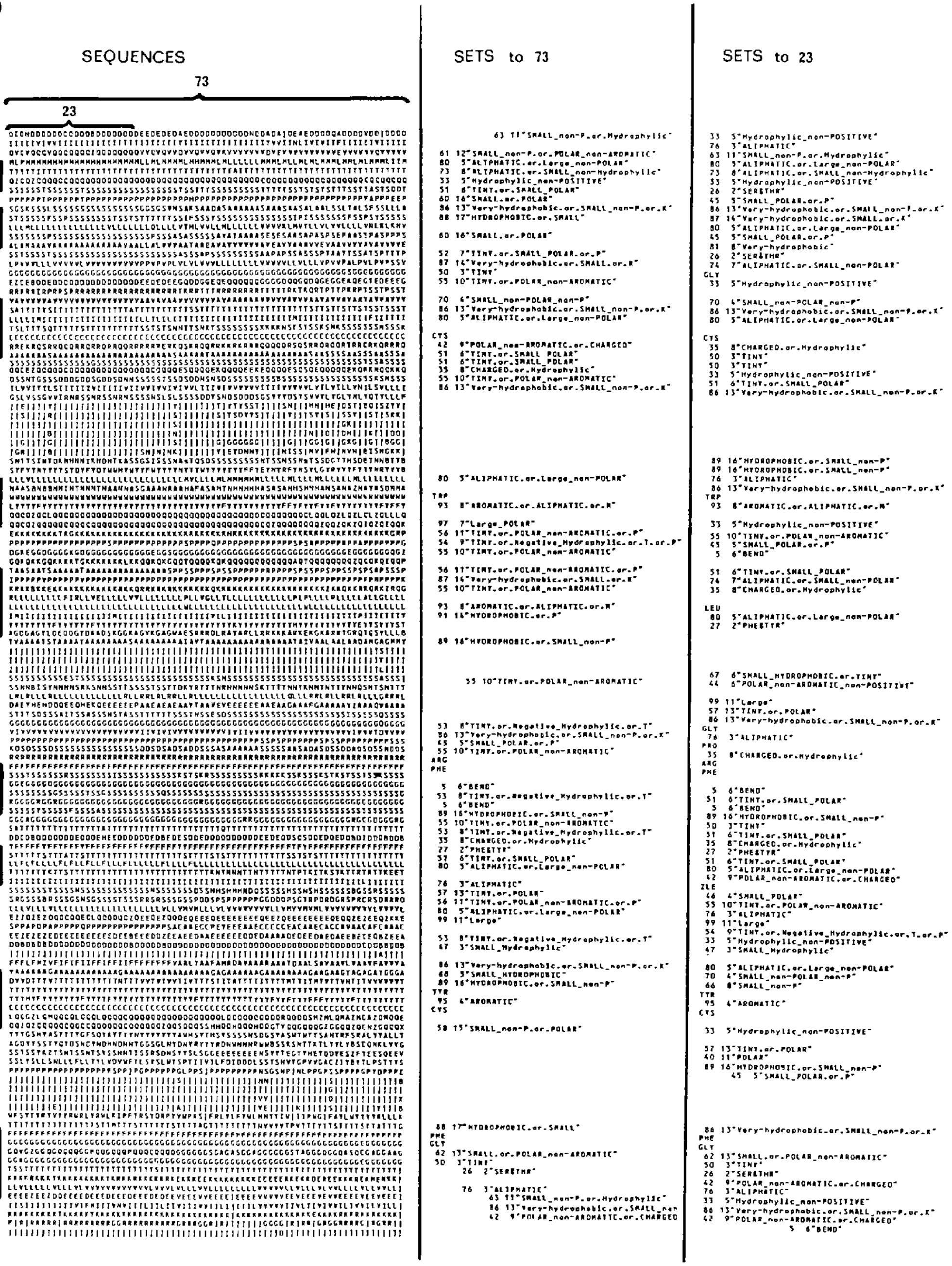
The classification of amino acids defined above, and its use in describing sequence alignments, allows the type of conservation observed in structural "micro-environments" to be rigorously quantified. The important aspect of the approach is that not only can the degree of conservation be measured, but the qualitative aspect of the conservation is also measured. Together these measures capture virtually all the useful information that can be extracted from a number of aligned sequences. Such information will be of use in designing new mutants as a protein engineer can analyse a sequence alignment to find, for every position, the range of possible amino acid changes that might be acceptable. On a wider front, the approach is being applied to the analysis of residue conservation in well defined structural motifs

