J Mol Evol (1983) 19:171–175

Journal of Molecular Evolution

What is a Conservative Substitution?

Simon French[1] and Barry Robson[2]

[1]Department of Decision Theory, University of Manchester, Manchester, M13 9PL, Great Britain
[2]Department of Biochemistry, University of Manchester, Manchester, M13 9PL, Great Britain

Summary. It is commonly recognised that many evolutionary changes of amino acid sequence in proteins are conservative: a substitution of one amino acid residue for another has a far greater chance of being accepted if the two residues are similar in properties. Here we investigate what properties are most important in determining the similarity of two amino acids, from the evolutionary point of view. Our results confirm earlier observations that the hydrophobicity and the molecular bulk of the side chain tend to be conserved. More importantly they also show that evolutionary pressures favour the conservation of secondary structure, and that all these properties can be arranged in a two dimensional diagram in which distances well preserve the observed substitution frequencies between amino acids. These results were obtained by a multi-dimensional scaling technique; and are independent of any prior opinions about conserved properties. Thus, it is demonstrated that all relations of importance to single amino acid substitutions can be represented by a single figure, which is much more comprehensible and useful than the usual tabular representation of substitution frequencies. Such a figure conveniently portrays the "stereochemical code" for conservative substitution.

Key words: Amino acid substitution – Protein evolution – Conservation of secondary structure – Hydrophobicity – Bulk – Multidimensional scaling

Introduction

Dayhoff et al. (1972) have collated much data concerning the amino acid sequences of proteins. From their work and that of others it is apparent that natural selection has favoured changes in protein sequence in which certain physical and chemical properties of residues are conserved ('conservative substitution'). The question that concerned us was whether the relevant properties of have been properly and completely identified. We have been brought to the conclusion that interesting details have not been appreciated, except perhaps by inspection of similar sequences which does not allow all the significant properties to be considered together in quantitative, objective, and useful manner. We have sought a more objective, quantitative approach, finally using a data-analytic method not widely exploited by evolutionary molecular biologists; this study therefore serves also to bring this technique to their attention.

The topic of conservative substitutions is of interest for three reasons. First, there is the obvious need for such information in order to consider the similarity and relatedness of sequences. Second, it has often been hypothesised that natural selection pressures mainly favour the conservation of 3-dimensional structure while allowing for extensive substitution (cf. 'neutralist theories'). If this is so, the few chemical and physical properties conserved would presumably be those that most generally determine the 3-dimensional structure of a protein. Third, such knowledge is a pointer to which properties of residues must be modelled in computer simulation of protein folding. In speaking of 3-dimensional structure, we include the secondary structure on which the gross 3-dimensional structure depends.

Our analysis starts from that of Dayhoff et al. (1972) (their table 9.10) which constitutes a 'relatedness odds matrix'. The elements of this matrix give the ratio of two probabilities: the probability that two residues at the same locus in two proteins are the consequence of common ancestry, and the probability that the relation occurred only by chance. The data were derived from comparing sequences within the cytochrome c, haemo-

Offprint requests to: B. Robson

globin, myoglobin, virus coat, chymotrypsinogen, glyceraldehyde 3-phosphate dehydrogenase, clupeine insulin, and ferrodixin families of proteins. By combining several quite different families, they have obtained an account of the selective pressures on proteins in general rather than in specific instances. The measure of Dayhoff et al. thus provides a matrix (σ_{ij}), where the elements have the properly that $\sigma_{ij} > \sigma_{kl}$ if amino acids i and j appear more similar to each other from the viewpoint of evolutionary pressure than amino acids k and l.

Dayhoff et al. have noted that this similarity data points naturally to a classification of amino acids into 5 groups:

Hydrophilic:	Ala, Pro, Gly, Glu, Asp, Gln, Asn, Ser, Thr
Sulphydryl:	Cys
Aliphatic:	Val, Ile, Leu, Met
Basic:	Lys, Arg, His
Aromatic:	Phe, Tyr, Trp.

Here we extend Dayhoff et al.'s analysis through the statistical technique of multidimensional scaling. We refine their grouping and show that this new grouping corresponds to a very high degree with one deduced by Robson and Suzuki (1976). This latter classification grouped amino acid residues according to their tendencies to be involved in different forms of secondary structure. This correspondence between the two classifications is the first objective evidence *from substitution probabilities* for the reasonable conjecture that natural selection strongly favours the maintenance of the intrinsic stability of secondary structure features.

