Proc. Natl. Acad. Sci. USA
Vol. 82, pp. 5255–5259, August 1985
Biochemistry

Identical short peptide sequences in unrelated proteins can have different conformations: A testing ground for theories of immune recognition

(protein structure/sequence comparisons/antipeptide antibodies)

IAN A. WILSON*, DANIEL H. HAFT*, ELIZABETH D. GETZOFF*, JOHN A. TAINER*, RICHARD A. LERNER*, AND SYDNEY BRENNER†

*Department of Molecular Biology, Research Institute of Scripps Clinic, La Jolla, CA 92037; and †Medical Research Council Laboratory of Molecular Biology, Cambridge, CB2 2QH, United Kingdom

Contributed by Sydney Brenner, April 8, 1985

ABSTRACT **The ability of antibodies raised against disordered short peptides to interact frequently with their cognate sequences in intact folded proteins has raised a major theoretical issue in protein chemistry. We propose to address this issue by using antibodies raised against peptides with identical sequences, but different conformations, in pairs of unrelated proteins of known three-dimensional structure. The general search method presented here enabled us to detect candidate sequences for such immunological studies.**

Short synthetic peptides can generate antibodies that react with cognate sequences in intact folded proteins with a surprising degree of success (1–4). This finding raises the question of how such small oligopeptides, unlikely to have fixed conformations, can nevertheless select antibodies with high affinities for the more ordered and constrained sequences in intact proteins. Models which assume that polyclonal antisera are collections of antibodies with specificities to many conformations of the peptide immunogen do not readily provide an answer to this question. It seems that the result must be explained either by defined conformations adopted by the peptide in solution or when linked to a carrier or by some special property, either of the antigenic sequence in the protein or of the binding sites in the antibody. One postulate has been that antibodies made against a peptide have binding sites that recognize one or a few conformations of the peptide. Antibody reaction with the protein would depend on the ease with which the cognate sequence can assume one of these conformations. Thus, the most effective antigenic peptides would be those corresponding to the more mobile and flexible regions of the intact molecule. Recently, there have been reports that a number of protein epitopes are of this type (5, 6). Another, almost complementary, postulate is that the immune system evolved to defend the organism against infectious agents and would therefore have generated binding sites biased towards the recognition of natural structures including those of protein molecules. This makes the range and number of conformations of the immunogenic peptide irrelevant since the preexisting bias will ensure the selection of the appropriate set of conformations.

One way of testing these hypotheses is to use as test antigens identical peptide sequences found in different conformations in unrelated proteins. The first model predicts that one antibody will react with both conformations if local flexibility allows the sequence in both proteins to assume a similar structure; the second predicts that a single antibody would react with only one of the conformations of the identical sequence in different proteins.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Here, we analyze protein sequences and structures to search for examples that could be used to test these hypotheses.

METHOD

The 2511 amino acid sequences in the Dayhoff Protein Sequence Database‡ include a total of 470,158 residues, and, therefore, a similar number of short, overlapping peptides of given length (see Table 1). A FORTRAN program was designed to create from the database complete lists of all occurring peptides 4 to 10 residues long [in one-letter code (7)], together with the amino acid position in the protein sequence. Each list was alphabetized in less than 1 hr with the VAX/VMS SORT utility on a VAX 11-750 computer: all polypeptides that occurred more than once in the database appeared as consecutive repeated entries in the sorted list.

Redundancy of identical sequences of five or more residues within the database usually results from homology rather than analogy. Analogy, as we define here, refers to sequence identity in proteins for which there is no obvious long-range overlap or evolutionary relationship. To eliminate identities resulting from protein sequence homology, the database was divided into families of known or postulated homology, and the list was searched for repeated entries in different families.