FIG. 6. (a) Alignment of the variable domains of the immunoglobulin kappa-chains (light-variable domain) found in the PIR database. The sequences run down the page and are represented in the one letter amino acid code (insertions are indicated by a bar "|"). The numbers identify the position in the sequence of REI and next to these, heavy bars indicate regions of β-structure. To the right of the sequences the amino acid set (see Fig. 5) which best describes the alignment is indicated both by name and number. This description is indented in proportion to the number of insertions. The number just to the left of the set description is the number of amino-acids which constitute the set. This gives a measure of specificity and is the number plotted in Fig. 7. Absolutely conserved positions are indicated by the residue name only. The set assignments are divided into two regions. The outer assignments derive from the 23 sequences with greatest homology to REI while the inner is derived from the entire 73 sequences. The sequence names and their relative homology can be found in (b) which is a matrix of homology between every pair of sequences. The homology is calculated as a percentage of residue identity matches over all matched pairs in two given sequences and is entered in the matrix at a position cross-referenced by the two sequence names. For clarity, homologies over 70% are filled solid, those in the 60s are a dot, those in the 50s a "5" and those below 50% are blanked out. The entries in the matrix have been ordered such that most similar sequences tend to be adjacent on the diagonal. This is achieved by minimising the second moment of homology about the diagonal; i.e.

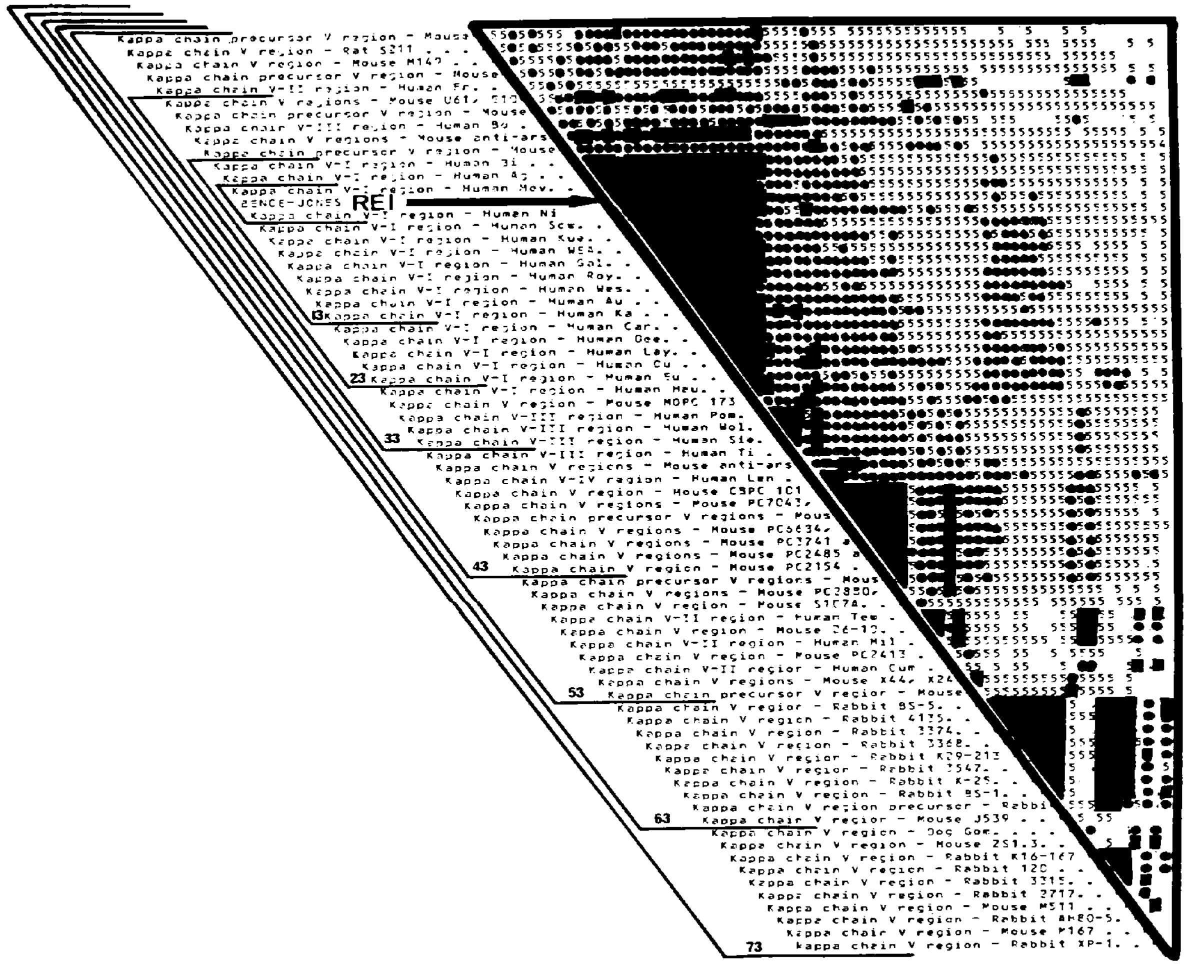
$$\sum_{i=1}^{N} \sum_{j=1}^{N} H_{ij}(i-j)^2 \rightarrow \min$$

