BUTTERWORTH HEINEMANN

Systematic representation of protein folding patterns

Arthur M. Lesk

Department of Haematology, University of Cambridge Clinical School, MRC Centre, Cambridge CB2 2QH, England

A tabular representation of protein folding patterns is described, which comprises information about the order along the chain of helices and strands of sheet, identifies the elements of secondary structure that interact, and indicates their relative orientation. These tableaux are intelligible to both people and computers, and support the application of algorithms for identification of proteins with similar folding patterns. Their inclusion in a database of protein structures would support investigations of structural relationships at the topological level.

Keywords: protein structure, topology, folding pattern, classification, tableau representation.

The great increase in known protein structures creates severe challenges in organizing and classifying their folding patterns. Here we present a concise general representation of any protein folding pattern, as a tableau showing the helices and sheets and their interactions. It satisfies the dual needs of intelligibility to people and computers, is suitable for application of pattern-recognition algorithms, and thereby provides a framework for information retrieval, database searching, and investigations of the range and relationships among protein topologies. Of particular significance is the possibility of generating systematically a complete catalogue of possible folding patterns and, by comparison with observed structures, deriving "selection rules" governing which of the a priori possible folds actually exist.

Protein structures show a wide variety of topologies.[1,2] Related proteins retain, but unrelated proteins may share, similar folds. Analysis of protein folding patterns has traditionally proceeded by letting the atomic coordinates represent the structure to computer programs, and a picture or pictures represent the structure to people. Visual examination of pictures has led to qualitative classifications of protein-folding patterns.[3,4] However, these have often been ad hoc or even anecdotal rather than systematic. There have been several attempts to classify certain sets of protein structures in a complete and systematic way. Presnell and Cohen examined four-helix bundles.[5] Murzin and Finkelstein developed a polyhedral model for the possible packings of α-helices around a hydrophobic core.[6] Ptitsyn et al. enumerated the possible Greek key folds formed by β sheets.[7] Finkelstein enumerated 60 different topologies of double β-sheet "sandwich" proteins with 4 strands in each sheet.[8] Murzin et al.,[9,10] using ideas of McLachlan,[11] investigated the possible topologies of β-barrels. More generally, numerous computational approaches based on numerical analysis of coordinate sets can identify structural similarities,[12–33] measure structural divergence, and thereby induce a classification of protein topologies.[24,33]

What is unsatisfactory about this situation is that neither the coordinate sets nor pictures embody explicitly the features that define a folding pattern: the assembly of the chain into helices and sheets and the geometry of their interactions. To abstract this essential information, Grindley and co-workers treated the interactions among secondary structure elements as a graph.[29] They have used algorithms for matching graphs and subgraphs to solve such problems as identifying common substructures. The representation introduced here is a development of that approach.

REPRESENTATION OF PROTEIN FOLDING PATTERNS AS TABLEAUX DISPLAYING INTERACTIONS OF HELICES AND SHEETS

A protein folding pattern is defined by the order along the chain of the helices and strands of sheet, and the geometry of interaction of the pairs of secondary structure elements that are in contact. This information can be encapsulated in a tableau. Figure 1A shows the structure of horse acylphosphatase.[33a,34] In Figure 1B the two helices and five strands of sheet are listed in order of appearance along the main diagonal of a matrix. (They are repeated, for clarity only, along the tops of the rows and sides of the columns). Each off-diagonal position in the matrix is either blank if the corresponding pair of secondary structure elements is not in contact, or contains a two-character symbol reflecting the geometric relationship between secondary structure elements that are in contact. The matrix is symmetric.

