Proc. Natl. Acad. Sci. USA
Vol. 74, No. 7, pp. 2820–2824, July 1977
Biochemistry

Structural convergence during protein evolution

(single-base substitutions/genetic code/amino acid substitutions/Darwinian evolution/protein secondary structure)

F. R. SALEMME, MICHAEL D. MILLER, AND STEVEN R. JORDAN

Department of Chemistry, University of Arizona, Tucson, Arizona 85721

Communicated by Martin D. Kamen, May 6, 1977

ABSTRACT **Several recent protein crystallographic structure determinations have demonstrated the existence of considerable tertiary structural similarity among proteins otherwise having little similarity in either amino acid sequence or biological function. In order to assess the possibility that such proteins may have arisen through processes of divergent evolution from a common ancestor, a graphical presentation is given which correlates the pattern of allowed single base substitutions defined by the genetic code with the associated changes in the structural properties of the encoded amino acids. The results show that while a large degree of structural conservation is evident due to codon synonomy, there is, in general, little tendency for the code to be structurally conservative in the majority of the cases where codon single-base changes result in amino acid substitutions. The possible consequences of this pattern of potential amino acid substitutions are discussed in relation to protein evolutionary processes.**

Some of the most convincing evidence for the common evolutionary origin of living organisms stems from observations of their fundamental biochemical and structural similarity. In recent years, extensive investigations have clearly shown that the presumed evolutionary relationship among members of protein families having functional and sequential similarities is also manifest by extensive similarities in the proteins' tertiary structures. Examples of such protein families include the cytochromes *c* (1, 2), the dehydrogenases (3–5), the oxygen-binding globins (6–9), and several of the serine proteases (10–12).

It is additionally evident, however, that in several instances considerable tertiary structural similarity exists among proteins having little or no similarity in either their amino acid sequences or their biological functions. Examples include similarities between the nucleotide-binding domain of the dehydrogenases with regions of subtilisin and flavodoxin (4), and overall structural similarity between immunoglobulin Fab fragments and superoxide dismutase (13). Although cases are known where structurally and sequentially related molecules have undergone considerable functional alteration during evolution (e.g., lysozyme and lactalbumin) (14), it is not clear whether those molecules showing similar tertiary structures in the absence of sequential or functional similarities result from processes of functionally divergent or structurally convergent evolution.

The work described here is an assessment of the extent to which single-base codon transitions defined by the genetic code result in the preservation of the structural properties of the