Method

Dayhoff et al. (1972) gave a similarity matrix, an ordering of elements reflecting the ordering of pairwise similarities between objects, here amino acid residues. The occurrence of such data is commonplace in psychology and sociology. Within those disciplines a family of statistical techniques, known collectively as multidimensional scaling, have been developed to explore and analyse similarity matrices. Surveys of these methods may be found in Shepard (1974), Shepard et al. (1972) and Sibson (1972). Briefly, the similarity matrix of Dayhoff et al. is analysed as follows. Using iterative optimisation techniques described in Kruskal (1964) and Guttman (1968) a set of 20 points (one for each amino acid residue is found in m dimensions such that nearer two points are, the more similar are the corresponding amino acids to evolutionary pressures. Essentially, this is comparable to finding the geographical distribution of towns from only an ordering of the (approximately determined) intertown distances (Kendall (1971)) and, furthermore, without knowing that the solution is two dimensional. More precisely, the optimisation is a best fit (in a particular least squares sense Kruskal (1964)) of the interpoint distances to the negatives of the measures of Dayhoff et al. asking only that

$$d_{ij} > d_{kl} \Leftrightarrow \sigma_{ij} < \sigma_{kl}$$

Where the d are the interpoint distances corresponding to the σ. Because this demands only that the σ are reasonably ordered and does *not* assume any functional relationship between the d_{ij} and σ_{ij}, this method is known to be very robust.

Results

A representation was readily obtained in two dimensions without any evidence that the use of a higher dimension would display any further information (Fig. 1) (the obtained stress (Kruksal 1964) was 9% and the Monte Carlo test procedure of Spence and Graef (1974) suggested clearly that a 2-dimensional representation was adequate).

Since the optimisation technique underlying multidimensional scaling is iterative, it requires an initial configuration. To avoid any possible bias we started with ten district random configurations. In each case the result converged to one with no significant difference from the one shown in Figure 1.

As expected (Dayhoff et al. (1972), Dickerson and Geiss (1969)) conservation of the hydrophobic nature of the residue is the most visually apparent feature. All points lie fairly close to a curve, the distance along which (from charged sidechain such as lysine, arginine, histidine to nonpolar aromatic residues, tyrosine and tryptophan) correlated well visually with increasing hydrophobicity. However the "horse shoe" shape of the curve also suggests a property of secondary importance, namely bulk which increases towards the right of the diagram. These two properties, hydrophobicity and bulk, are the only two amino acid properties that can be clearly seen to vary systematically along a trajectory, linear or otherwise, on the diagram. An automatic search for other properties of importance was also undertaken by analysis of the variation of many properties (Jungck (1978)) using a method developed by Carroll (1972) using a program in the MDS(X) suite for multidimensional scaling. However this failed to discover any further systematic variation. Thus it appears that the representation can answer the very general question: what amino acid properties tend to be conserved in evolution. Hydrophobicity and molecular bulk are the ones that we observed.

However the innovation in the diagram is that it can answer more *specific* questions than that general one. Namely, what amino acid properties are conserved in evolutionary change starting from a specific amino acid? Closer inspection of Figure 1 reveals features that at first glance seem curious. For example, the proximity of glycine and proline and of alanine and glutamic acid at the left side of the diagram is quite inconsistent with their bulk or degree of hydrophobicity. However, these apparent anomalies are in the nature of groupings which are strikingly similar to those obtained by Robson and Suzuki (1976) who undertook a clustering analysis of