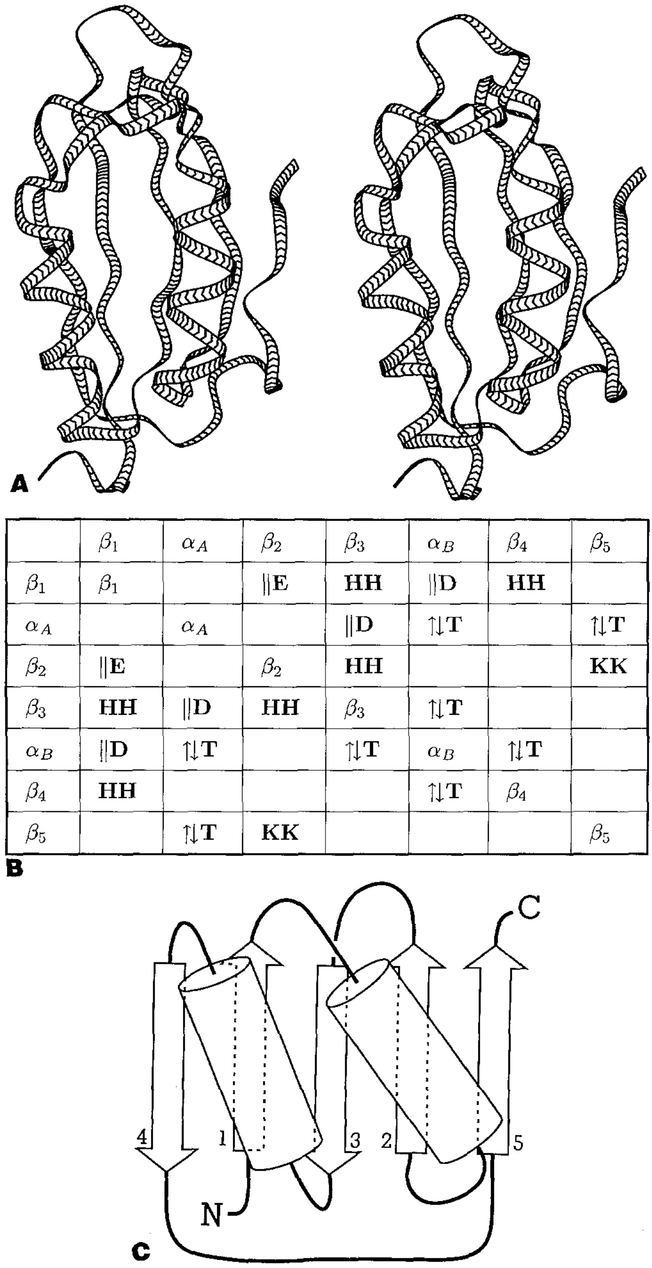
The encoding of the relative geometry is crucial.

Associate with each element of secondary structure a vec-

Address reprint requests to Dr. Lesk at the Department of Haematology, University of Cambridge Clinical School, MRC Centre, Hills Road, Cambridge CB2 2QH, England.
Received 8 February 1995; accepted 9 February 1995.

Journal of Molecular Graphics 13:159–164, 1995
© 1995 by Elsevier Science Inc.
655 Avenue of the Americas, New York, NY 10010

0263-7855/95/$10.00
SSDI 0263-7855(95)00037-7



	β_1	α_A	β_2	β_3	α_B	β_4	β_5
β_1	β_1		‖E	HH	‖D	HH	
α_A		α_A		‖D	↑↓T		↑↓T
β_2	‖E		β_2	HH			KK
β_3	HH	‖D	HH	β_3	↑↓T		
α_B	‖D	↑↓T		↑↓T	α_B	↑↓T	
β_4	HH				↑↓T	β_4	
β_5		↑↓T	KK				β_5

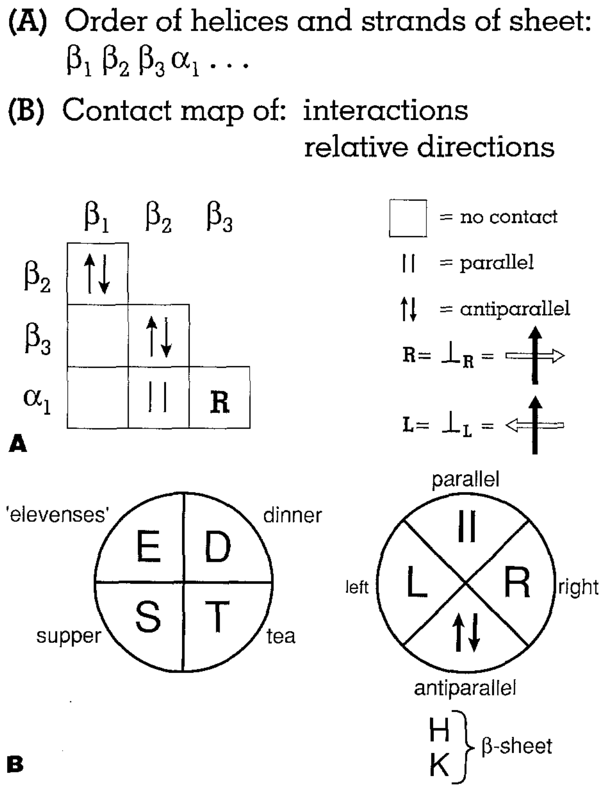
Figure 1. (A) The structure of horse acylphosphatase (Protein Data Bank[42] code 1APS). The herringbone pattern indicates the direction of the chain. (B) The abstract representation of its folding pattern as a tableau of interacting helices and sheets. The elements of the tableau indicate the relative geometry of the interacting units of secondary structure (see Figure 2). (C) A simplified representation of the structure, recoverable from the tableau in (B).