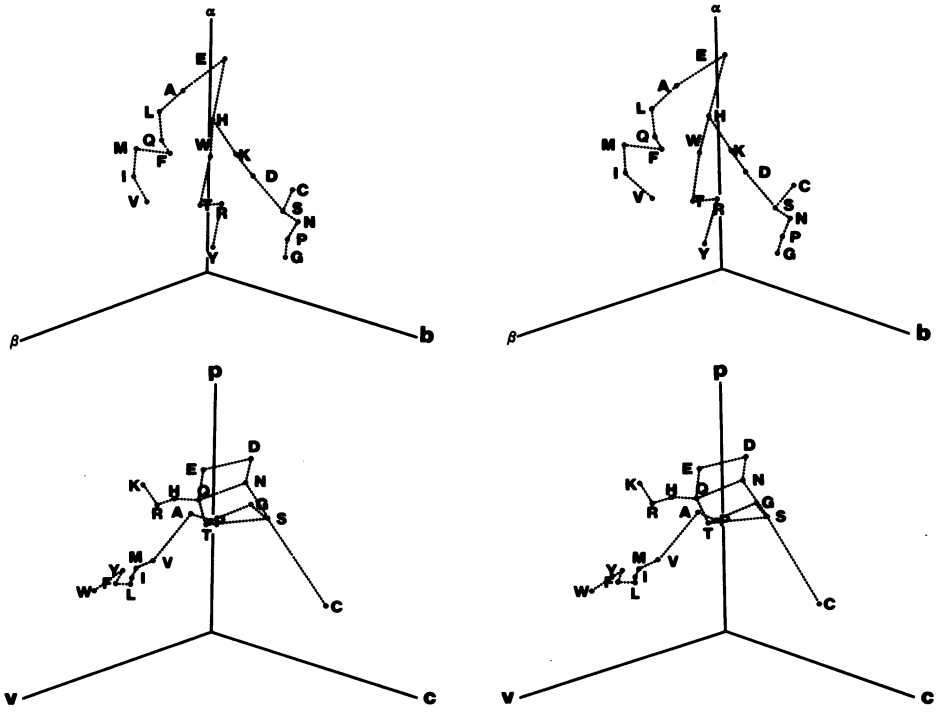

FIG. 1. Stereoscopic views of three-dimensional conformational spaces illustrating the structural properties of the 20 naturally occurring amino acids. *Upper* diagram shows distribution of amino acids according to their probabilities of forming α-helical (α), β-pleated sheet (β), or hairpin bend (*b*) secondary structures (20). *Lower* diagram plots side chain composition (*c*), polarity (*p*), and volume (*v*) (16). Dotted lines show representative nearest neighbor connectivity of the points in both spaces. Single-letter code (14): A, alanine; C, cysteine; D, aspartic acid; E, glutamic acid; F, phenylalanine; G, glycine; H, histidine; I, isoleucine; K, lysine; L, leucine; M, methionine; N, asparagine; P, proline; Q, glutamine; R, arginine; S, serine; T, threonine; V, valine; W, tryptophan; Y, tyrosine.

