Proc. Natl. Acad. Sci. USA
Vol. 80, pp. 1964–1968, April 1983
Evolution

Modular structural units, exons, and function in chicken lysozyme

(module/compact structural unit/origin of intron/protein evolution)

MITIKO GŌ

Department of Biology, Faculty of Science, Kyushu University 33, Fukuoka 812, Japan

Communicated by Walter Gilbert, January 5, 1983

ABSTRACT **By the application of the same algorithm for finding compact structural units encoded by exons as applied previously to hemoglobin, five units, M1–M5, were identified in chicken egg white lysozyme. They consist of residues 1–30, 31–55, 56–84, 85–108, and 109–129, respectively. I call these compact structural units "modules." As in hemoglobin, modules thus identified correspond well to exons—i.e., modules M1, M2 plus M3, M4, and M5 correspond to exons 1, 2, 3, and 4 of the lysozyme gene, respectively. Localization of the catalytic sites glutamic acid-35 and aspartic acid-52 on the module M2 suggests that this module might have worked as a functional unit in a primitive lysozyme. The good correspondence between exons and modules reinforces the idea of "proteins in pieces," which was derived from the fact of "genes in pieces." The evolutionary origin of the introns in globins and lysozyme is discussed.**

Intervening sequences (introns) interrupt the coding regions (exons) of many eukaryotic genes (1). The origin of intervening sequences during evolution is one of the most interesting subjects in molecular biology (2–4). The evolutionary advantage of the existence of introns has been discussed by Gilbert and Tonegawa and colleagues (5, 6). Their hypothesis is that exons correspond to functional units of protein molecules and new functional proteins have evolved by selection of various combinations of the functional units that are produced by unequal crossing-over on introns.

On the other hand, a protein must possess a stable specific conformation to carry out its biological function. Blake argued that, if exons encode structural units as well as functional units, then combinations of such exons would have the advantage of producing stable functional proteins (7).

I have proposed a method to define compact structural units as least extended conformations in globular proteins (8), and now I call these compact structural units "modules." Four modules, F1–F4, have been identified in Hb α- and β-chains. Modules F1, F2 plus F3, and F4 have been shown to correspond to exons 1, 2, and 3 (9–12), respectively, in the genes of the mouse. This finding led to the prediction that one more intron may have been present in an ancestral globin at the position corresponding to the junction between modules F2 and F3—i.e., somewhere in a region between residues 66 and 71 of the Hb α-chain. Shortly after the report of this correlation, a third intron was found by Jensen *et al.* (13) exactly in the predicted region—i.e., between codons 68 and 69, in the gene of leghemoglobin from soybean. This protein is believed to be an ancestral form of Mb and of Hb α- and β-chains (14, 15). Therefore, it has been revealed that the four exons in the leghemoglobin gene correspond exactly to the modules F1–F4, respectively, identified in the Hb α- and β-chains. It is desirable to see whether the correspondence between modules and exons is a phenomenon specific to globins or is a common one seen also in other proteins.

Jung *et al.* have determined the nucleotide sequence of the chicken egg white lysozyme gene (16). They found that the gene has four exons, corresponding to residues −18 to 28, 28 to 82, 82 to 108, and 108 to 129. Residues −18 to −1 compose the signal peptide of prelysozyme; thus three introns are located on the codons 28, 82, and 108 within the coding regions for lysozyme. X-ray analysis of chicken egg white lysozyme by Blake *et al.* (17) made clear the three-dimensional structure of the enzyme.

METHODS AND RESULTS

Five Modules of Chicken Egg White Lysozyme. Atomic coordinates of chicken egg white lysozyme determined by x-ray crystallographic studies (17) were supplied by the Protein Data Bank (18). By using the coordinates of the C^α atoms, protein folding structure can be presented on a two-dimensional plane by the so-called C^α–C^α distance map (19–22). The modules, or the least extended structural units, were defined by inspecting the pattern of C^α–C^α distance map (8). Modules are defined in two steps. First, the residue pairs separated more than a certain distance (23 Å was used in the case of

