BUTTERWORTH
HEINEMANN

A novel representation of protein structure

Thomas W. Barlow and W. Graham Richards

Physical Chemistry Laboratory, Oxford, England

Using a nonlinear mapping technique, we demonstrate that proteins folded in two dimensions display the same overall structural features as their three-dimensional counterparts. The two-dimensional representation of protein structure provides a novel way to visualize structural as well as distance information. It may also provide a link for deriving three-dimensional structure from amino acid sequence.

Keywords: protein structure, nonlinear mapping, distance matrix

INTRODUCTION

The prediction of protein structure remains one of the grand challenge problems. Since the experiments of Anfinsen et al. involving protein denaturation and renaturation[1] it has been assumed that all the information required to predict the three-dimensional structure of a protein is inherent within the primary sequence of its amino acids. Yet attempts to predict three-dimensional protein structure have proved difficult. The problem is essentially one of transforming a one-dimensional pattern (amino acid sequence) into a three-dimensional pattern (tertiary structure). Is there a "half-way" pattern in two dimensions that might simplify the problem? In other words, can we usefully represent a three-dimensional protein structure in two dimensions?

One way of doing this is to create a distance matrix.[2–5] This is the matrix of all interamino acid distances in a protein. It is a useful tool both for comparing protein conformations and for identifying homology between different proteins. One valuable attribute of a distance matrix is that the three-dimensional parent structure can be readily regenerated using distance geometry techniques.[6,7]

Here, we demonstrate an alternative method that folds an amino acid chain in two-dimensional space. A nonlinear mapping technique is used to generate two-dimensional plots of protein structure in which the distances between amino acids reflect as closely as possible the corresponding distances in three dimensions. This novel representation provides structural information about a protein in a more intuitive way than does a distance matrix.

NONLINEAR MAPPING

The mapping is performed using an algorithm originally developed by Sammon[8] for multidimensional data analysis. It involves minimization of the difference between the distance matrix of the original structure and that of a novel two-dimensional structure.

A protein of N amino acids can be described by N three-dimensional vectors, $P_i = (x_i, y_i, z_i)$, $i = 1, \ldots, N$. P_i represents the α-carbon coordinates of residue i. We define N corresponding vectors in two-dimensional space: $W_i = (w_{i1}, w_{i2})$, $i = 1, \ldots, N$. Initially the two-dimensional coordinates are randomly assigned. Let d^*_{ij} be the distance between residue i and residue j in the three-dimensional structure. That is,

$$d^*_{ij} = [(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2]^{1/2} \qquad (1)$$

Let d_{ij} be the corresponding interamino acid distance in the two-dimensional representation.

We can now define an error, $E(m)$, which describes how well the interamino acid distances in the two-dimensional representation compare to those in the three-dimensional structure.

$$E(m) = \frac{1}{\sum\limits_{i<j}^{N} [d^*_{ij}]} \sum_{i<j}^{N} \frac{[d^*_{ij} - d_{ij}(m)]^2}{d^*_{ij}} \qquad (2)$$

This error can be minimized by an iterative process. In Eq. (2), m labels the iteration number. The two-dimensional vectors, W_i, are modified to minimize the error by a steepest descent method. The new components of W_p are given by

$$w_{pq}(m + 1) = w_{pq}(m) - \eta\Delta_{pq}(m) \qquad (3)$$

Color plates for this article are on p. 354.

Address reprint requests to Dr. Richards at the Physical Chemistry Laboratory, South Parks Road, Oxford OX1 3QZ, England.

Received 20 August 1995; accepted 28 September 1995.