tor from its N to C terminus by fitting a least-squares line to the axis of each helix or to the C_α atom positions of each strand of sheet. For each pair of secondary structure elements in contact, Ω is the angle between the corresponding vectors after projection onto the plane perpendicular to the line between them; its sign is defined by the convention that a clockwise rotation of the nearer vector with respect to the farther is positive.[35] Because proteins with the same folding pattern show considerable variability in the angles between packed elements of secondary structure, it is necessary to characterize the geometries of interaction by broad rather than narrow categories. One such classification into four groups is as follows (see Figure 2A):

1. Nearly parallel ($\Omega = 0° \pm 45°$)
2. Perpendicular, "right-handed" (if the more distant vector points "up," the nearer vector points "right"; $\Omega = 90° \pm 45°$)
3. Perpendicular, "left-handed" (if the more distant vector points "up," the nearer vector points "left"; $\Omega = -90° \pm 45°$)
4. Antiparallel $\Omega = 180° \pm 45°$

Now, the problem with this or any other *single* classification into discrete ranges of angles is that variations in geometry among proteins with similar topologies may change the angles from one category to another, especially if the values are near a boundary. It would not be possible reliably to identify similar folding patterns by demanding that pairs of interacting helices and sheets retain the same class of geometry in different proteins. Note that a finer classification would aggravate rather than alleviate the problem.

A way out of this difficulty is to use a *double* classification of the relative geometry, using overlapping ranges of angles (see Figure 2B). The possible values of the interaxial



	β_1	β_2	β_3
β_2	↑↓		
β_3		↑↓	
α_1		||	R

Figure 2. (A) Definition of protein folding pattern and definition of classes of relative geometry of interacting helices and/or sheets. Because of shifts in relative geometry of helices and sheets during evolution of protein families, this set of classes is too simple. A satisfactory definition of classes of relative geometry of elements of secondary structure is shown in (B). (B) Double quadrant coding of overlapping classes of relative geometries.

angles are divided into quadrants in two ways, one corresponding to Figure 2A and the other with the boundaries rotated by 45°. Suitable mnemonic symbols distinct from those already used can be adduced by regarding the circle as a clock, and taking initial letters of the meal typically consumed (in Britain) during the time period in question (see Figure 2).