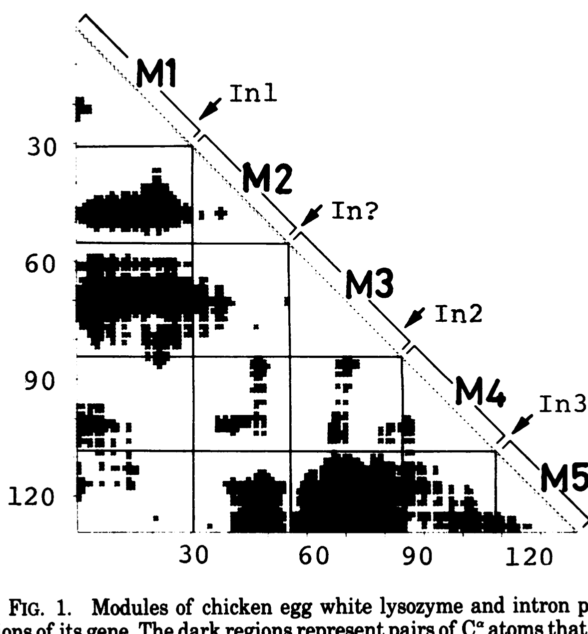

FIG. 1. Modules of chicken egg white lysozyme and intron positions of its gene. The dark regions represent pairs of C^α atoms that are separated more than 23 Å. Both ordinate and abscissa are residue numbers. Five modules, M1–M5, are identified by using the least extended criterion—i.e., by drawing a pair of horizontal and vertical straight lines that meet on the diagonal in the map in such a way as to keep away from the dark regions. Intron positions (16) are marked by arrows, together with the predicted position of another intron at the junction between modules M2 and M3.

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.

chicken egg white lysozyme; see Fig. 1) are marked and shown as dark areas. These areas will be called "well-separated regions." Second, a pair of horizontal and vertical straight lines is drawn to go through points of a certain residue number i on ordinate and abscissa, respectively. These two straight lines meet at the same point on the diagonal. Then the residue number i is moved over all residue numbers, and those residue numbers for which the pair of straight lines scarcely crosses the well-separated (shaded) regions are searched for. The C^{α}–C^{α} distance map of the chicken egg white lysozyme is shown in Fig. 1. Five segments, M1–M5, are characterized as modules by drawing lines somewhere between residues 27–35,