655 Avenue of the Americas, New York, NY 10010

0263-7855/95/$10.00
SSDI 0263-7855(95)00083-6

where η is a learning rate parameter set at 0.3 and

$$\Delta_{pq}(m) = \frac{\partial E(m)}{\partial w_{pq}(m)} \Big/ \left| \frac{\partial^2 E(m)}{\partial w_{pq}(m)^2} \right| \tag{4}$$

Equation (3) is further constrained so that none of the two-dimensional vectors, W_i become identical. This would cause problems in the determination of the partial derivatives. The partial derivatives are readily determined as follows:

$$\frac{\partial E}{\partial w_{pq}} = \frac{-2}{\sum\limits_{i<j}^{N} d^*_{ij}} \sum_{\substack{j=1 \\ j\neq p}}^{N} \left[\frac{d^*_{pj} - d_{pj}}{d_{pj} d^*_{pj}}\right] (w_{pq} - w_{jq}) \tag{5}$$

$$\frac{\partial^2 E}{\partial w^2_{pq}} = \frac{-2}{\sum\limits_{i<j}^{N} d^*_{ij}} \sum_{\substack{j=1 \\ j\neq p}}^{N} \frac{1}{d_{pj} d^*_{pj}} \left[(d^*_{pj} - d_{pj}) - \frac{(w_{pq} - w_{jq})^2}{d_{pj}} \left(1 + \frac{d^*_{pj} - d_{pj}}{d_{pj}}\right) \right] \tag{6}$$

Repeated evaluation of Eq. (2) followed by modification of the two-dimensional coordinates by Eq. (3) produces a nonlinear two-dimensional mapping of the three-dimensional protein structure. The resulting two-dimensional coordinates have no direct physical significance. However, these coordinates can be plotted to obtain a two-dimensional representation of protein structure in which the distances between coordinates reflect the distances between α-carbons in the original structure. See Figure 1.[9]

EXAMPLES

In this study, the Sammon[8] software package has been used to construct a series of two-dimensional representations of three-dimensional protein structures obtained from the Brookhaven Protein Database.[10] In each case structural features inherent in three dimensions have been reproduced in the two-dimensional maps. Secondary structure was assigned for each protein using QUANTA.[11] These assignments are displayed in the two-dimensional figures. Actual numbers of each secondary structure type are recorded in Table 1.

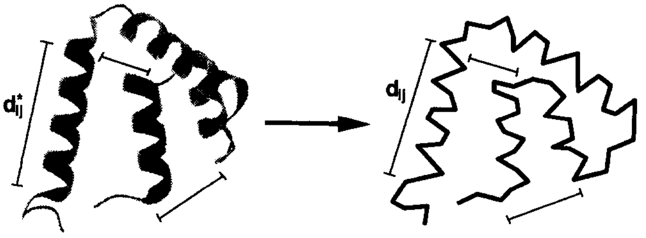

Figure 1. Nonlinear mapping of three-dimensional structure of rabbit uteroglobin.[9] Mapping preserves as much as possible the inter-C_α distances. For the two-dimensional representation on the right, the chain order is shown by joining each of the minimized two-dimensional coordinates with a line.

Color Plate 1 shows nonlinear plots of six β-sheet proteins.[12–18] The parallel strands making up the β-sheet secondary structure are easily distinguished in these two-dimensional representations. It is interesting to note that turns in a three-dimensional protein structure also correspond to turns in the nonlinear map.

The helix is a more difficult structure to transform from three dimensions to two. A sheet, after all, is virtually a two-dimensional structure to start off with. The nonlinear mapping process transforms the three-dimensional helix into a two-dimensional zig-zag. Color Plate 2 shows a series of α-helical proteins.[19–24] Maps of these proteins are readily differentiated from those of β-sheet proteins in Color Plate 1. Different families of α-helical proteins are also easily distinguished from one another. Thus hemoglobin and myoglobin maps are similar to one another, yet they are obviously different from the maps of other α-helical proteins.

Color Plates 1 and 2 show two-dimensional maps of protein structure for proteins that favor a particular secondary structure type. These proteins were chosen to see what happens to such structural features during the nonlinear mapping process. Perhaps surprisingly, these features were always preserved. Color Plate 3 shows a series of larger proteins[25,26] with mixed secondary structure. Again, structural features are easy to pick out. It is extraordinary that so much of the three-dimensional structure of a protein could be preserved in a two-dimensional representation. Color Plates 1–3 show only a sample of the proteins we have mapped. In a study of more than 40 different proteins, secondary structure is clearly maintained in every case.