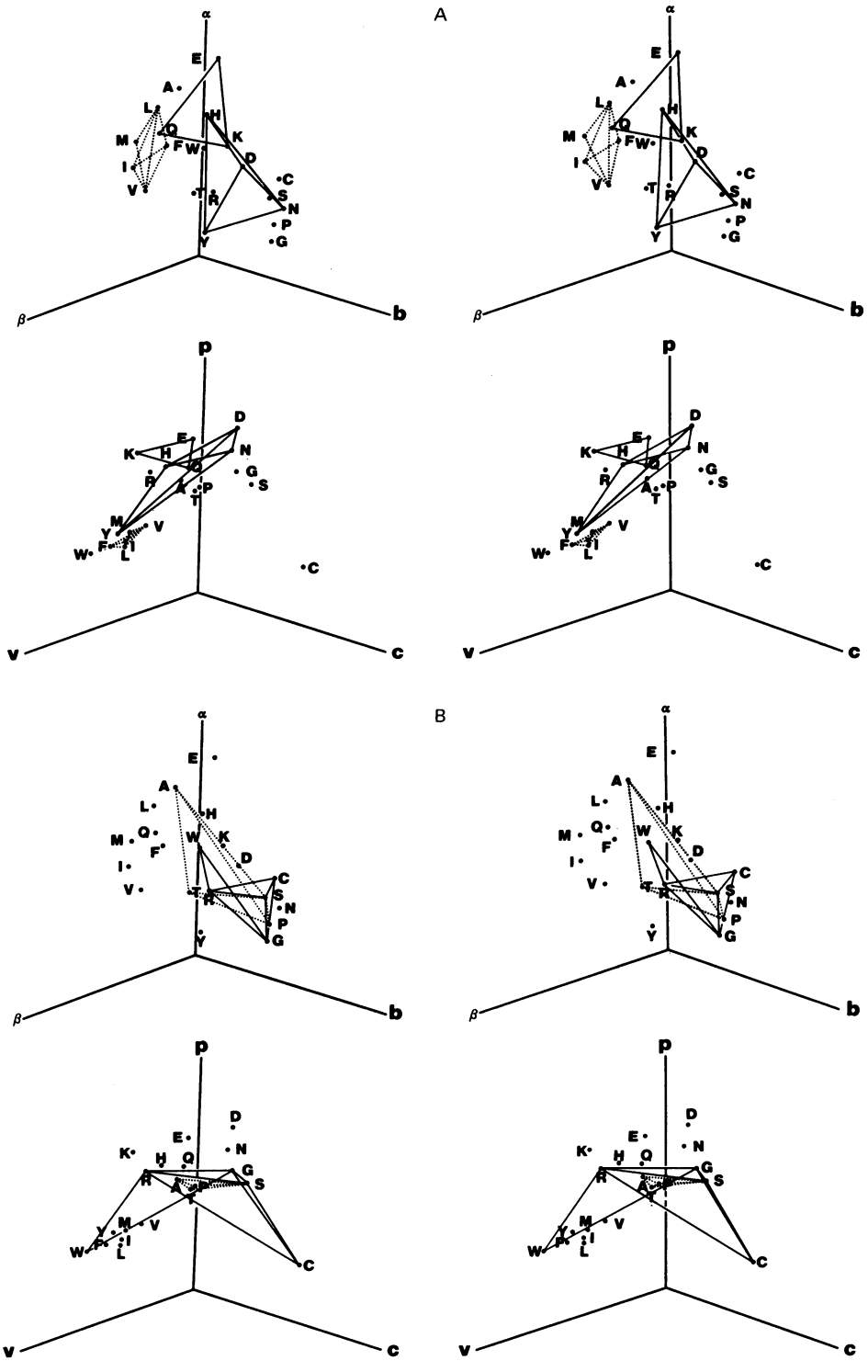

FIG. 2. First-base transitions and associated structural changes. (*A*) First-base changes holding U (dotted) or A (solid) constant as second-codon base. Amino acid structural properties associated with codon first-base changes holding U constant as second base are conservatively preserved in both spaces. (*B*) First-base changes holding C (dotted) or G (solid) constant as second base.

encoded amino acids. Although aspects of this subject have been considered previously (15–17), the present graphical treatment gives a more uniform depiction of the relative amino acid structural differences, and moreover provides additional insight concerning the overall pattern of amino acid structural changes accompanying single-base codon transitions. This information is useful in evaluating the extent of structural conservation inherent in the genetic code, and consequently, the relative likelihood that proteins showing extensive structural similarities in the absence of any sequence homology have resulted from processes of divergent or convergent evolution.

METHODS

The basic approach utilized in this study involves the characterization of each amino acid in terms of its known physical or structural properties, which are subsequently treated as *xyz* coordinates in a Cartesian coordinate system. This gives a set of 20 points (corresponding to the 20 amino acids) whose rela-