where H_{ij} is the percentage homology between proteins at positions i and j and N is the number of proteins. The increments of ten sequences each corresponding to a plot in Fig. 7 are indicated.

(a)



(b)


Kappa chain precursor V region - Mouse
Kappa chain V region - Rat S211 . . .
Kappa chain V region - Mouse M142 . .
Kappa chain precursor V region - Mouse
Kappa chain V-II region - Human Fr. .
Kappa chain V regions - Mouse U61, S13
Kappa chain precursor V region - Mouse
Kappa chain V-III region - Human Bo .
Kappa chain V regions - Mouse anti-ars
Kappa chain precursor V region - Mouse
Kappa chain V-I region - Human Bi . .
Kappa chain V-I region - Human Ag . .
Kappa chain V-I region - Human Mev. .
BENCE-JONES REI
Kappa chain V-I region - Human Ni
Kappa chain V-I region - Human Scw. .
Kappa chain V-I region - Human Kue. .
Kappa chain V-I region - Human WEA. .
Kappa chain V-I region - Human Gal. .
Kappa chain V-I region - Human Roy. .
Kappa chain V-I region - Human Wes. .
Kappa chain V-I region - Human Au . .
13
Kappa chain V-I region - Human Ka . .
Kappa chain V-I region - Human Car. .
Kappa chain V-I region - Human Dee. .
Kappa chain V-I region - Human Lay. .
Kappa chain V-I region - Human Cu . .
23
Kappa chain V-I region - Human Su . .
Kappa chain V-I region - Human Heu. .
Kappa chain V region - Mouse MOPC 173
Kappa chain V-III region - Human Pom.
Kappa chain V-III region - Human Wol.
33
Kappa chain V-III region - Human Sie.
Kappa chain V-III region - Human Ti .
Kappa chain V regions - Mouse anti-ars
Kappa chain V-IV region - Human Len .
Kappa chain V region - Mouse CBPC 101
Kappa chain V regions - Mouse PC7043,
Kappa chain precursor V regions - Mous
Kappa chain V regions - Mouse PC6634,
Kappa chain V regions - Mouse PC3741 a
Kappa chain V regions - Mouse PC2485 a
43
Kappa chain V region - Mouse PC2154 .
Kappa chain precursor V regions - Mous
Kappa chain V regions - Mouse PC2880,
Kappa chain V region - Mouse S107A. .
Kappa chain V-II region - Human Tew .
Kappa chain V region - Mouse 26-10. .
Kappa chain V-II region - Human Mil .
Kappa chain V region - Mouse PC2413 .
Kappa chain V-II region - Human Cum .
Kappa chain V regions - Mouse X44, X24
53
Kappa chain precursor V region - Mouse
Kappa chain V region - Rabbit BS-5. .
Kappa chain V region - Rabbit 4135. .
Kappa chain V region - Rabbit 3374. .
Kappa chain V region - Rabbit 3368. .
Kappa chain V region - Rabbit K29-213
Kappa chain V region - Rabbit 3547. .
Kappa chain V region - Rabbit K-25. .
Kappa chain V region - Rabbit BS-1. .
Kappa chain V region precursor - Rabbi
63
Kappa chain V region - Mouse J539 . .
Kappa chain V region - Dog Gom. . . .
Kappa chain V region - Mouse ZS1.3. .
Kappa chain V region - Rabbit K16-167
Kappa chain V region - Rabbit 120 . .
Kappa chain V region - Rabbit 3315. .
Kappa chain V region - Rabbit 2717. .
Kappa chain V region - Mouse M511 . .
Kappa chain V region - Rabbit AH80-5.
Kappa chain V region - Mouse M167 . .
73
kappa chain V region - Rabbit XP-1. .