Obviously any mapping that lowers the dimensionality of a problem will result in a loss of information. One way to determine the error introduced by the nonlinear mapping procedure is to compare the distance matrix of the two-dimensional structure with that of the original structure. Comparison of distance matrices confirms that although distances tend to be slightly compressed in the two-dimensional structure, it is clear that overall spatial relationships are preserved. This is demonstrated in Table 1 by the low values of the second rms error calculated for all distances greater than 12 Å.

CONCLUSIONS

In the introduction, we asked whether it was possible to usefully represent a three-dimensional protein structure in two dimensions; the answer would appear to be "yes." The Sammon algorithm provides a method for approximating three-dimensional protein structure by a two-dimensional plot. Using this novel representation, the overall structure of a protein as well as distance relationships between α-carbons become apparent at a glance. Comparison of the original distance matrix with the approximate one derived from the two-dimensional representation has shown them to be surprisingly similar. Secondary structure is also clearly maintained in the two-dimensional plot. We are currently working on a method that uses these plots as a stepping

Table 1. Mapped proteins,[a] with the number of residues classified by secondary structure type[b]

Code	Unit	Residue	Helix	Sheet	Coil	rms (Å)	rms12 (Å)
1REI	A	107	0	60	47	3.1	2.3
1PFC		111	4	35	72	2.3	1.7
2PAB	A	114	7	61	46	3.3	2.8
2RHE		114	6	58	50	3.4	2.6
2MCP	H	222	6	117	99	2.5	1.8
2STV		184	10	91	83	2.9	2.6
2MLT	A	26	23	0	3	0.6	0.5
1PPT		36	19	0	17	0.7	0.7
2CCY	A	127	86	2	39	2.7	2.0
2HMQ	A	113	77	0	36	2.8	2.0
1ECD		136	99	0	37	2.9	2.2
1MBD		153	112	0	41	2.8	2.1
1FX1		147	55	33	59	3.2	2.3
8API	A	340	105	120	115	4.6	3.8

[a] See Color Plates 1–3.

[b] The root mean square deviation (rms) has been calculated for each nonlinear map. rms = $(1/N)[\Sigma_{i \neq j}^{N} (d_{ij}^{*} - d_{ij})^2]^{1/2}$. The root mean square deviation for distances greater than 12 Å (rms12) has also been calculated.

stone between one-dimensional sequence and three-dimensional structure.

ACKNOWLEDGMENTS

Software for the Sammon nonlinear mapping algorithm was obtained from Teuvo Kohonen's ftp site cochlea.hut.fi (130.233.168.48) in the directory /pub/som_pak. The author is grateful for the support he receives from Kobe Steel Ltd. and from the ORS award scheme through CVCP.

REFERENCES

1 Anfinsen, C., Haber, E., Sela, M., and White, F.H., Jr. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proc. Natl. Acad. Sci. U.S.A.* 1961, **47,** 1309–1314

2 Phillips, D.C. Development of crystallographic enzymology. *Biochem. Soc. Symp.* 1970, **31,** 11–28

3 Nishikawa, K. and Ooi, T. Comparison of homologous tetiary structure of proteins. *J. Theoret. Biol.* 1974, **43,** 351–374

4 Liebman, M.N. Quantitative analysis of structural domains in proteins. *Biophys. J.* 1980, **32,** 213–215

5 Sippl, M.J. On the problem of comparing protein structures. Development and applications of a new method for the assessment of structural similarities of polypeptide conformations. *J. Mol. Biol.* 1984, **156,** 359–388

6 Havel, T.F., Kuntz, I.D., and Crippen, G.M. The theory and practice of distance geometry. *Bull. Math. Biol.* 1983, **45,** 665–720