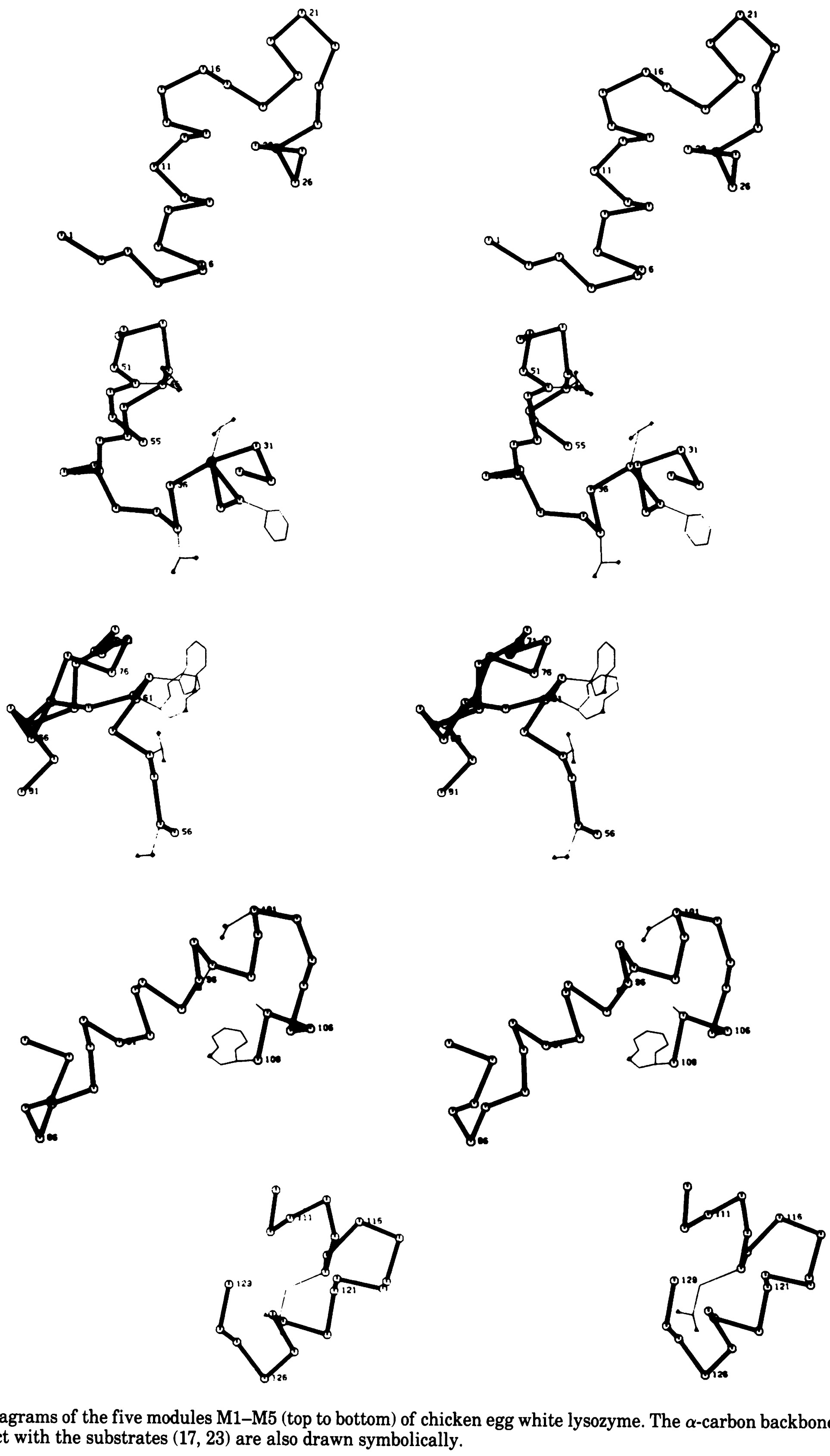

FIG. 2. Stereo diagrams of the five modules M1–M5 (top to bottom) of chicken egg white lysozyme. The α-carbon backbone is represented. The side chains in contact with the substrates (17, 23) are also drawn symbolically.

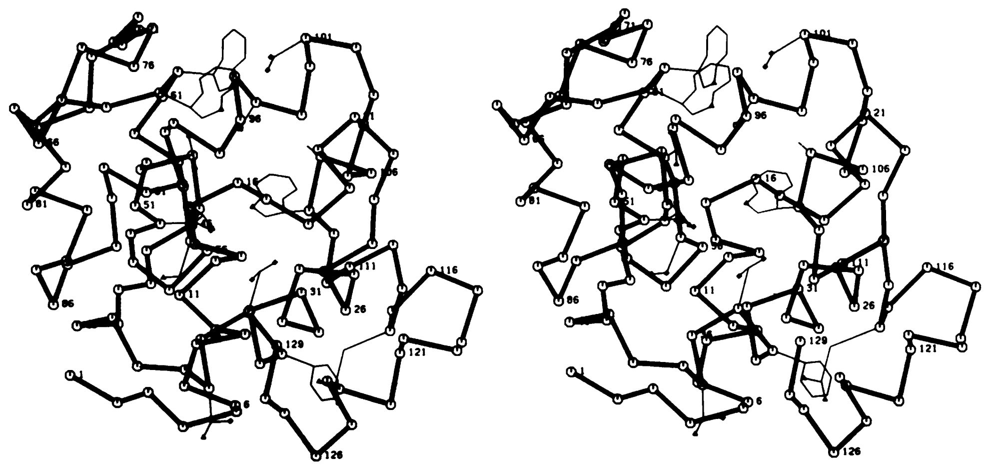

FIG. 3. Stereo diagram of the whole chain of chicken egg white lysozyme. See legend for Fig. 2.

53–57, 80–88, and 105–111. In Fig. 1 modules M1–M5 are defined as the segments 1–30, 31–55, 56–84, 85–108, and 109–129, respectively.