RESULTS AND DISCUSSION

Analysis of the Protein Sequence Databank

Grouped by strong homology, the number of independent sequences reduces to about 500 different families, which still appear partially related when sequence comparisons are made. To identify truly analogous sequences in unrelated proteins, families with significant long-range sequence overlaps were grouped into 24 larger superfamilies, similar to a division already constructed for the Protein Sequence Database.‡

To compare the number of matching unrelated peptides found with the expected number from statistical probability, the number of independent peptides of given length (tetramer to octamer) present in the database must be estimated. Statistically, the expected number of analogous matches is roughly proportional to the square of the number of independent nonhomologous peptides of that same length in the database (Table 1). This effective size (n) is then the number of different sequences plus the number of analogous repeats

‡Barker, W. C., Hunt, L. T., Orcutt, B. C., George, D. G., Yeh, L. S., Chen, H. R., Blomquist, M. C., Johnson, G. C., Seibel-Ross, E. I. & Dayhoff, M. O. (1984) *Protein Sequence Database* (National Biomedical Research Foundation, Washington, DC).

Table 1. Analysis of short peptides in the Dayhoff Protein Sequence Database

	Tetramers	Pentamers	Hexamers	Heptamers	Octamers
Theoretical different	160,000	3,200,000	6.4×10^7	1.28×10^9	2.56×10^{10}
Total in database	459,701	456,684	453,723	450,797	447,900
Different in database	111,572	302,836	347,736	357,517	363,077
Unique in database	31,309	225,287	292,795	307,510	315,958
Repeated in database	80,263	77,549	54,941	50,007	47,119
Effective independent peptides in database	251,438	334,402	350,244	357,606	363,095
Most common sequence	GPPG	HGKKV	AHGKKV	KAHGKKV	KYIPGTKM
No. of occurrences	149	87	65	54	52
Identical sequences in superfamilies					
One occurrence	44,155	274,565	345,267	357,331	363,060
Two occurrences	28,997	25,319	2,433	183	16
Three occurrences	17,515	2,585	33	3	1
Four occurrences	9,936	311	3	0	0
More than four occurrences	10,969	56	0	0	0
Expected occurrences of two or more	268,785	25,673	1,521	86	5

The total number of tetramer to octamer sequences was calculated from the individual sequences present in the database. The number of theoretically possible different sequences is based on the 20 common amino acids in proteins. The total number of different sequences found in the database illustrates, except for tetramers, the limited sample of sequences in the database compared to the number possible. Most longer sequences are represented only once, and almost all the repeats of sequences are in homologous proteins. The most common sequence for tetramers is in collagen, that for pentamers to heptamers is in hemoglobin, and that for octamers is in cytochrome *c*. Analysis of numbers of identical sequences found in unrelated proteins of the 24 superfamily groups is indicated and compared to the statistical expectation of finding such identical sequences.

only. For example, the effective size is about 75% of the total number of hexapeptide sequences in the database (Table 1). For a database with an effective size of n independent m-mers and equal odds for all 20 amino acids, the expected number of analogous matches is $\frac{1}{2}n(n-1)(0.05)^m$ or approximately half the square of the effective size, divided by the number of

