Received for publication 17 December 1979.

REFERENCES

1. Richardson, J. S. 1979. The anatomy and taxonomy of protein structures. *Adv. Prot. Chem.* In press.

QUANTITATIVE ANALYSIS OF STRUCTURAL DOMAINS IN PROTEIN

M. N. Liebman, *The Institute for Cancer Research, Philadelphia, Pennsylvania 19111 U.S.A.*

X-ray crystallographic studies have provided the three-dimensional structures of more than 100 proteins, including both structurally and functionally related families of macromolecules. To attempt to understand the relationship between structure, function, evolution, and macromolecular recognition and specificity, it has been of interest to compare the structures of related proteins or polypeptide folding domains. The statistical, root-mean-square deviation has provided a semiquantitative measure of structural agreement after the superposition of the segments to be compared. A new method is reported which permits the quantitative separation and comparison of the contributions of secondary, tertiary, and quaternary structure without the requirement of direct superposition technique.

The observation of several polypeptide folding patterns (~40–100 residues in length) reveals both intrinsic functional properties (e.g., nucleotide binding domain), and apparent structural stability (e.g., β-barrels, immunoglobulin fold, hemerythrin fold). It is of interest to be able to compare these analogous features at all structural levels: primary, secondary (structural elements of the domain), tertiary (intra-domain organization), and quaternary (inter-domain and intermolecular packing). The representation of the protein structures by distance matrix methods (1, 2, 3) has already provided qualitative methods for indicating structural domains (4), intra- and inter-molecular symmetry (5), quantitative assignment of structural insertions and deletions in the polypeptide (5), and protein:protein interactions(5). Quantitative examination of idealized secondary and tertiary structural interactions have also used this method (5).

Distance representation involves the construction of a square matrix of n cells, where n is the number of amino acids in the protein. The elements of this matrix, (i–j), contain the distance between the i-th and j-th alpha carbon along the polypeptide. Selective contouring of this matrix reveals levels of structural organization by pattern recognition (1, 2, 3, 4, 5). Comparison between structures using this representation can be achieved without superpositions of the three-dimensional coordinates because the distance matrices are internally referenced and thus independent of molecular rotation or translation (5).

It has been recently shown (6) that the comparison of two protein structures by use of a root-mean-square statistic is highly dependent on the nature of the secondary structures within the proteins. This reflects the correlated nature of the polypeptide chain caused by the chemical linkage, and also is indicative of the difference between topological and topographical identities. Thus it is inadequate to describe the difference between sperm whale

myoglobin, met vs. deoxy, in agreement with an rms value of 0.15 Å. While this indicates the high degree of similarity between the two forms, no information is available as to whether the differences are uniformly distributed throughout the structure, all contained within one helix undergoing distortion, two helices moving closer together, etc.

We note, in approaching this problem, that the internal organization of the distance matrix, with secondary structure occurring along the diagonal, and tertiary and quaternary structure more distant, suggests structural comparison by partitioning (Fig. 1). A simple algorithm has been developed to separate analytically the virtually-bonded alpha carbon chain into one of several broad secondary structural groupings. In addition, a domain-identifying algorithm has been developed which operates in a similar manner to that recently reported (7), but one which functions in distance space by simple comparison of matrix rows and columns. The application of these procedures thus subtends the linear sequence into secondary structures and domains. This subtended distance matrix then reveals tertiary structure partitions which contain helix-helix, helix-sheet, sheet-sheet, turn-turn, domain-domain, etc., interactions, as well as partitions of the secondary structures themselves. The comparison between two macromolecules can then be achieved by taking the difference of their respective distance matrices, and determination of those partitions which show significant differences. In addition, by use of vector differences, it is possible to differentiate between structures moving apart or together. This technique has been further utilized in comparing evolutionarily related proteins where insertions and deletions have occurred in the amino acid sequence. The topographical mapping algorithm (5) enables the comparison to reveal to what extent these modifications in the primary structure have affected higher levels of organization in either a direct or indirect manner. Results of some of the comparisons are given in Table I. This analysis has been extended to compare immunoglobulins, proteases, dehydrogenases, globins, lysozymes, cytochromes, flavodoxins, and hemerythrins.