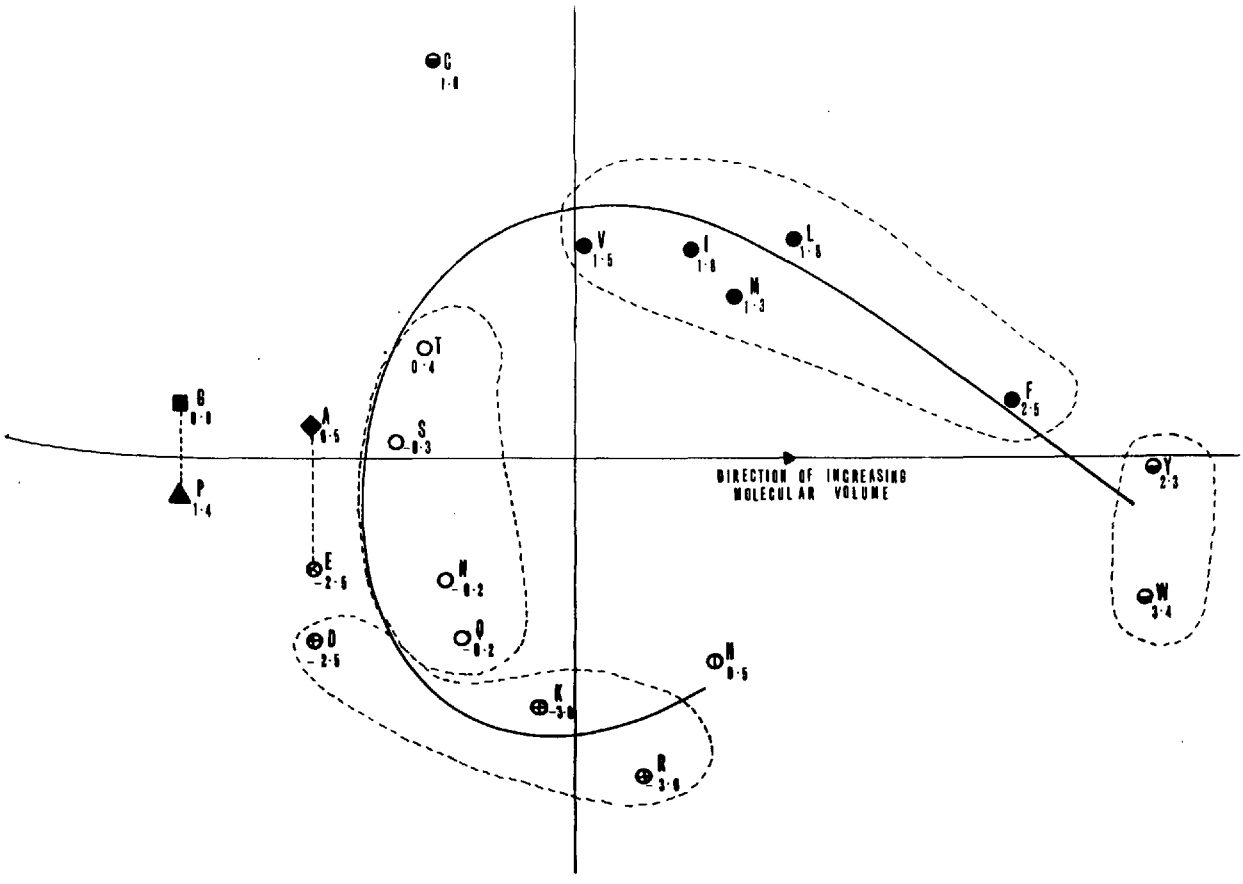

Fig. 1. Two-dimensional scaling plot (see text) of the odds-relatedness matrix of Dayhoff et al. (1972). The symbols correspond to those of Robson and Suzuki (1976). • Hydrophobic residues; ○ Hydrophobic residues which have ability to form hydrogen bonds; ○ Residues which may receive *or* donate hydrogen bonds; ○ Residues which may receive *and* donate hydrogen bonds; ■ Gly; ▲ Pro; ♦ Ala; ○ Glu ; ○ His (see text). The numbers associated with each amino acid are their hydrophobicities as given by Levitt (1976). The two indicated directions give general trends in the diagram of increasing molecular weight and volume. Note that the axes have no *a priori* significance in this technique, much as would also arise if a map of Britain were constructed from tables of distances between all towns and villages. That is to say, such a distance table contains no information about North-South, East-West axes (though the direction of South might be deduced *a posterior* in the grounds that a warmer climate encouraged habitation; in a similar way the properties of importance are deduced above leaving in mind that important trends need not be linear or even lie on a curve)

the 20 amino acid residues in a space whose dimensions corresponding to the helix, extended chain, and coil forming power of a residue. The work took no account whatsoever of evolutionary relationships between proteins or residues, but only between sequence and conformation. Their figures are reproduced for comparison in Fig. 2. The similar groupings obtained in Fig. 1 demonstrate for the first time that the preferences for different types of backbone conformational (secondary) structure are also a property of considerable importance for evolutionary pressures on amino acid mutations. As discussed by Robson and Suzuki, this effect arises as a result of sidechain-backbone interactions in a way largely determined by the nature or absence of hydrogen bonding groups in the sidechain. Glycine, proline, alanine and glutamic acid were treated as special cases on the grounds of special stereochemical effects and this is strongly supported by the present study.