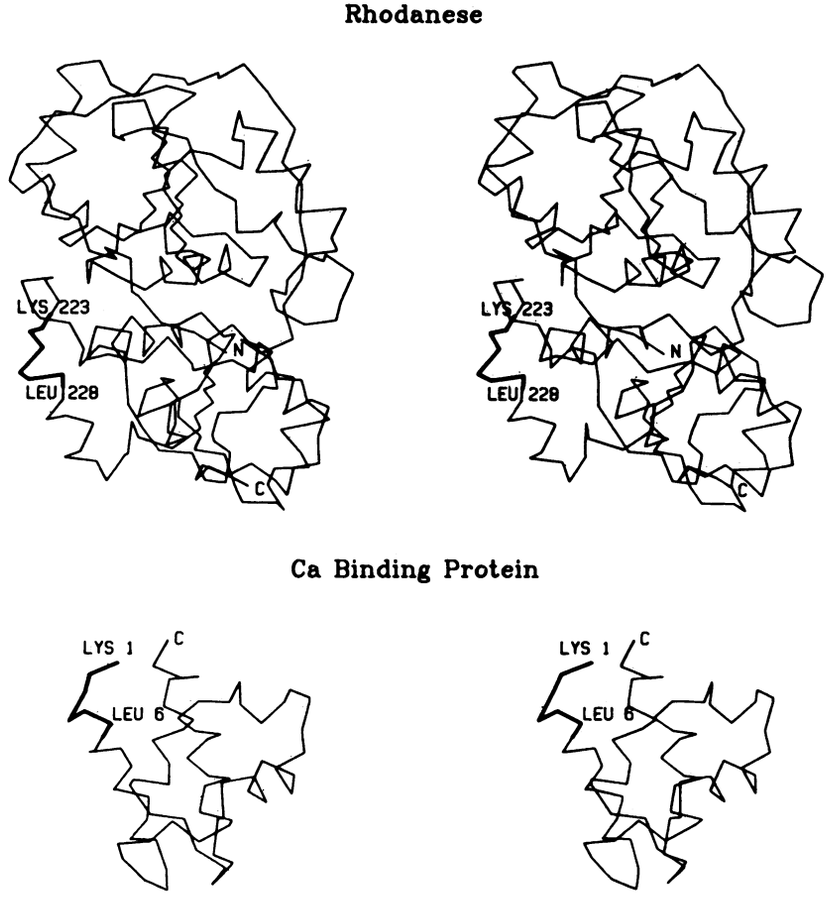

FIG. 1. Comparison of the conformation of the sequence KSPEEL in rhodanese (13) and in calcium-binding protein (12). Stereo traces of the polypeptide chains (α carbons only) show the peptide location in thicker lines and residue numbers in their respective proteins. Both of the peptide conformations are helical. The coordinates are taken from the Brookhaven databank. Figs. 1 and 2 were generated from representations on the Evans and Sutherland multipicture system, using the GRAMPS (14) and GRANNY (15) computer programs.

m-mers possible. In the case of hexamers, this prediction value is 958. A somewhat corrected calculation (used in Table 1), using the same effective size but the global rms frequency of amino acids rather than 1/20 to calculate match odds, predicts $\frac{1}{2}n(n-1)(0.054)^6 = 1521$ analogous hexapeptide matches. If a local rather than global rms frequency could be used for each protein, the prediction would be higher, and probably closer to the frequencies actually found in the database (Table 1).

The number of identical sequences represented at least twice by analogy among the 24 superfamilies was 17 octamers, 186 heptamers, 2469 hexamers, 28,271 pentamers, and 67,417 tetramers. Except for tetramers, these numbers agree well with, although they exceed, the predictions shown in Table 1. This excess is caused in part by highly degenerate sequences containing mainly hydrophobic, basic, or proline residues (Table 2). Otherwise, the excess over expectation does not appear to warrant postulation of structural or functional bias for particular amino acid sequences.

Our analysis is more extensive than a related study published recently (8) on a database of 10,000 residues. The longest matching sequence found in that study (which was intentionally confined to proteins of known three-dimensional structure) was a pentamer, although some unpublished data indicated some matching hexamers.

Analysis of Identical Hexamers and Heptamers in Unrelated Proteins with Known Structures

To compare the three-dimensional conformations of identical sequences in unrelated proteins, a list of identical sequences was compiled for the 223 coordinate sets of 79 different proteins with either published and deposited coordinates [Brookhaven protein databank, 1984; (9)] or published high-resolution structures. Only 16 pairs were identified that had both an amino acid sequence and a backbone trace of the polypeptide chain. Detailed comparisons of polypeptide conformation, including side chains, were possible for only five pairs of these sequences.

These matching sequence pairs can be grouped generally into either like or unlike structures. This approach is somewhat subjective, but when coordinates were available, a least-squares optimization was computed for the superposition of the polypeptide α-carbon atoms, main-chain atoms, or both. Quantitative evaluation of the similarity of the different peptide conformations showed that the group of like structures superimposed such that their rms difference in equivalent atomic position was approximately 2 Å or less. The unlike structures had an rms difference in equivalent position of greater than 3 Å.

The like structures (Table 3) consisted mainly of α-helical structures, with six out of eight pairs of identical sequences having helical conformations in both proteins. For example, the sequence Lys-Ser-Pro-Glu-Glu-Leu (KSPEEL) was α-helical in both calcium-binding protein (12) and rhodanese (13) (Fig. 1). On the other hand, the sequence Gln-Val-Lys-Glu-Val-Lys was β-strand in both glutathione reductase (16) and inorganic pyrophosphatase (17) (Table 3). Examples of very different secondary structure for identical hexamers were also found. The sequence Asn-Ala-Ala-Ile-Arg-Ser (NAAIRS) was α-helical in phosphofructokinase (18) but β-strand in thermolysin (19) (Fig. 2). Similarly, the sequence Val-Glu-Leu-Ile-Arg-Gly was helical in influenza neuraminidase (20) but β-strand in tobacco mosaic virus coat protein (21).