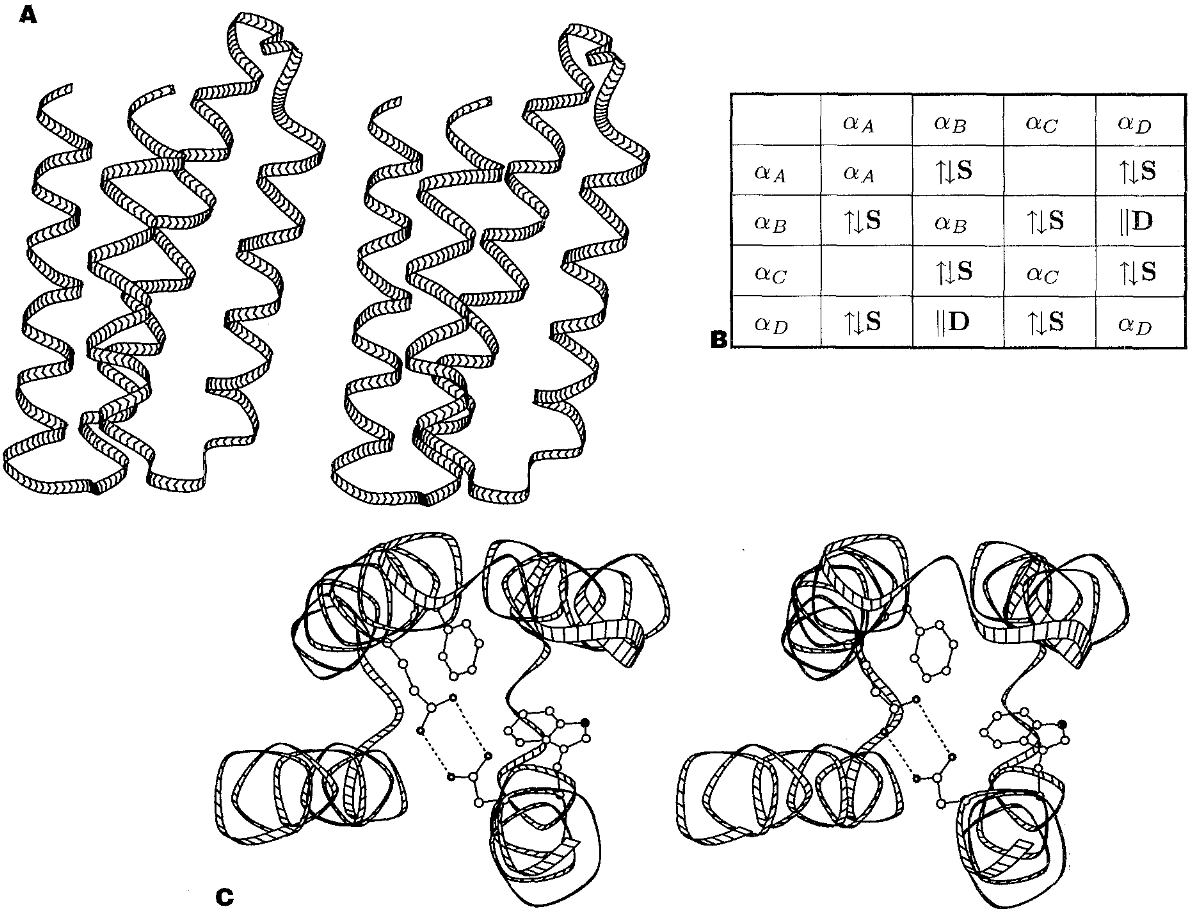
Ω (range, in degrees)		Symbol[a]	
0–90		D	
	45–135		R
90–180		T	
	135–225		O (for opposite) or ⇅
180–270		S	
	225–315		L
270–360		E	
	315–45		P or ‖

[a]We use O for opposite rather than A for antiparallel in order to reserve the letterA to stand for α helix when using restricted character sets.

Now encode each interaxial orientation by a two-letter code: an ordered pair of letters, one from either division of the circle into quadrants. For instance, $\Omega = 30^\circ$ corresponds to DR. If the two-letter codes for corresponding pairs of secondary structures in different proteins agree in either position, the difference in interaxial angles must be $\leq 90^\circ$. If they agree in both positions, the interaxial angles can differ by no more than 45°. Conversely, if they differ in both positions the difference in interaxial angles must be at least 45°. (Consider $\Omega = 44^\circ$, encoded D ‖, and $\Omega = 91^\circ$, encoded TR.) These facts will be useful in identifying similar substructures. Although the choice of quadrants gives reasonable ranges of angles based on our understanding of how protein structures evolve,[2] the double-encoding technique could be used to define classes with any required range, and indeed could be generalized to classes with unequal ranges of angles. The double-encoding technique is a versatile one and could be applied also, for example, to matching discretized distance matrices.