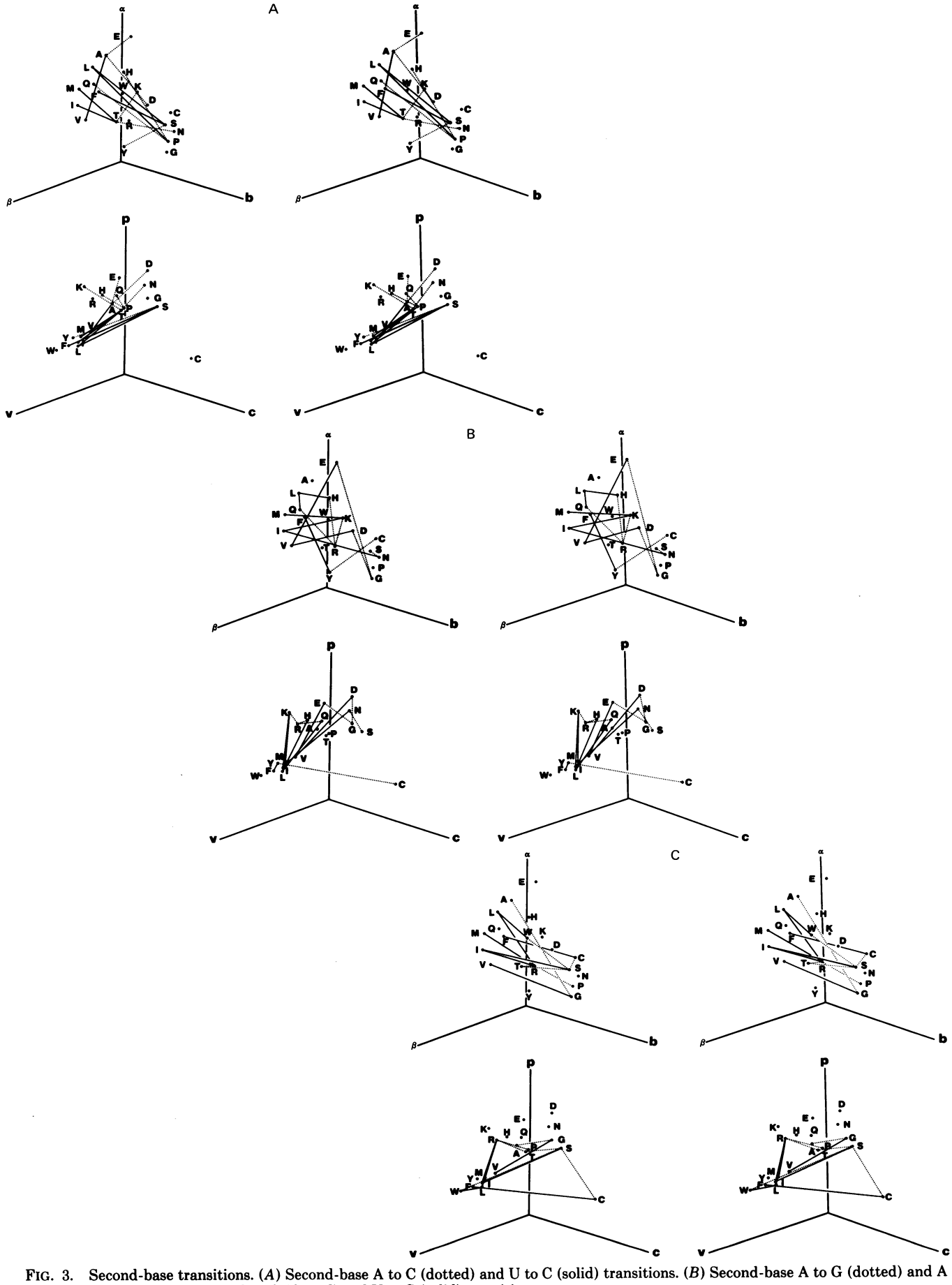

FIG. 3. Second-base transitions. (*A*) Second-base A to C (dotted) and U to C (solid) transitions. (*B*) Second-base A to G (dotted) and A to U transitions. (*C*) Second-base C to G (dotted) and U to G (solid) transitions.