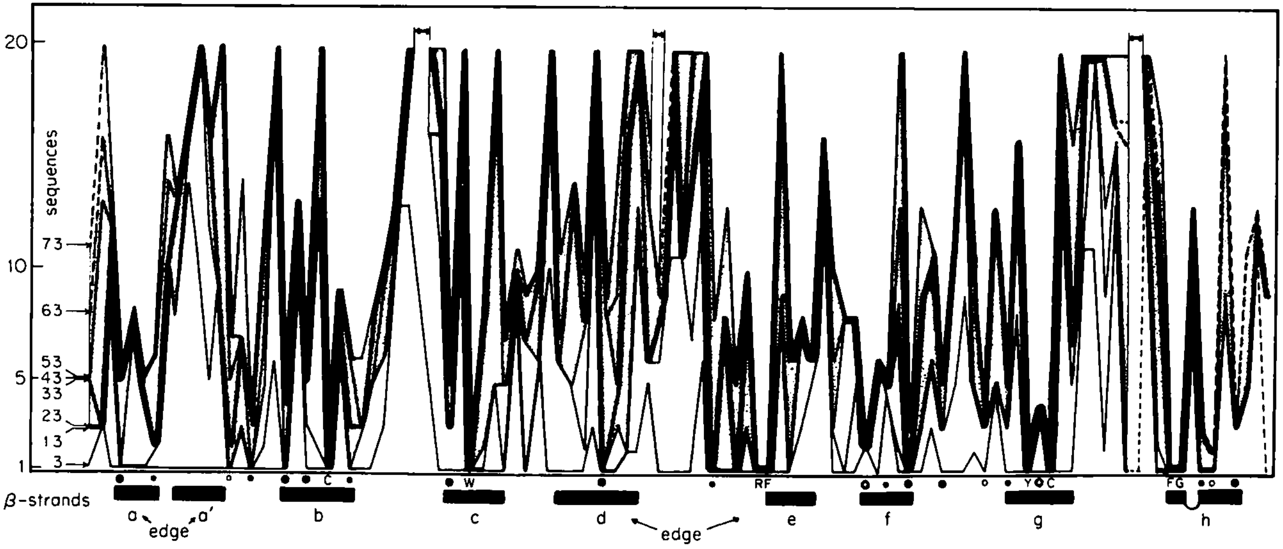


FIG. 7. Plot of degree of conservation along the sequence of aligned immunoglobulin sequences. Conservation is measured by the number of members in the assigned minimal set. This ranges from 1 (absolute conservation of type) to 20 (the set of all amino acids). Eight graphs are plotted, showing the effect of introducing more sequences into the alignment. These progress from three aligned sequences in the steps of ten to the final alignment of 73 sequences. The sequences are added in order of decreasing homology to the Bence Jones protein REI (as measured by residue identity). Insertions required to maintain the alignment occurred in the hyper-variable loops which are indicated by arrows at the top of the plot. The strands of β-structure observed in REI are indicated along the sequence by bars. (Those which lie on the edge of the two β-sheets are also indicated.) Above this, residue types that are conserved in all the sequences are indicated using the one letter code, highly conserved aromatic residues are indicated by a small hexagon, conserved hydrophobic residues by a large black dot, conserved small hydrophobic residues by a smaller dot and conserved small hydrophylic residues by a small open dot (see Fig. 5 for the detailed set assignment). To clarify the graph the line becomes thicker where more than one plot runs together and the region between 73 and 23 sequences is shaded. Considering both these features and bearing in mind that the plots must be monotonic decreasing with the number of sequences, it is possible to identify a single plot at most positions. The dotted lines indicate that a few sequences are not aligned in that region. The sequences are all light-variable domains of the immunoglobulin κ-chains found in the Protein Information Resource databank (Barker *et al.*, 1984). They come from a variety of species but mainly man, rabbit and mouse.