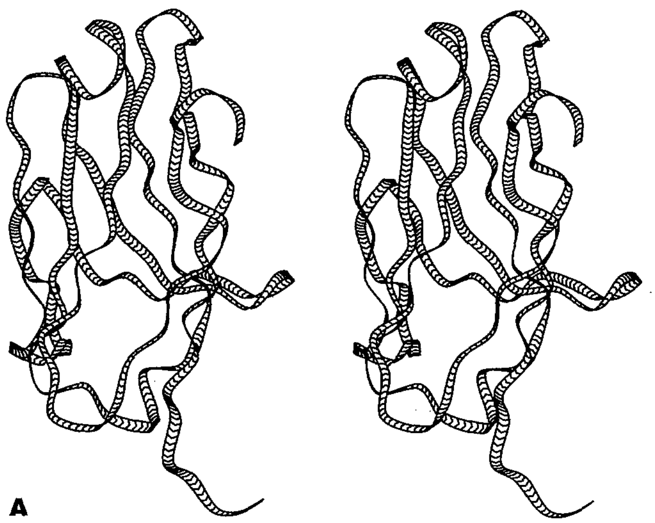
For strands of sheet in parallel and antiparallel orientations it is useful to distinguish those that are hydrogen bonded together and interact laterally to form the β sheet, from those that are in different β sheets packed face to face. In these cases, use the codes HH and KK for antiparallel and parallel β sheet interactions, respectively; the letters suggest the orientation of the hydrogen bonds.

Using this two-character code, any protein folding pattern can be represented as a symmetric tableau in which the rows and columns correspond to helices and strands of sheet, and the nonblank entries show geometries of interaction. A concise way of writing the tableau is then available using Forsyth notation (used in presenting positions of



	α_A	α_B	α_C	α_D
α_A	α_A	⇅S		⇅S
α_B	⇅S	α_B	⇅S	‖D
α_C		⇅S	α_C	⇅S
α_D	⇅S	‖D	⇅S	α_D

Figure 3. (A) Structure of hemerythrin, a four-helix bundle (Protein Data Bank code 1HMD). (B) Tableau representation of its folding pattern. (C) View down the long axis of the bundle, showing interactions between helices.



	β_1	β_2	β_3	β_4	β_5	β_6	β_7
β_1	β_1	KK				∥D	∥D
β_2	KK	β_2	∥E		KK	∥E	∥D
β_3		∥E	β_3		∥E	KK	∥E
β_4				β_4	KK		
β_5		KK	∥E	KK	β_5	∥E	∥E
β_6	∥D	∥E	KK		∥E	β_6	KK
β_7	∥D	∥D	∥E		∥E	KK	β_7

B

	β_1	β_2	β_3	β_4	β_5	β_6	β_7
β_1	β_1	KK					
β_2	KK	β_2			KK		
β_3			β_3			KK	
β_4				β_4	KK		
β_5		KK		KK	β_5		
β_6			KK			β_6	KK
β_7						KK	β_7

C

Figure 4. (A) Structure of immunoglobulin V_L domain 2RHE. (B) Tableau representation of the core of the domain: a four-stranded and a three-stranded β sheet. (C) A simplified tableau containing only the lateral interactions between strands in the same β sheets.

pieces on the chess board). Rows of the tableau are transcribed in order, with numerals signifying a number of blank spaces and a slash mark (/) separating lines. Reconstruction of the tableau from this abbreviation is straightforward. In this notation, and using only printable characters, the elements on and below the main diagonal in the tableau for acylphosphatase (Figure 1B) appear as follows:

B / 1 A / PE 1 B / HH PD HH B / PD OT 1 OT A / HH 3 OT B / 1 OT KK 3 B

This concise notation for the tableau could easily be included in any computer-readable database of protein structures, including but not limited to the Protein Data Bank entries themselves.

SOME REPRESENTATIVE EXAMPLES

Figures 3–5 illustrate a variety of protein structures and the corresponding tableaux. The simplest is the "four-helix bundle" of hemerythrin (Figure 3).[36] The elements of the tableau in positions adjacent to the main diagonal show the interactions between helices consecutive in the chain. These correspond to the interactions between helix A and helix B, helix B and helix C, and helix C and helix D. They are all ⇅ S, showing that each pair of consecutive helices forms an antiparallel hairpin. Helix A and helix D also form an antiparallel pair. Across the diagonals of the bundle, helices B and D interact, with their axes parallel, but helices A and C are not in contact. Figure 3C shows the view down the axis of the molecule, illustrating these contacts.

Figure 4 shows the core of the immunoglobulin domain from the typical β-sandwich protein V_L RHE.[37] The β sheet structure is shown by the cells containing KK: these indicate the assembly of the strands into fully antiparallel sheets. (Figure 4C contains a subtableau, limited to the cells containing KK.) Four of the strands—1, 3, 6, and 7—have only one partner; this "end-group analysis" shows that these must be edge strands, implying that there are two β sheets present. By aligning each strand with its partner(s), the strand order in the sheets can be inferred: 1, 2, 5, 4 and 3, 6, 7. Returning to the complete tableau (Figure 3B) we can conclude that the two sheets are packed face to face and infer their relative orientation.