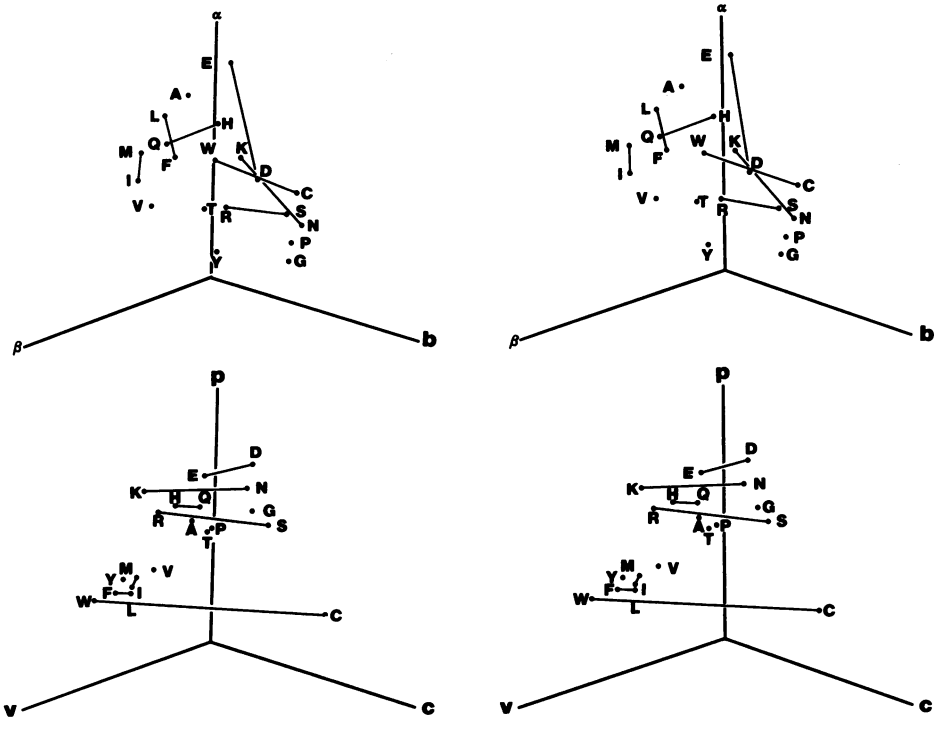

FIG. 4. Third-base transitions. These appear to strongly conserve side-chain polarity.

tive spatial relationships are an accurate representation of the structural differences between amino acids. It is then straightforward to connect those points corresponding to amino acid substitutions resulting from single-codon base substitutions to assess the degree of accompanying amino acid structural conservation.

Two sets of structural parameters are utilized to characterize individual amino acids.

The first set ("*cpv*") assigns values representative of composition (a function of the ratio of non-carbon to carbon side-chain atoms best related to the potential of an internal residue for hydrogen bond formation), side chain polarity, and volume to each amino acid (16). The physical rationale for the selection of these parameters stems from the observations that protein molecules are close-packed structures, having interiors composed of either apolar residues or hydrogen-bonded polar uncharged residues, with charged amino acid side chains located on the molecular exterior (18, 19).

The second parameter set ("$\alpha\beta b$") incorporates values representative of the probability that each amino acid may form α-helical, β-pleated sheet, or hairpin bend secondary structures (20, 21). These hydrogen-bonded secondary structures make up the majority of most known protein structures and, in general, constitute the features by which structural similarity between proteins is most readily recognized.

For graphical convenience, members of the respective parameter sets have been adjusted to the same relative scale according to the equation $p = P/D_p$, in which p is the scaled parameter value, p is the raw parameter value (16, 21), and D_p is the average parameter difference taken over all 190 pairs of amino acids. The scaled values for the two parameter sets are treated as independent variables (i.e., xyz) and plotted in three-dimensional Cartesian coordinate systems (Fig. 1). The values of the parameters associated with each amino acid define points in the respective *cpv* and $\alpha\beta b$ spaces representative of their structural properties. Distances between any two points associated with different amino acids are consequently a measure of the similarity or difference in their overall structural properties. By connecting points corresponding to those amino acid substitutions resulting from single base changes, the accompanying structural alterations that may result may be conveniently represented.

RESULTS AND DISCUSSION

Fig. 1 shows the distribution of the amino acids in the respective *cpv* and $\alpha\beta b$ structural parameter spaces. Although there are some similarities in the distribution of the points defining the amino acids in these spaces (e.g., phenylalanine, methionine, isoleucine, and leucine are close neighbors in both spaces), there does not appear to be any uniform transformation that will make all of the points defined in the two spaces coincident.

Figs. 2, 3, and 4 show the connectivity patterns arising from first-, second-, and third-base codon substitutions, respectively. From inspection of these drawings, it is apparent that with the exception of first-base codon transitions holding A or U constant as second base (Fig. 2A), the majority of the allowed transitions do not occur between nearest neighbors in these spaces.