Junctions Between the Modules Correspond to the Intron Positions Except for the Junction Between M2 and M3. The residues 28, 82, and 108 at the intron positions (16) fall exactly in the regions of intermodule joints, 27–35, 80–88, and 105–111 (Fig. 1). The stereo diagrams of the modules M1–M5 and of the whole chain of chicken egg white lysozyme are shown in Figs. 2 and 3, respectively. The side chains that bind a substrate are also drawn symbolically in the figures. The character of the modules as the least extended conformational units is easily recognized in Fig. 2.

No intron is found in the region of the junction between modules M2 and M3—i.e., in the region of residues 53–57. The correspondence between the modules and the exons is perfect except that M2 plus M3 corresponds to exon 2. This pattern of correspondence is quite similar to that of Hb α- and β-chains (9–12, 24–26), in which two contiguous modules, F2 and F3, correspond to one central exon (8).

Most Modules Are Linked to Each Other by Disulfide Bonds. The sizes of the five modules M1–M5 are 28, 27, 26, 27, and 21 residues, respectively. Those modules have close relationships with the locations of S—S bonds and with the secondary structures. Fig. 4 is a schematic representation of the modules, distribution of S—S bonds, and the secondary structures observed in lysozyme (17). Four S—S bonds are located between residues 6 and 127, 30 and 115, 64 and 80, and 76 and 94. It is worth mentioning that three out of the four S—S bonds bridge different modules, and only one is located within a module. One side of the modules in small one-layer polypeptide chains, such as lysozyme and Hb α- and β-chains, is exposed to the surface and therefore should be mainly polar; the other side should be mainly nonpolar. By assembling preexisting modules in such a way that their nonpolar sides come in contact with each other, the modular structure of proteins has the advantage of yielding new stable globular proteins in the process of evolution (8). In chicken egg white lysozyme, the stability of the assembly of the five modules is enhanced by the four S—S bonds linking each of those modules together. Lysozyme is an extraordinarily stable protein due to the four disulfide bridges (27). It appears that although the disulfide bridges are not necessary to determine the three-dimensional structure in lysozyme, they do con-

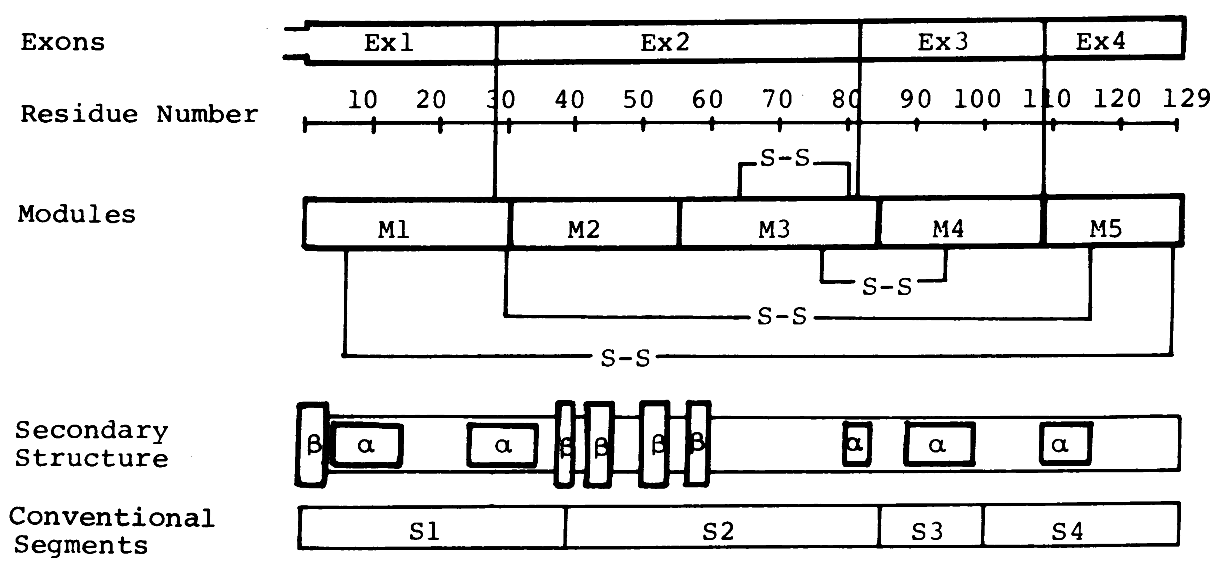

FIG. 4. Correlations between the structure of chicken egg white lysozyme and the exons in its gene. The diagram shows the segments encoded by the exons (16), modules, S—S bonds, α-helix and β-sheet structures (27), and the conventional structural segments (17, 28–31). It can be seen that the junctions between modules correspond to the intron positions (shown as junctions between the exons) better than the junctions of the conventional structural segments except for the absence of an intron corresponding to the junction between M2 and M3.