This method is an attempt to establish guidelines to understand better the levels of both intra- and intermolecular organization. The analysis is directed towards understanding the relationship between structure and function, macromolecular recognition and specificity, and evolutionary controls and constraints. It is hoped that by developing the capability to observe

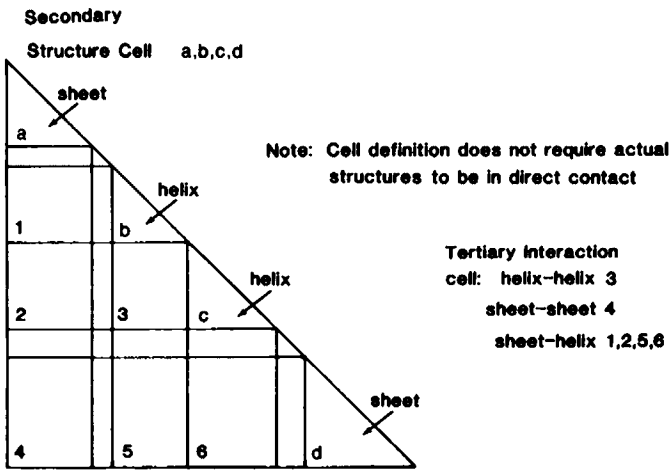

Figure 1 Partitioning of distance matrix.

TABLE I
COMPARISON OF DISTANCE COMPARISON STATISTICS WITH RMS ANALYSIS

Proteins Compared (8)	No. Initial		No. Final		Difference distance cell averages				
	Res.	Rms	Res.	Rms	All	2 only	H/S/M	3 only	Rest
Myoglobin, Sperm Whale									
Deoxy vs. Met	153	0.15A	—	—	0.08A	0.03A	0.05(H)	0.09A	0.11A
Myoglobin, Met									
Seal vs. Sperm Whale	153	1.94A	142	1.38A	0.91A	0.18A	0.30(H)	1.01A	1.39A
Cytochrome C, Albacore									
Red vs. Oxid	103	1.11A	99	0.93A	0.38A	0.16A	0.31(H)	0.41A	0.42A
Lysozyme, Hen Egg White							0.11(H)		
Tricl vs. Mono	129	0.51A	128	0.49A	0.20A	0.11A	0.16(S) 0.17(M)	0.21A	0.24A
Lysozyme, Hen Egg White							0.09(H)		
Inact vs. Act	129	0.38A	127	0.36A	0.16A	0.10A	0.15(S) 0.14(M)	0.17A	0.18A
Hemerythrin									
B vs. Aquo Met	113	2.95A	104	2.04A	1.11A	0.36A	0.38(H)	1.33A	1.86A
Flavodoxin									
Oxid vs. Semiqui	138	0.29A	133	0.25A	0.10A	0.06A	—	0.10A	—
Calcium Binding Parvalbumin B									
Set 6H vs 6A	108	0.22A	—	—	0.07A	0.04A	—	0.08A	—
Immunoglobulin Light Chains									
Rhe vs. Rei	113	★	88	1.09A	0.65A	0.19A	0.30(S)	0.70A	0.84A

H—Denotes helix-helix packing
S—Denotes sheet-sheet packing
M—Denotes mixed-helix-sheet packing
★—Denotes alignment based on structural insertion/deletion analysis

the details of structural modification or perturbation, it will become possible to learn how these macromolecules function and dysfunction in normal and diseased states.

This research was supported by National Institutes of Health grant in Biomathematics CA-22780.

Received for publication 17 December 1979.

REFERENCES

1. Phillips, D. C. British Biochemistry Past and Present, edit. Goodwin. Academic Press, London (1970).
2. Nishikawa, K., T. Ooi, Y. Isogai, and N. Saito. 1972. *J. Phys. Soc. Japan.* **32**:1331.
3. Kuntz, I. D. 1975. *J. Amer. Chem. Soc.* **97**:4362.
4. Rossmann, M. G., and A. Liljas. 1974. *J. Mol. Biol.* **85**:177.
5. Liebman, M. N. Submitted for publication.
6. McLachlan, A. 1979. *J. Mol. Biol.* **128**:49.
7. Rose, G. D. 1979. *J. Mol. Biol.* **134**:447.
8. Bernstein, F. C., T. F. Koetzle, G. J. B. Williams, E. F. Meyer, Jr., M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and T. Tasumi. 1977. *J. Mol. Biol.* **112**:535.