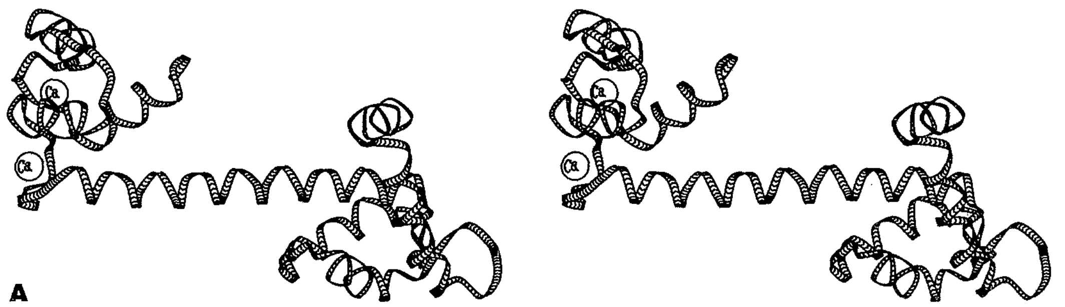
In the example of acylphosphatase (Figure 1B), reasoning analogous to that of the last paragraph shows that there is a single antiparallel sheet. The two helices are packed against the sheet, and, because they are packed against each other also, they must lie on the same side of the sheet with their axes parallel to the direction of the strands against which they are packed. It would be easy to produce the sketch of Figure 1C from the tableau of Figure 1. However, an enantiomorph with the helices on the other side of the sheet is also consistent with the tableau.

Figure 5 shows troponin c,[38] illustrating the point that a protein that consists of domains has a tableau that breaks into "blocks"; troponin c is admittedly an extreme example; other multidomain proteins show more interaction between secondary structures in different domains.

COMPARISON OF FOLDING PATTERNS

The measurement of similarity of folding patterns between two protein structures expressed as tableaux can be addressed by methods that detect the maximal common substructure.

If two proteins have the same set of secondary structures appearing in the same order along the chain, and equivalent elements of secondary structure interact with the same geometry—to within 90°—then they will have tableau of the same size, and corresponding entries will either both be blank or will agree in at least one position of the two-letter code. If corresponding nonblank entries all agree in both



	α_A	α_B	α_C	α_D	α_E	α_F	α_G	α_H
α_A	α_A	⇅T			LD			
α_B	⇅T	α_B	⇅T		⇅T			
α_C		⇅T	α_C	⇅T	‖E			
α_D			⇅T	α_D	⇅T			
α_E	LD	⇅T	‖E	⇅T	α_E			⇅T
α_F						α_F	⇅T	
α_G						⇅T	α_G	
α_H					⇅T			α_H

B

Figure 5. (A) Structure of troponin C (Protein Data Bank code 4TNC). (B) Tableau representation of the folding pattern of troponin C, showing that in the case of two noninteracting domains the tableau breaks into blocks.

positions, then the geometry will be preserved to within 45°.

If two proteins share a common core,[39] but each is decorated by different peripheral secondary structures, then the maximal common substructures can be found. From each molecule, consider the possible subsets of helices or strands of sheet, and the subtableaux that represent them and their interactions. These are obtained from the original tableau by crossing out certain rows and the corresponding columns. If two subtableaux have the same size and entries equal in one or both positions, the corresponding substructures have the same folding pattern. The mathematical problem is to find the largest equal subtableaux, which correspond to the maximal common substructure (there may, in principle, be more than one). A solution to this problem has been described.[40] This approach would also permit probing a database of folding patterns for proteins containing a query structure.

Some general rules about the assembly of secondary structural elements in proteins have simple expressions in terms of tableaux. For instance, elements of secondary structure formed from regions adjacent in the sequence tend to be in contact in three dimensions. This is equivalent to saying that positions of the tableaux adjacent to the diagonal are likely to contain non-null entries. The observation that the packing of ridges into grooves at the interfaces between interacting helices produces preferred classes of interaxial angles[35]—centered at −52°, +23°, and −105°—can be expressed by the "selection rule" that angles encoded by D and R will be rare in helix–helix packings.