(super-secondary structures) commonly found in globular proteins (Sibanda & Thornton, 1985; Taylor *et al.*, in preparation). In these structures it is found that residue variation is restrained at particular locations in the motifs for general structural reasons. The observed patterns of conservation can then be used to predict the occurrence of the structural motif in a sequence of unknown structure using pattern recognition techniques such as the template matching method of Taylor & Thornton (1983, 1984) and Taylor (1986).

This work was begun while the author was a research fellow at IBM-UK Scientific Centre, Winchester, and the author thanks IBM for support and computing facilities. The author is currently supported by the SERC as an advanced fellow and thanks Prof. T. L. Blundell and Dr J. M. Thornton for valuable discussion.

REFERENCES

Baker, E. N. & Hubbard, R. E. (1984). *Prog. Biophys. mol. Biol.* **44,** 97.

Barker, W. C., Hunt, L. T., Orcutt, B. C. *et al.* (1984). *Protein Identification Resource.* Release 3.0.

Blundell, T. L., Sibanda, L. & Pearl, L. (1983). *Nature, Lond.* **304,** 273.

Chou, P. Y. & Fashman, G. D. (1974). *Biochemistry* **13,** 211.

Cohen, F. C., Sternberg, M. J. E. & Taylor, W. R. (1981). *J. mol. Biol.* **148,** 253.

Cohen, F. C., Sternberg, M. J. E. & Taylor, W. R. (1982). *J. mol. Biol.* **156,** 821.

Dayhoff, M. O. (1972). *Atlas of Protein Sequence and Structure.* Washington DC: National Biomedical Research Foundation.

Dayhoff, M. O. (1978). *Atlas of Protein Sequence and Structure.* Supplement 3. Washington, DC: National Biomedical Research Foundation.

Dickerson, R. E. & Geis, I. (1969). *The Structure and Action of Proteins.* Ch. 1. New York: Harper & Row.

Epp, O., Lattman, E. E., Schiffer, M., Huber, R. & Palm, W. (1975). *Biochemistry* **14,** 4943.

Fitch, W. M. (1966). *J. mol. Biol.* **16,** 9.

French, S. & Robson, B. (1983). *J. mol. Evol.* **19,** 171.

Garnier, J., Osguthorp, D. J. & Robson, B. (1978). *J. mol. Biol.* **120,** 97.

Grantham, R. (1974). *Science* **185,** 862.

Jiménez-Montaño, M. A. (1984). *Bull. math. Biol.* **46,** 641.

Klapper, M. H. (1971). *Biochim. biophys. Acta* **229,** 557.

McLachlan, A. D. (1971). *J. mol. Biol.* **61,** 409.

McLachlan, A. D. (1972). *J. mol. Biol.* **64,** 417.

Rees, A. R. C. & Sternberg, M. J. E. (1984). *From Cells to Atoms.* Ch. 1. Oxford: Blackwell Scientific.

Sander, C. & Schultz, G. E. (1979). *J. mol. Evol.* **13,** 245.

Schulz, G. E. & Schirmer, R. H. (1979). *Principles of Protein Structure.* Ch. 1. New York: Springer-Verlag.

Sibanda, B. L. & Thornton, J. M. (1985). *Nature, Lond.* **316,** 107.

Sneath, D. H. A. (1966). *J. theor. Biol.* **12,** 157.

Swanson, R. (1984). *Bull. Math. Biol.* **46,** 187.

Taylor, W. R. (1981). D. Phil. thesis, University of Oxford.

Taylor, W. R. (1986). *J. mol. Biol.* **188** (in press).

Taylor, W. R. & Thornton, J. M. (1983). *Nature, Lond.* **301,** 540.

Taylor, W. R. & Thornton, J. M. (1984). *J. mol. Biol.* **173,** 487.

Taylor, W. R., Thornton, J. M., Barlow, D. J., Sibanda, L. & Edwards, M. (in preparation).