There are, however, interesting differences. These authors classified residues according to whether sidechains were non-hydrogen bonding (filled circles in Fig. 1), could receive *and* donate a hydrogen bond (crossed circles), or could receive *or* donate a hydrogen bond (open circles). Histidine is 10% protonated at neutral pH and with reservations was assigned to the group of residues whose sidechains can both receive and donate hydrogen bond. From this point of view of evolutionary pressures, Fig. 1 places it firmly alongside lysine and arginine, which are close to fully charged at neutral pH. Indeed, it reveals that evolutionary pressure places much greater emphasis on whether a sidechain is negatively

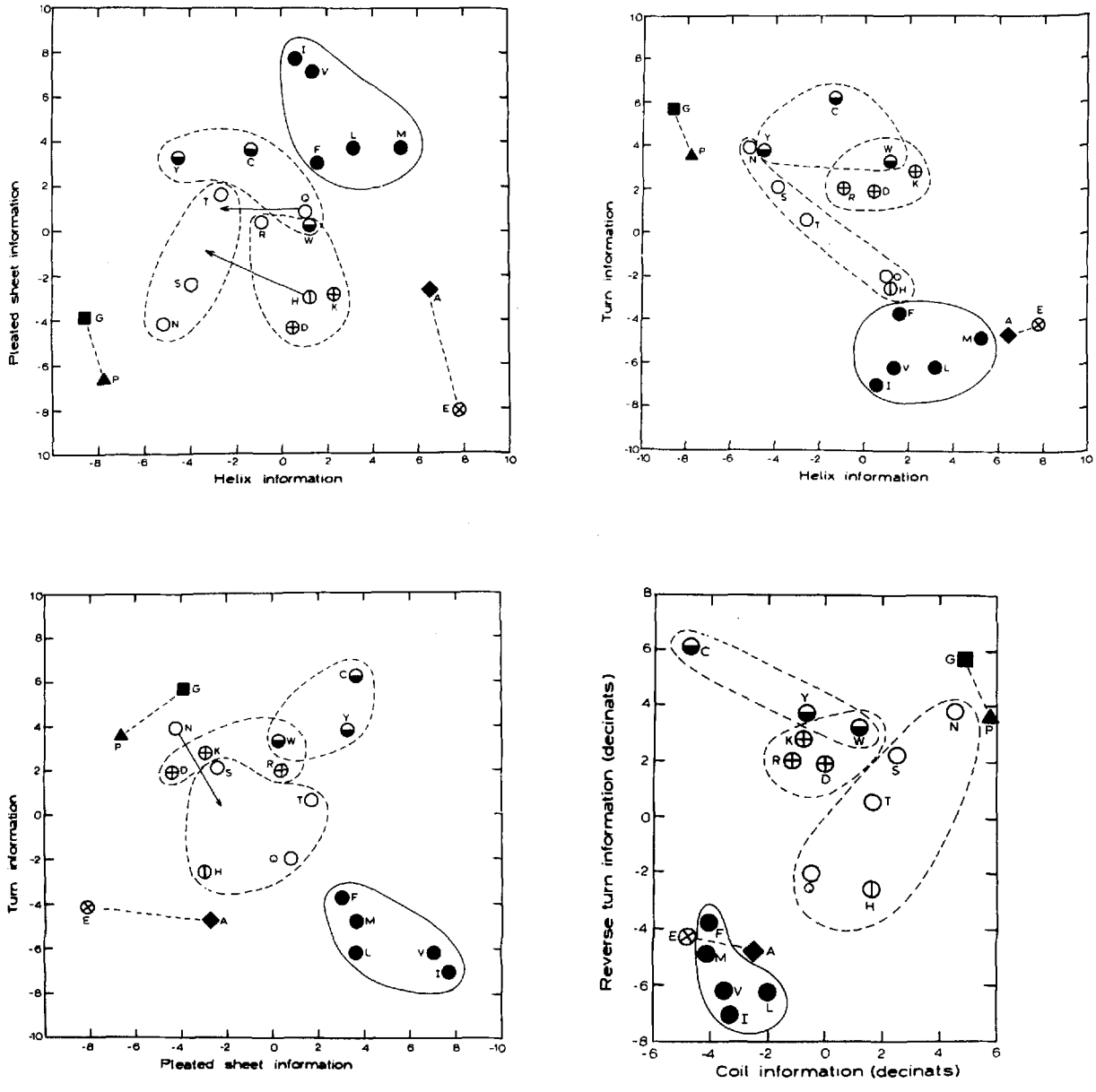

Fig. 2. The groupings of Robson and Suzuki based on conformational tendencies and physicochemical properties alone, i.e. without reference to comparison of homologous sequences. Symbols as Fig. 1

charged (glutamate and aspartate) or positively charged (lysine, arginine, histidine) than did the tentative assignments of Robson and Suzuki based on clustering analysis and the sidechain-backbone hydrogen bonding interactions. Cysteine also deviates from the largely nonpolar but weakly hydrogen bonding group to which it was assigned by the cluster analysis of Robson and Suzuki, but this may be expected from the point of view of evolutionary pressure because of its special role in forming covalent disulphide bridges in some cases. Use of tables of substitution distances obtained independently for intracellular and extracellular proteins might well clarify this point, though this would depart from the idea of seeking "gross" global determinants of substitution frequencies independent of any kind of family grouping, and independent of any specific interactions peculiar to a protein class. On the whole, however, the agreement is remarkable and this illustrates the value of multidimensional scaling in revealing patterns which may be meaningful to the observer.