CONCLUSION

This article has described and illustrated the properties of a tabular representation of protein-folding patterns. Any protein fold defined by packings of helices and sheets, which includes all but a few of the known structures, can be represented as such a tableau. Software has been constructed for generation and analysis of tableaux, including the determination of the maximal common subtableau[x] of two protein structures.

We suggest that these tableaux provide a useful framework for investigations aimed at analyzing the global properties of the set of protein-folding patterns.

ACKNOWLEDGMENTS

I thank Dr. A. Pastore for helpful discussion, Prof. A.M. Geoffrion for the use of the program RIP30C, and the Kay Kendall Foundation for generous support.

REFERENCES

1 Branden, C. and Tooze, J. *Introduction to Protein Structure*. Garland Publishing, Inc., New York, 1991

2 Lesk A.M. *Protein Architecture; A Practical Approach*. IRL Press, Oxford, 1991

3 Levitt, M. and Chothia, C. Structural patterns in globular proteins. *Nature* (*London*) 1976, **261,** 552–558

4 Richardson, J. The anatomy and taxonomy of protein structure. *Adv. Protein Chem.* 1981, **34,** 167–339

5 Presnell, S.R. and Cohen, F.E. Topological distribution of 4-α-helix bundles. *Proc. Natl. Acad. Sci. U.S.A.* 1989, **86,** 6592–6596

6 Murzin, A.G. and Finkelstein, A.V. The general architecture of α-helical globules. *J. Mol. Biol.* 1988, **204,** 749–770

7 Ptitsyn, O.B., Finkelstein, A.V., and Falk (Bendzko), P. Principal folding pathway and topology of all-β proteins. *FEBS Lett.* 1979, **101,** 1–5

8 Finkelstein, A.V. In: *PRODES90: Protein Design on Computers*. EMBL, BIOcomputing Technical Document 6, p. 139, 1991

9 Murzin, A.G., Lesk, A.M., and Chothia, C. Principles determining the structure of β sheet barrels in proteins. I. A theoretical analysis. *J. Mol. Biol.* 1994, **236,** 1369–1381

10 Murzin, A.G., Lesk, A.M., and Chothia, C. Principles determining the structure of β sheet barrels in proteins. II. The observed structures. *J. Mol. Biol.* 1994, **236,** 1382–1400

11 McLachlan, A.D. Gene duplications in the structural evolution of chymotrypsin. *J. Mol. Biol.* 1979, **128,** 49–79

12 Matthews, B.W. and Rossman, M.G. Comparison of protein structures. *Methods Enzymol.* 1985, **115,** 397–420

13 Levine, M., Stuart, D., and Williams, J. A method for the systematic comparison of the three-dimensional structures of proteins and some results. *Acta Crystallogr.* 1984, **A40,** 600–610

14 Liebman, M.N., Venanzi, C.A., and Weinstein, H. Structural analysis of carboxypeptidase A and its complexes with inhibitors as a basis for modelling enzyme recognition and specificity. *Biopolymers* 1985, **24,** 1721–1758

15 Abagyan, R.A. and Maiorov, V.N. A simple qualitative representation of polypeptide chain folds: Comparison of protein tertiary structures. *J. Biomol. Struct. Dynam.* 1988, **5,** 1267–1279

16 Zuker, M. and Somorjai, R.L. The alignment of protein structures in three dimensions. *Bull. Math. Biol.* 1989, **51,** 55–78

17 Willet, P. *Three-Dimensional Chemical Structure Handling*. Research Studies Press, Taunton, 1991, Chap. 3

18 Mitchell, E.M., Artymiuk, P.J., Rice, D.W., and Willett, P. Use of techniques from graph theory to compare secondary structure motifs in proteins. *J. Mol. Biol.* 1989, **212,** 151–166

19 Karpen, M.E., de Haseth, P.L., and Neet, K.E. Comparing short protein substructures by a method based on backbone torsion angles. *Proteins Struct. Funct. Genet.* 1989, **6,** 155–167