Table 1. Modules, segments encoded by exons, and functions

Module	Residues	Segment encoded by exon*	Catalytic site†	Contact sites of the substrate†	Contact rings of the substrate†
M1	1–30	1–28			
M2	31–55	28–82	Glu-35, Asp-52	Phe-34, Glu-35, Asn-37, Asn-44, Asp-52	D, E, F
M3	56–84			Gln-57, Asn-59, Trp-62	C, D, E
M4	85–108	82–108		Ile-98, Asp-101, Ala-107, Trp-108	A, B, C
M5	109–129	108–129		Arg-114	F

* From Jung *et al.* (16).
† From Imoto *et al.* (27) and Kelly *et al.* (23).

tribute significantly to the maintenance of the stable native conformation once it is formed (32).

All of the four junctions between the modules M1, M2, M3, M4, and M5 fall on segments of α-helix or β-sheet. The boundary M1–M2 (residues 30–31) is located on an α-helix. The boundaries M3–M4 (residues 84–85) and M4–M5 (residues 108–109) fall on the COOH and NH_2 termini of α-helices. The boundary M2–M3 (residues 55–56) is located on the turn that links two strands (residues 50–54 and 57–60) into an antiparallel β-sheet.

Catalytic Sites Are Localized on One Module and Binding Sites Are Located on Different Modules. Lysozyme attacks many bacteria by dissolving the mucopolysaccharide of the cell wall (33). The bacterial cell wall polymer is an alternating polymer of *N*-acetylglucosamine (GlcNAc) and *N*-acetylmuramic acid (MurNAc). Lysozyme cuts an alternating -GlcNAc-MurNAc-GlcNAc-MurNAc- polymer between C1 of MurNAc and the chain-linking oxygen.

Jung *et al.* (16) discussed the exon–intron boundaries of the lysozyme gene in relation to the function of lysozyme. They noted that exon 2, which encodes Trp-28 through Ala-82, carries the catalytic center Glu-35 and Asp-52 and also the surrounding functionally important residues that exist in the three-dimensional structure on both sides of the crevice.

Exon 2 corresponds to modules M2 (residues 31–55) plus M3 (residues 56–84). However, both of the catalytic sites Glu-35 and Asp-52 are localized on module M2. This localization of both of the catalytic sites on module M2 implies a possible role of this module as a functional unit in a primitive lysozyme.

Contact sites to the hexasaccharide substrate have been deduced by x-ray crystallographic analysis and by model building (23). The modules M1–M5 are shown in Table 1 in relationship to the locations of the catalytic and contacting sites as well as contacting rings of the substrate. The enzyme cuts the bond between carbohydrate rings D and E of the substrate. Comprehensive analysis of the thermodynamic quantities and of the nature and number of the interatomic contacts between enzyme and saccharide bound (27) shows that the substrate binds to the enzyme predominantly at ring sites C–E. The modules M2 and M3 contribute mainly to bind the substrate, the modules M4 and M5 contribute to a lesser amount, and the module M1 contributes nothing to the substrate binding.

DISCUSSION

Five structural modules, M1–M5, defined as compact structural units from the tertiary structure of chicken egg white lysozyme, have been shown to have a remarkable correspondence with the exons of its gene, as shown in Fig. 1 and Table 1. However, no intron has been found in the joint region of the modules M2 and M3—i.e., residues 53–57. I expect that an intron exists or existed at the corresponding position in the ancestral gene of chicken egg white lysozyme or in the lysozyme genes of contemporary lysozymes other than chicken egg white.