Not only the conformation of the same sequence in unrelated proteins but also the accessibility of these sequences in the protein structure is important. Generally speaking, the accessible surface areas (22) calculated for the identical peptides found in unrelated proteins were approximately equivalent (data not shown). However, the surface areas of the matching peptides (calculated for monomers) of thermolysin and phosphofructokinase differed significantly.

Table 2. Identical octamers and larger sequences with two or more occurrences in unrelated proteins

Sequence	No. of first residue	Protein and origin
Repetitive sequences		
DAAAAAAATAAA	59	Hexon-associated protein (IX), adenovirus 7 and 3
DAAAAAAATAA	26	Antifreeze peptide 4 precursor, flounder
AAAAAAATAAA	71	Antifreeze peptide 4 precursor, flounder
AAAAAAAT	717	Late 100K protein, adenovirus 5 and 2
AAAAAAAT	28	Cytochrome *c* peroxidase precursor, baker's yeast
APAPAPAP	13	Myosin L1 (A1) and L4 (A2), chicken
	201	Outer membrane protein A precursor, *Escherichia coli*
EPEPEPEP	189	Early 32K, 26K, and 13K proteins, adenovirus 2 and 5
	75	Gene *tonB* protein, *E. coli*
VPPPPPPP	387	Terminal protein precursor, adenovirus 2
	24	Proline-rich peptide P-B, human
Nonrepetitive sequences		
AAKPKKAA	173	Histone H1, trout
	28	Nonhistone chromosomal protein H6, rainbow trout
AFLLLLSL	193	Cytochrome oxidase polypeptide I, baker's yeast
	11	Gene *E* protein, bacteriophage G4
AYLVGLFE	98	Histone H3, bovine, etc.
	241	Hypothetical *cobA* intron protein, *Aspergillus nidulans*
EPVPGDPD	359	Coat protein VP1, polyoma virus
	89	Ribulose bisphosphate carboxylase, maize
GVANLDNL	74	Globin β chain, musk shrew
	637	Large T antigen, polyoma virus
IPKDIQLA	130	Acetolactate synthase II, *E. coli*
	119	Histone H3(4), mouse
IPSGVDAG	46	Plastocyanin, vegetable marrow
	146	α-Crystallin A chain, bovine, etc.
IWYNNNVI	60	*Ea47* gene protein, bacteriophage λ
	401	Triacylglycerol lipase, porcine
LSSSTQAS	344	Gene *IV* protein, bacteriophages fd, M13, and f1
	175	Heat-shock-related 70C protein, fruit flies
TFISRHNS	182	Large T antigen, simian virus 40
	4	Gene *G* protein, bacteriophage ϕX174
VLLLSLIG	73	Tetracycline resistance protein (transposon Tn*10*), *E. coli*
	4	α-Amylase precursor, pancreatic, mouse

The computer search of the Dayhoff protein sequence databank identified 15 octamers present in unrelated proteins. The sequence number and identification of the individual proteins (or families) are tabulated. No shorter sequences than octamers are included due to the high numbers of matching sequences found in shorter sequence lengths (see Table 1). The sequences are shown in one-letter code (7).