20 Taylor, W.R. and Orengo, C.A. Protein structure alignment. *J. Mol. Biol.* 1989, **208,** 1–22

21 Orengo, C. and Taylor, W.R. A rapid method for protein structure alignment. *J. Theor. Biol.* 1990, **147,** 517–551

22 Šali, A. and Blundell, T. Definition of general topological equivalence in protein structures. A procedure involving comparison of properties and relationships through simulated annealing and dynamic programming. *J. Mol. Biol.* 1990, **212,** 403–420

23 Vriend, G. and Sander, C. Detection of common three-dimensional substructures in proteins. *Proteins Struct. Funct. Genet.* 1991, **11,** 52–58

24 Orengo, C., Brown, N.P., and Taylor, W.R. Fast structure alignment for protein databank searching. *Proteins Struct. Funct. Genet.* 1992, **14,** 139–167

25 Johnson, M.S. Comparisons of protein structures. *Curr. Opin. Struct. Biol.* 1991, **1,** 334–344

26 Holm, L., Ouzonis, C., Sander, C., Tuparev, G., and Vriend, G. A database of protein structure families with common folding motifs. *Protein Sci.* 1992, **1,** 1691–1698

27 Alexandrov, N.N., Takahashi, K., and Gō, N. Common spatial arrangement of backbone fragments in homologous and non-homologous proteins. *J. Mol. Biol.* 1992, **225,** 5–9

28 Kolaskar, A.S. and Kulkarni-Kale, V. Sequence alignment approach to pick up conformationally similar protein fragments. *J. Mol. Biol.* 1992, **223,** 1053–1061

29 Grindley, H., Artymiuk, P.J., Rice, D., and Willett, P. Identification of tertiary structure resemblance in proteins using a maximal common subgraph isomorphism algorithm. *J. Mol. Biol.* 1993, **229,** 707–721

30 Orengo, C.A. and Taylor, W.R. A local alignment method for protein structure motifs. *J. Mol. Biol.* 1993, **233,** 488–497

31 Orengo, C.A., Flores, T.P., Taylor, W.R., and Thornton, J.M. Identification and classification of protein fold families. *Protein Eng.* 1993, **6,** 485–500

32 Subbiah, S., Laurents, D.W., and Levitt, M. Structural similarity of DNA-binding domains of bacteriophage repressors and the globin core. *Curr. Biol.* 1993, **3,** 141–148

33 Holm, L. and Sander, C. Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* 1993, **233,** 123–138

33a Bernstein, F.C., Koetzle, T.F., Williams, G.J.B. Meyer, E.F. Jr., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., and Tasumi, M. The protein databank: A computer-based archival file for macromolecular structure. *J. Mol. Biol.* 1977, **112,** 535–542

34 Pastore, A., Saudek, V., Ramponi, G., and Williams, R.J.P. Three-dimensional structure of acylphosphatase. Refinement and structure analysis. *J. Mol. Biol.* 1992, **224,** 427–440

35 Chothia, C.H., Levitt, M., and Richardson, D. Helix to helix packing in proteins. *Proc. Natl. Acad. Sci. U.S.A.* 1977, **74,** 4130–4134

36 Stenkamp, R.E., Sieker, L.C., Jensen, L.H. McCallum, J.D., and Sanders-Loehr, J. Active site structures of deoxyhemerythrin and oxyhemerythrin. *Proc. Natl. Acad. Sci. U.S.A.* 1985, **82,** 713–716

37 Furey, W. Jr., Wang, B.C., Yoo, C.S., and Sax, M. Structure of a novel Bence–Jones protein (Rhe) fragment at 1.6 Å resolution. *J. Mol. Biol.* 1983, **167,** 661–692

38 Satyshur, K.A., Rao, S.T., Pyzalska, D., Drendel, W., Greaser, M., and Sundaralingam, M. Refined structure of chicken skeletal muscle troponin c in the two-calcium state at 2-Å resolution. *J. Biol. Chem.* 1988, **263,** 1628–1647

39 Chothia, C., and Lesk, A.M. Relationship between the divergence of sequence and structure in proteins. *EMBO J.* 1986, **5,** 823–826

40 Lesk, A.M. Boolean programming formulation of some pattern-matching problems in molecular biology. *J. Chem. Soc. Faraday Trans.* 1993, **89,** 2603–2607

41 Chothia, C. and Finkelstein, A.V. The classification and origins of protein folding patterns. *Annu. Rev. Biochem.* 1990, **59,** 1007–1039