The conventional structural segments of chicken egg white lysozyme from its tertiary structure are four: residues 1–39, 40–85, 86–100, and 101–129 (17, 28–31). They are determined by watching the three-dimensional structure with particular attention to contacts between segments and by taking into account the secondary structures. The correlation between these structural segments and the exon/intron structure of chicken egg white lysozyme gene was discussed by Jung *et al.* (16). The comparison between the modules and the conventional structural segments in chicken egg white lysozyme is given schematically in Fig. 4 in relation to their correlation with the exon/intron structure of its gene. The number of structural segments is the same as the number of exons. In contrast to the correlation between intron positions and intermodule regions except for the absence of an intron corresponding to the junction between M2 and M3, however, the boundaries between the conventional structural segments, 39–40, 85–86, and 100–101, correlate less with the intron positions, 28, 82, and 108.

Modules are defined as compact structural units in a globular protein or in a protein domain. Monomeric proteins and subunits of oligomeric proteins are often subdividable into structural domains (34–36). Structural domains are somewhat separated in space from other parts of a protein and can be easily recognized by inspection of its three-dimensional structure. Structural domains are usually made of 100–200 residues. A module is a contiguous piece of polypeptide chain that assumes a compact or least extended conformation in a small protein or in a domain of a large protein. A module itself is less globular than a simple protein or a domain. The contact area between the modules is not small, because modules assemble to form a compact globular structure. Lengths of modules are in the range of 20–40 residues in Hb and lysozyme.

The good correspondence between the exons and the modules in Hb and in lysozyme almost excludes the possibility that the introns were inserted after a contiguous gene had been completed during an early stage of biological evolution. The probability of introns' being inserted by pure chance at the exact positions corresponding to the junctions of modules is extremely small. Instead, the presence of exons as mini-genes coding protein modules in proto-organisms in early stages of evolution is more likely (2, 3). In the following period of biological evolution, some introns at positions corresponding to intermodules may have been lost. In the globin gene an intron at the position corresponding to the junction of modules F2 and F3 is found in the leghemoglobin gene (13) but not in the genes of Hb α- and β-chains. It could have been lost during evolution either on the lineage from the leghemoglobin–Mb ancestor to Mb or on the lineage from the Mb–Hb ancestor to the Hb α-chain–β-chain ancestor (14). The analysis of the Mb gene will give the answer to this question in the future.

Does a protein module encoded by an exon maintain by itself the stable compact conformation observed in native protein? Wetlaufer discussed the advantage of rapid self-assembly by a modular folding process in proteins (31). From a collection of experimental studies on the folding of protein fragments, Wetlaufer concluded that the minimal size of a peptide that can assume a stable compact structure is in the range of 20–40 residues (37). This minimal size of peptides coincides with the size of the modules observed in globin and lysozyme. A module may have no rigid specific conformation but instead show structural softness, as observed in glucagon, which is 29 residues long (38). Conformation of modules in a native protein may be stabilized only by the interactions with other modules, but the propensity for the conformation to be realized in a native protein seems to be carried in the module itself. Experimental work on the folding of the modules of globin and lysozyme is desirable to resolve the question of intrinsic stability and softness of the modules.

I thank Mr. H. Mizumo for the use of the computer programs for the distance map and stereo diagram and Profs. W. Gilbert, H. Matsuda, K. Hamaguchi, and Dr. N. Gō for valuable discussions and encouragement. This work was supported in part by grants-in-aid from the Ministry of Education, Japan.