Table 3. Analysis of identical peptides in unrelated proteins of known structure

Peptide	Hydrophobicity	Type: Predicted	Type: Protein structure	No. of first residue	Protein and origin	Resolution, Å	PDB code	Dayhoff code
					Like structures			
AALKAA	0.34	A	A	133	Subtilisin, *Bacillus amyloliquefaciens*	2.5	2SBT	SUBS N
			A	258	Glyceraldehyde-3-*P* dehydrogenase, *Bacillus stearothermophilus*	2.7	0GD1	DEBSGF
DALKAAG	0.14	A	A	138	Myoglobin, sea hare *Aplysia limacina*	3.6	0MBA	GGGA-A
			A	157	L-Arabinose-binding protein, *E. coli*	2.4	1ABP	RBEC
DGAILV	0.62	L	B	100	Elongation factor Tu, *E. coli*	2.6	0ETV	EFECT
			B	20	Thioredoxin, *E. coli*	2.8	1SRX	TXEC
DGLAHL	0.32	A	A	38	Glyceraldehyde-3-*P* dehydrogenase, *B. stearothermophilus*	2.7	0GD1	DEBSGF
			A	73	Globin β chain, human	2.1	1HHO	HBHU
KSEAQA	−0.34	A	A	27	Ribonuclease, *B. amyloliquefaciens*	2.2	0RNB	NRBS N
			A	146	Nuclease, *Staphylococcus aureus* Foggie strain	1.5	2SNS	NCSA F
KSPEEL	−0.33	A	A	1	Calcium-binding protein, bovine intestinal	2.0	1ICB	KLBOI
			A	223	Thiosulfate S-transferase (rhodanese), bovine	2.5	1RHD	ROBO
QVKEVK	−0.41	A	B	250	Glutathione reductase, human erythrocyte	2.0	2GRS	RDHUU
			B	132	Inorganic pyrophosphatase, baker's yeast	3.0	1PYP	PWBY
SVVRKAI	−0.01	A	A	22	Phosphofructokinase, *B. stearothermophilus*	2.4	0PFK	KIBSFF
			A	18	D-Galactose-binding protein, *E. coli*	3.0	0GBP	JGEC
					Unlike structures			
FTAFKD	0.09	L	L	111	Mn superoxide dismutase, *B. stearothermophilus*	2.9	—	DSBSNF
			A	103	Dihydrofolate reductase, *Lactobacillus casei*	1.7	3DFR	RDLBD
GIIQGD	0.33	L	A	116	D-Galactose-binding protein, *E. coli*	3.0	0GBP	JGEC
			L	51	Coat protein, satellite tobacco necrosis virus	3.0	—	VCTN-S
NAAIRS	−0.14	AL	B	97	Thermolysin, *Bacillus thermoproteolyticus*	1.6	3TLN	HYBS T
			A	17	Phosphofructokinase, *B. stearothermophilus*	2.4	0PFK	KIBSFF
PGTAPK	−0.04	L	L	64	Thioredoxin, *E. coli*	2.8	1SRX	TXEC
			BL	42	Immunoglobulin λ V-I region, human Newm	2.0	3FAB	L1HUNM
TNNQRI	−0.60	L	A	34	Phosphoglycerate kinase, horse muscle	3.0	2PGK	KIHUG
			L	181	Glyceraldehyde-3-*P* dehydrogenase, *B. stearothermophilus*	2.7	0GD1	DEBSGF
TYGTGS	0.16	L	B	118	Pepsinogen A, porcine	3.0	1PEP	PEPG
			L	452	Neuraminidase, influenza virus A/NT/60/68	2.9	—	NMIV2
VELIRG	0.12	A	B	424	Neuraminidase, influenza virus A/NT/60/68	2.9	—	NMIV2
			A	130	Coat protein, tobacco mosaic virus vulgare	2.8	—	VCTMVU
YLTIHS	0.34	B	A	231	Citrate synthase, porcine	2.7	1CTS	YKPG
			B	190	Carboxypeptidase B, bovine	2.8	1CPB	CPBOB

Pairs of identical hexapeptides and heptapeptides are shown for proteins whose three-dimensional structures have been published and/or deposited in the Brookhaven databank. The Brookhaven databank (PDB) code, Dayhoff protein sequence code, number of the first residue, and reported resolution of the three-dimensional structure determination are given. Type refers to the conformation of the peptide in the protein structure or its predicted conformation (10): A, α-helix; B, β-strand; and L, bend or loop. The average hydrophobicity per residue (11) for each peptide is also shown.