7 Sippl, M.J. and Scheraga, H.A. Solution of the embedding problem and decomposition of symmetric matrices. *Proc. Natl. Acad. Sci. U.S.A.* 1985, **82,** 2197–2201

8 Sammon, J.W. A nonlinear mapping for data structure analysis. *IEEE Trans. Comput.* 1969, **C-18**(5), 401–409

9 Morize, I., Surcouf, E., Vaney, M.C., Buehner, M., and Mornon, J.P. Refinement of the c222 crystal form of oxidized uteroglobin at 1.34 angstroms resolution. *J. Mol. Biol.* 1987, **194,** 725–731.

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

11 QUANTA 4.0. Molecular Simulations, Inc. Waltham, Massachusetts, 1994

12 Epp, O., Lattman, E.E., Schiffer, M., Huber, R., and Palm, W. The molecular structure of a dimer composed of the variable portions of the Bence–Jones protein refined at 2.0 angstroms resolution. *Biochemistry* 1975, **14,** 4943–4952

13 Bryant, S.H., Amzel, L.M., Phizackerley, R.P., and Poljak, R.J. Molecular replacement structure of guinea pig igg1 pfc(prime) refined at 3.1 angstroms resolution. *Acta Crystallogr. Sect. B* 1985, **41,** 362–368

14 Blake, C.C.F., Geisow, M.J., Oatley, S.J., Rerat, B., and Rerat, C. Structure of prealbumin, secondary, tertiary and quaternary interactions determined by Fourier refinement of 1.8 angstroms. *J. Mol. Biol.* 1978, **167,** 339–356

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

16 Padlan, E.A., Cohen, G.H., and Davies, D.R. On the specificity of antibody/antigen interactions. *Ann. Immunol. (Paris), Sect. C* 1985, **136,** 271–283

17 Jones, T.A. and Liljas, L. Structure of satellite tobacco necrosis virus after crystallographic refinement at 2.5 angstroms. *J. Mol. Biol.* 1984, **177,** 735–767

18 Kabsch, W. and Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-

bonded and geometrical features. *Biopolymers* 1983, **22,** 2577–2637

19 Terwilliger, T.C. and Eisenberg, D. The structure of melittin. *J. Biol. Chem.* 1982, **257,** 6010–6015

20 Blundell, T.L., Pitts, J.E., Tickle, I.J., Wood, S.P., and Wu, C.-W. X-Ray analysis [1.4-angstroms resolution) of avian pancreatic polypeptide, small globular protein hormone. *Proc. Natl. Acad. Sci. U.S.A.* 1981, **78,** 4175–4179

21 Finzel, B.C., Weber, P.C., Hardman, K.D., and Salemme, F.R. Structure of ferricytochrome c(prime) from *Rhodospirillum molischianum* at 1.67 angstroms resolution. *J. Mol. Biol.* 1985, **186,** 627–643

22 Holmes, M.A. and Stenkamp, R.E. The structures of met and azidomet hermerythrin at 1.66 angstroms resolution. *J. Mol. Biol.* 1991, **220,** 723–735

23 Steigemann, W. and Weber, E. Structures of erythorcrugrin in different ligand states refined at 1.4 angstroms resolution. *J. Mol. Biol.* 1979, **127,** 309–338

24 Phillips, S.E.V. Structure and refinement of oxymyoglobin at 1.6 angstroms resolution. *J. Mol. Biol.* 1980, **142,** 531–554

25 Watenpaugh, K.D., Sieker, L.C., and Jensen, L.H. Flavodoxin at 2.0 angstroms resolution. *Proc. Natl. Acad. Sci. U.S.A.* 1973, **70,** 3857–3860

26 Engh, R., Loebermann, H., Schneider, M., Wiegand, G., and Huber, R. The s-variant of human alpha-1-antitrypsin, structure and implications for function and metabolism. *Protein Eng.* 1989, **2,** 407–419