Fig. 5 shows plots of the number of single-base changes defined by the genetic code versus corresponding overall differences (i.e., distances between plotted points) in amino acid structural properties in both the *cpv* and $\alpha\beta b$ spaces. Both plots show large peaks near their origins, predominantly resulting from transitions between synonomous codons whose associated structural difference is zero. However, the remainder of the transitions do not show marked tendencies towards the preservation of amino acid structural properties in either space. This can be seen by the close correspondence of the average distance between each amino acid and its eight nearest neighbors (d_8) and the mean structural difference (d_s) for all codon single-base changes, including those 26% of the total shown resulting from transitions between synonomous codons. Although it may be argued that the simultaneous preservation of all amino acid structural properties provides excessively rigid criteria by which to judge amino acid structural similarity, it is notable that with few exceptions (see Fig. 4), the structural differences observed

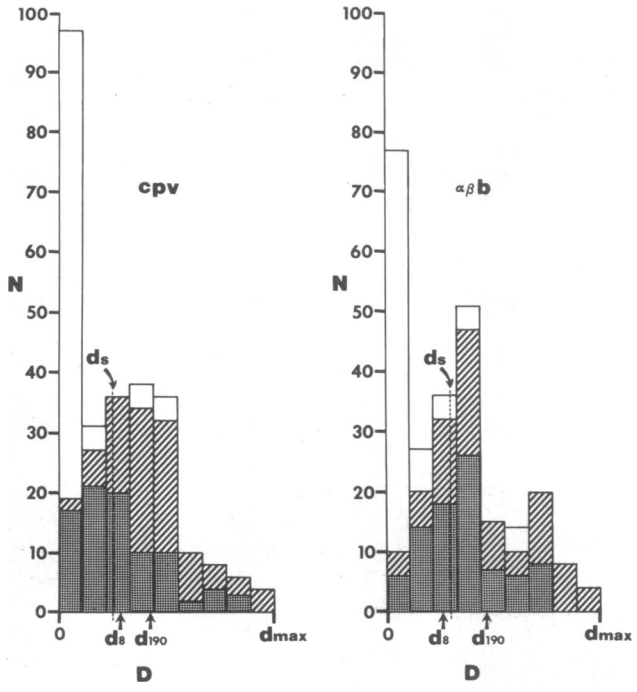

FIG. 5. Plots of number (N) of single-base codon substitutions defined by the genetic code versus associated amino acid structural differences (d) for *cpv* and *αβb* spaces. Each codon interconversion is counted once. Transitions to or between codons not coding for amino acids are omitted. Distance ranges for each space were divided into nine equal increments from $d = 0$ to $d = d_{max}$, the maximum distance between any two amino acids in the respective spaces. Indicated are d_s, the average value for all structural differences associated with single-base substitutions; d_8, the average structural difference between any amino acid and its eight nearest neighbors (on the average, any amino acid is related to 7.5 others by codon single-base substitutions), and d_{190}, the average structural difference between all pairs of amino acids in the respective spaces. The value of d_{190} is shown only to give an idea of the overall average range of values in the conformational spaces. It is not directly comparable to d_s or d_8 because its calculation assumes that all amino acid interconversions are both possible and equally probable, in contrast to what is actually allowed by the genetic code. First-base changes are represented by cross-hatched bars, second-base changes by obliquely striped bars, and third-base changes by open bars.

persist in the two-dimensional projections of these spaces along their axes.

Summarizing, it would appear that, with the exception of synonomous codon transitions, the overall consequences arising from minimal base substitutions favor protein structural diversification rather than structural preservation. Although it is possible that proteins having similar tertiary structures in the absence of any observable sequence homology or functional role may be related by processes of divergent evolution, the structural interconversions defined by the genetic code do not favor such processes. It is at least equally likely that there may, in fact, exist a relatively small number of preferred tertiary conformations that polypeptides may assume (22), subject either to requirements for facile kinetic pathways for folding, or the final attainment of a stable close-packed structure. Evidence for the latter possibility stems from observations of chirality invariance in super-secondary structures among virtually all known protein structures (23, 24), together with evidence that considerable tertiary structural similarity exists in proteins such as lactate dehydrogenase and carboxypeptidase despite connectivity differences in their polypeptide chains.