The hexamer sequence, NAAIRS, in thermolysin (646 Å^2, 3-Å radius probe) is part of a central strand of a five-stranded β-sheet on the protein surface (Fig. 2). In phosphofructokinase, the same hexamer sequence (495 Å^2 for the monomer) is in the subunit interface and in an α-helix sandwiched between two other helical structures. In another case, similar accessible surface areas are present for the same hexapeptide sequence in calcium-binding protein and rhodanese (Fig. 1), even in very different locations of the polypeptide chain. In calcium-binding protein, the sequence KSPEEL is the amino terminus but in rhodanese the peptide sequence is in the middle of the structure. Nevertheless, both sequences are relatively exposed in the monomers although the sequence in calcium-binding protein is significantly less tethered due to its amino-terminal location.

Clearly, the peptide secondary structure in the intact protein is influenced both by its neighboring amino acids and by tertiary interactions from nonadjacent amino acids. Secondary structure predictions based on empirical parameters (10) show that, for about half of these identical peptides in unlike structures, the predicted secondary structure correlates with the actual structure when adjacent residues are included in the calculation. In these instances, the adjacent residues are important in influencing the secondary structure found in the different proteins, whereas in others the tertiary interactions must influence the peptide conformation in the folded protein.

This study has shown that common sequences of up to eight residues do occur in unrelated proteins. However, sequence identity does not ensure similarity of shape (Figs. 1 and 2). Hence, sequence-specific antibodies can be generated to test binding to identical sequences contained in unrelated proteins. Such information can be used to test theories on immune recognition and to explore the importance of conformation, accessibility, hydrophobicity, hydrophilicity, and mobility in antigenic recognition.

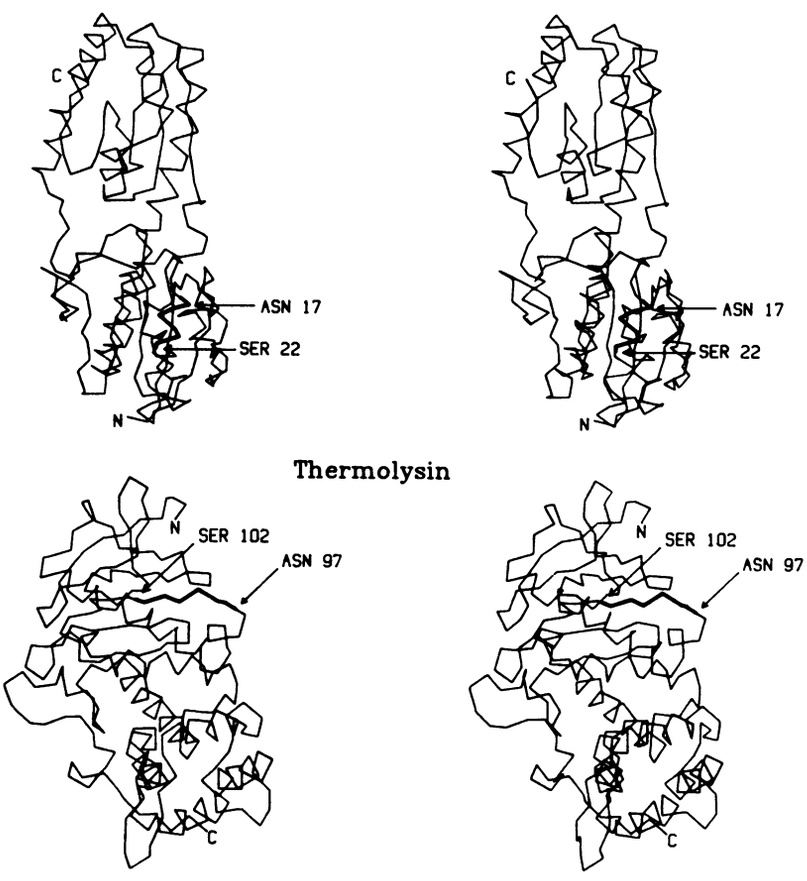

FIG. 2. Comparison of the conformation of the sequence NAAIRS in phosphofructokinase (18) and in thermolysin (19). Conventions as in Fig. 1. The peptide conformation is α-helical in phosphofructokinase and β-strand in thermolysin. The thermolysin coordinates are taken from the Brookhaven databank and the phosphofructokinase coordinates were kindly provided by Philip Evans (Medical Research Council Laboratory of Molecular Biology, Cambridge, England).