The similarity between alanine and glutamate (AE) and proline and glycine (PG) in terms of substitution distances may seem surprising in view of the fact that the former are strong helix formers, the latter strong helix breakers. A preiliminary view might be that molecular bulk dominates here, perhaps along with a few physical properties of less general importance. However, other types of secondary structure tendency must be

considered, and it may be that the ability to disrupt extended (primarily plented sheet) structure, to introduce local bends *in it, and to demarcate its boundaries*, is of prime evolutionary importance. These aspects are now under investigation.

Conclusions

Evolution of proteins in general has tended to conserve (1) the degree of hydrophobicity of a residue, (2) the conformational preferences of its backbone and (3) its bulk. All these are continuous properties and the extent to which a substitution is conservative is correspondingly a matter of degree. Since the maximum distance in Figure 1 is between glycine and tryptophan, changes between residues at less than one third this distance might be conveniently classified as "good" conservative substitutions. Because most substitutions which would conserve bulk involve greater distances in Fig. 1 and indeed constitute "bad" conservative substitution, the dominance of importance seems to be in the order given above, with bulk playing a subserviant if significant role.

This work emphasizes the value of multidimensional scaling in reaching conclusions without any initial *a priori* assumptions. Jorre and Currow (1975) have applied the technique to a similar problem but their analysis had a strong theoretical input which modelled their prior beliefs about relationships. Moreover it was confined to a single well-defined protein family, the cytochrome c group, and therefore considered only evolutionary pressures relating to the structure, stability and function of cytochrome c. Hence they arrived at somewhat different conclusions and answered a different question. The advantage of the present work is, again, that it applies to the conversation of substitutions of proteins in general, using extensive data from which effects peculiar to conformations of specific families have presumably been almost entirely averaged out.

Acknowledgements. The programs in our analysis were from the MDS(X) package developed by A.P.M. Coxon and funded by the Social Science Research Council. Computing facilities were provided by the University of Manchester Regional Computer Centre.

One of us (SF) is most grateful to Dr. C.C.F. Blake for encouraging him to work in this area. The other (BR) is grateful for S.R.C. funding relevant to the discovery of properties relating to protein folding simulations.

After preparation of this manuscript Dr. W. Taylor has drawn to our attention to his very similar results and conclusions independently obtained (Taylor 1982). We are grateful to him for useful discussions.

References

Carroll JD (1972) Individual differences and multidimensional scaling. In: Shepard RN, Romney AK, Nerlove SB, Multidimensional scaling: Theory and Applications in the Behavioural sciences. Seminar Press, London, pp 105–155

Dayhoff MO, Eck RV, Park CM (1972) A model of evolutionary change in proteins. In: Dayhoff MO (ed) Atlas of protein sequence and structure. National Biomedical Research Foundation, Georgetown University, Washington DC, pp 89–99

Dickerson KE, Geis I (1969) The structure and action of proteins. *Harper and Row, New York*

Guttmann L (1968) A general non-metric technique for finding the smallest co-ordinate space for a configuration of points. Psychometrika 33:469–506

Jorre RP, Curnow RN (1975) A model for the evolution of proteins. Biochimie 57:1147–1156

Jungck JR (1978) The genetic code as a periodic table. J Mol Evol 11:211–224

Kendall DG (1971) Construction of maps from odd bits of information. Nature 231:158–159

Kruskal JB (1964) Non-metric multidimensional scaling. Psychometrika 29:1–27

Levitt M (1976) A simplified representation of protein conformations for rapid simulation of protein folding. J Mol Biol 104:59–107

Robson B, Suzuki E (1976) Conformational properties of amino acid residues in globular proteins. J Mol Biol 107:327–356

Shepard RN (1974) Representation of structure in similarity data: problems and prospects. Psychometrika 39:373–421

Shepard RN, Romney AK, Nerlove SB (1972) Multidimensional scaling: Theory and applications in the behavioural sciences. Vols I and II. Seminar Press, London

Sibson R (1972) Order in variant methods for data analysis (with discussion) J Roy Statist Soc B34:311–349

Spence I, Graef J (1974) The determination of the underlying dimensionality of an empirically obtained matrix of proximities. Multivariate Behavioural Research 9:331–342

Taylor, W (1982) Private Communication

Received July 20/Accepted November 1, 1982