In conclusion, it is of interest to note that although there are many known examples of structural convergence in the evolution of whole organisms (e.g., the independent evolution of a streamlined body form by both porpoises and icthyosaurs), they are, in contrast to the situation for proteins, invariably associated with convergence upon similar function. The apparent disparity between the eventual outcome of whole organism versus protein convergent evolutionary processes results from the fact that protein functionality is usually defined in a chemical, rather than a structural context. However, the processes are similar in the sense they both result in the attainment of structures that are optimal adaptations to their physical environment.

This research has been supported by grants from the National Institutes of Health (GM 21534), the National Science Foundation (BMS 75-06558), and the Research Corporation.

The costs of publication of this article were defrayed in part by the payment of page charges from funds made available to support the research which is the subject of the article. This article must therefore be hereby marked "advertisement" in accordance with 18 U. S. C. §1734 solely to indicate this fact.

1. Salemme, F. R., Kraut, J. & Kamen, M. D. (1973) *J. Biol. Chem.* **248,** 7701–7716.
2. Dickerson, R. E., Timkovich, R. & Almassy, R. J. (1976) *J. Mol. Biol.* **100,** 473–491.
3. Rossmann, M. G. & Liljas, A. (1974) *J. Mol. Biol.* **85,** 177–181.
4. Rao, S. T. & Rossmann, M. G. (1973) *J. Mol. Biol.* **76,** 241–256.
5. Buehner, M., Ford, G. E., Moras, D., Olsen, K. W. & Rossmann, M. G. (1974) *J. Mol. Biol.* **90,** 25–49.
6. Hendrickson, W. A. & Love, W. E. (1971) *Nature New Biol.* **232,** 197–203.
7. Perutz, M. F., Muirhead, H., Cox, J. M. & Goaman, L. C. G. (1968) *Nature* **219,** 313–139.
8. Huber, R., Epp, O., Steigemann, W. & Formanek, H. (1971) *Eur. J. Biochem.* **19,** 42–50.
9. Kendrew, J. C. (1962) *Brookhaven Symp. Quant. Biol.* **15,** 216–225.
10. Shotton, D. M. & Watson, H. C. (1970) *Nature* **225,** 811–816.
11. Birktoft, J. J. & Blow, D. M. (1972) *J. Mol. Biol.* **68,** 187–240.
12. Stroud, R. M., Kay, L. M. & Dickerson, R. E. (1974) *J. Mol. Biol.* **83,** 185–208.
13. Richardson, J. S., Richardson, D. C., Thomas, K. A., Silverton, E. W. & Davies, D. R. (1976) *J. Mol. Biol.* **102,** 221–235.
14. Dayhoff, M. O., ed. (1972) *Atlas of Protein Sequence and Structure* (National Biomedical Research Foundation, Silver Spring, MD), Vol. 5.
15. Epstein, C. J. (1966) *Nature* **210,** 25–28.
16. Grantham, R. (1974) *Science* **185,** 862–864.
17. Sneath, P. H. A. (1966) *J. Theor. Biol.* **12,** 157–195.
18. Lee, B. K. & Richards, F. M. (1971) *J. Mol. Biol.* **59,** 379–400.
19. Shrake, A. & Rupley, J. A. (1973) *J. Mol. Biol.* **79,** 351–371.
20. Chou, P. Y. & Fasman, G. D. (1974) *Biochemistry* **13,** 211–222.
21. Chou, P. Y. & Fasman, G. D. (1974) *Biochemistry* **13,** 222–245.
22. Levitt, M. & Chothia, C. (1976) *Nature* **261,** 552–557.
23. Richardson, J. S. (1976) *Proc. Natl. Acad. Sci. USA* **73,** 2619–2623.
24. Sternberg, M. J. E. & Thornton, J. M. (1976) *J. Mol. Biol.* **105,** 367–382.