We thank Dr. Philip Evans, Dr. Anne Bloomer, Dr. T. Alwyn Jones, Dr. Martino Bolognesi, and Dr. Frances Jurnak for generously providing unpublished coordinates; Dr. James Hogle for helpful advice; Andrew R. Cherenson and Daniel Bloch for technical assistance; and Ann McDonald for manuscript preparation. Partial support from National Institutes of Health Grant AI 19499 (to I.A.W. and R.A.L.) is acknowledged. This is publication 3697-MB from the Research Institute of Scripps Clinic.

1. Green, N., Alexander, H., Olson, A. J., Alexander, S., Shinnick, T. M., Sutcliffe, J. G. & Lerner, R. A. (1982) *Cell* **28,** 477–487.
2. Niman, H. L., Houghten, R. A., Walker, L. E., Reisfeld, R. A., Wilson, I. A., Hogle, J. M. & Lerner, R. A. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 4949–4953.
3. Wilson, I. A., Niman, H. L., Houghten, R. A., Cherenson, A. R., Connolly, M. L. & Lerner, R. A. (1984) *Cell* **37,** 767–778.
4. Pfaff, E., Kuhn, C., Schaller, H., Leban, J., Thiel, H.-J. & Bohm, H.-O. (1985) in *Vaccines 85*, eds. Lerner, R. A., Chanock, R. M. & Brown, F. (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY), pp. 199–202.
5. Westhof, E., Altschuh, D., Moras, D., Bloomer, A. C., Mondragon, A., Klug, A. & Van Regenmortel, M. H. V. (1984) *Nature (London)* **311,** 123–126.
6. Tainer, J. A., Getzoff, E. D., Alexander, H., Houghten, R. A., Olson, A. J., Lerner, R. A. & Hendrickson, W. A. (1984) *Nature (London)* **312,** 127–134.
7. IUPAC–IUB Commission on Biochemical Nomenclature (1969) *Biochem. J.* **113,** 1–4.
8. Kabsch, W. & Sander, C. (1984) *Proc. Natl. Acad. Sci. USA* **81,** 1075–1078.
9. Bernstein, F. C., Koetzle, T. F., William, G. J. B., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977) *J. Mol. Biol.* **112,** 535–542.
10. Chou, P. Y. & Fasman, G. D. (1978) *Adv. Enzymol. Relat. Areas Mol. Biol.* **47,** 45–149.
11. Eisenberg, D., Weiss, R. M., Terwilliger, T. C. & Wilcox, W. (1982) *Faraday Symp. Chem. Soc.* **17,** 109–120.
12. Szebenyi, D. M. E., Obendorf, S. K. & Moffat, K. (1981) *Nature (London)* **294,** 327–332.
13. Ploegman, J. H., Drent, G., Kalk, K. H. & Hol, W. G. J. (1978) *J. Mol. Biol.* **123,** 557–594.
14. O'Donnell, T. J. & Olson, A. J. (1981) *Comput. Graph.* **15,** 133–142.
15. Connolly, M. L. & Olson, A. J. (1985) *Comput. Chem.*, in press.
16. Thieme, R., Pai, E. F., Schirmer, R. H. & Schulz, G. E. (1981) *J. Mol. Biol.* **152,** 763–782.
17. Arutiunian, E. G., Terzian, S. S., Voronova, A. A., Kuranova, I. P., Smirnova, E. A., Vainstein, B. K., Hohne, W. E. & Hansen, G. (1981) *Dokl. Akad. Nauk. SSSR* **258,** 1480–1486.
18. Evans, P. R. & Hudson, P. J. (1979) *Nature (London)* **279,** 500–504.
19. Holmes, M. A. & Matthews, B. W. (1982) *J. Mol. Biol.* **160,** 623–639.
20. Varghese, J. N., Laver, W. G. & Colman, P. M. (1983) *Nature (London)* **303,** 35–40.
21. Bloomer, A. C., Champness, J. N., Bricogne, G., Staden, R. & Klug, A. (1978) *Nature (London)* **276,** 362–368.
22. Connolly, M. L. (1983) *Science* **221,** 709–713.