1. Breathnach, R. & Chambon, P. (1981) *Annu. Rev. Biochem.* **50**, 349–383.
2. Darnell, J. E., Jr. (1978) *Science* **202**, 1257–1260.
3. Doolittle, W. F. (1978) *Nature (London)* **272**, 581–582.
4. Crick, F. (1979) *Science* **204**, 264–271.
5. Gilbert, W. (1978) *Nature (London)* **271**, 501.
6. Tonegawa, S., Maxam, A. M., Tizard, R., Bernard, O. & Gilbert, W. (1978) *Proc. Natl. Acad. Sci. USA* **75**, 1485–1489.
7. Blake, C. C. F. (1978) *Nature (London)* **273**, 267.
8. Gō, M. (1981) *Nature (London)* **291**, 90–92.
9. Leder, A., Miller, H. I., Hamer, D. H., Seidman, J. G., Norman, B., Sullivan, M. & Leder, P. (1978) *Proc. Natl. Acad. Sci. USA* **75**, 6187–6191.
10. Nishioka, Y. & Leder, P. (1979) *Cell* **18**, 875–882.
11. Konkel, D. A., Tilghman, S. M. & Leder, P. (1978) *Cell* **15**, 1125–1132.
12. Konkel, D. A., Maizel, J. V., Jr., & Leder, P. (1979) *Cell* **18**, 865–873.
13. Jensen, E. O., Paludan, K., Hyldig-Nielsen, J. J., Jorgensen, P. & Marcker, K. A. (1981) *Nature (London)* **291**, 677–679.
14. Dayhoff, M. O., Hunt, L. T., McLaughlin, P. J. & Jones, D. D. (1972) in *Atlas of Protein Sequence and Structure*, ed. Dayhoff, M. O. (Natl. Biomed. Res. Found., Washington, DC), Vol. 5, p. 2.
15. Hyldig-Nielsen, J. J., Jensen, E. O., Paludan, K., Wiborg, O., Garrett, R., Jorgensen, P. & Marcker, K. A. (1982) *Nucleic Acids Res.* **10**, 689–701.
16. Jung, A., Sippel, A. E., Grez, M. & Shutz, G. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 5759–5963.
17. Blake, C. C. F., Koenig, D. F., Mair, G. A., North, A. C. T., Phillips, D. C. & Sarma, V. R. (1965) *Nature (London)* **206**, 757–761.
18. Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Jr., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977) *J. Mol. Biol.* **112**, 535–542.
19. Phillips, D. C. (1970) in *British Biochemistry, Past and Present*, ed. Goodwin, T. W. (Academic, London), pp. 11–28.
20. Rao, S. T. & Rossmann, M. G. (1973) *J. Mol. Biol.* **76**, 241–256.
21. Nishikawa, K., Ooi, T., Isogai, Y. & Saito, N. (1974) *J. Phys. Soc. Jpn.* **32**, 1331–1337.
22. Kuntz, I. D. (1975) *J. Am. Chem. Soc.* **97**, 4362–4366.
23. Kelly, J. A., Sielecki, A. R., Sykes, B. D., James, M. N. G. & Phillips, D. C. (1979) *Nature (London)* **282**, 875–878.
24. van Ooyen, A., van den Berg, J., Mantei, N. & Weissmann, C. (1979) *Science* **206**, 337–344.
25. Patient, R. K., Elkington, J. A., Kay, R. M. & Williams, J. G. (1980) *Cell* **21**, 565–573.
26. Partington, G. A. & Baralle, F. E. (1981) *J. Mol. Biol.* **145**, 463–470.
27. Imoto, T., Johnson, L. N., North, A. C. T., Phillips, D. C. & Rupley, J. A. (1972) in *The Enzymes*, ed. Boyer, P. D. (Academic, New York), 3rd Ed., Vol. 7, pp. 665–868.
28. Liljas, A. & Rossmann, M. G. (1974) *Annu. Rev. Biochem.* **43**, 475–507.
29. Crippen, G. M. (1978) *J. Mol. Biol.* **126**, 315–332.
30. Rose, C. D. (1979) *J. Mol. Biol.* **134**, 447–470.
31. Wetlaufer, D. B. (1973) *Proc. Natl. Acad. Sci. USA* **70**, 697–701.
32. Thornton, J. M. (1981) *J. Mol. Biol.* **151**, 261–287.
33. Salton, M. R. J. & Ghuysen, J. M. (1959) *Biochim. Biophys. Acta* **36**, 552–554.
34. Rossmann, M. G. & Argos, P. (1981) *Annu. Rev. Biochem.* **50**, 497–532.
35. Schulz, G. E. & Schirmer, R. H. (1979) *Principles of Protein Structure* (Springer, New York), pp. 84–95.
36. Thomas, K. A. & Schechter, A. N. (1980) in *Biological Regulation and Development*, ed. Goldberg, R. F. (Plenum, New York), Vol. 2, pp. 43–100.
37. Wetlaufer, D. B. (1981) in *Advances in Protein Chemistry*, eds. Anfinsen, C. B., Edsall, J. T. & Richards, F. M. (Academic, New York), Vol. 34, pp. 61–92.
38. Epand, R. M. (1972) *Biochemistry* **11**, 3571–3575.