Proc. Natl. Acad. Sci. USA
Vol. 75, No. 1, pp. 390–394, January 1978
Genetics

Silkmoth chorion proteins: Sequence analysis of the products of a multigene family

(protein sequences/protein homology)

JEROME C. REGIER*, FOTIS C. KAFATOS*, ROBERT GOODFLIESH†‡, AND LEROY HOOD†

* Cellular and Developmental Biology, Harvard University Biological Laboratories, Cambridge, Massachusetts 02138, and † Division of Biology, California Institute of Technology, Pasadena, California 91125

Communicated by E. B. Lewis, September 2, 1977

ABSTRACT **Five polypeptide components have been isolated from the eggshell (chorion) of a silkmoth. Two are homogeneous on sodium dodecyl sulfate and isoelectric focusing gels, and three contain predominantly two proteins each. Amino acid analyses show that all five components are similar to each other. These proteins have been sequenced from the amino terminus. Homogeneous components yielded single sequences; heterogeneous components yielded two residues at some positions, consistent with their containing two major electrophoretic components. Striking similarities are apparent among all these sequences. These similarities can be increased dramatically by separating each of the three protein mixtures into two sequences and introducing a small number of gaps or insertions. This is due in part to bringing into register a portion that contains short repeating subunits found in all sequences. All proteins are also characterized by a region of high cysteine content near the amino terminus followed by a longer low-cysteine region. The data suggest that these proteins share a common evolutionary origin and are encoded by a multigene family.**

This report describes striking similarities among the NH_2-terminal sequences of a set of functionally related structural proteins found in the eggshell (chorion) of a silkmoth, *Antheraea polyphemus*. These similarities in sequence suggest that the proteins are related in an evolutionary sense and must be encoded by genes that have evolved by reduplication followed by diversification brought about by nucleotide substitutions and rearrangements.

Recently it has been proposed (1) that certain groups of functionally related RNAs or proteins may be encoded by families of genes that are physically clustered and evolutionarily related (informational multigene family). The immunoglobulin genes are a prototype of this type of gene family (1). Chorion genes are apparently linked (M. R. Goldsmith, personal communication) and thus, according to the evidence presented here, would seem to be a second example of an informational multigene family. This type of family is of substantial interest because it may permit rapid evolution without unduly high genetic load and because it may support the high information content necessary for some of the most complex aspects of the differentiated phenotype in eukaryotic organisms.

The sequence analyses of chorion proteins are part of a broad study of the insect chorion (2). Ultimately, we hope to understand the organization of the chorion genes, as related to their pattern of expression during development (3, 4), the degree of their evolutionary relatedness, and their functional similarity. The evolutionary relatedness is being investigated by protein sequencing, which is proceeding in parallel with sequencing studies on chorion DNA (5); the latter has been synthesized enzymatically with chorion mRNA as the template and cloned in bacteria, by using a procedure known to be of high fidelity (6).

We are also investigating the functional relationship between primary structure of the proteins and supramolecular structure of the chorion. To date, 15 proteins have been localized to specific substructures of the chorion. A structure-function analysis in this system should be of general interest because the chorion is a highly ordered protein assembly with many ultrastructural features common in the extracellular materials of vertebrates, invertebrates, and plants.

METHODS

Chorion Protein Purification. Pure eggshell fragments were obtained from laid eggs of the American silkmoth, *A. polyphemus*, and the eggshell (chorion) proteins were solubilized, carboxamidomethylated, and fractionated by differential precipitation and column isoelectric focusing (7). The purified components were characterized by sodium dodecyl sulfate ($NaDodSO_4$)/gel electrophoresis, gel isoelectric focusing, and amino acid analysis and named as described (7).

Automated Sequence Analyses. Automated sequence analyses of polypeptides were performed on Beckman sequencers (models 890 A and B, both updated, and 890C); 200–500 nmol of protein was sequenced per run.

Protein mixture A1,1- -a2,3 was sequenced twice, to positions 45 and 76. A modified 0.1 M Quadrol program (8) as described by Terhorst *et al.* (9) was used. Protein A4- -c1 was sequenced twice, through residues 70 and 86. Protein a4- -d1 was sequenced twice, to positions 26 and 52; a modified, *N,N*-dimethylbenzylamine program (10) and the 0.1 M Quadrol program were used. Protein mixture A2,3- -d5,6 was sequenced once through position 26 by using the dimethylbenzylamine program and twice through positions 28 and 40 by using the 0.1 M Quadrol program. Protein mixture A3,5- -d9 was sequenced twice through position 47 by using the dimethylbenzylamine program and once through position 45 by using the 0.1 M Quadrol program. The sequence fractions were identified by several independent methods as described (7).

RESULTS

Characterization of purified components

Fig. 1 (*left*) shows the isoelectric focusing profile of total chorion proteins and indicates the bands purified, identifying them according to a scheme already described (7). For identification, the isoelectric focusing pattern is divided according to landmark bands into segments, each named by a lower case letter (a to j, basic to acidic); individual bands within each segment are

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

Abbreviation: $NaDodSO_4$, sodium dodecyl sulfate.

‡ Present address: Upjohn Company, Kalamazoo, MI.

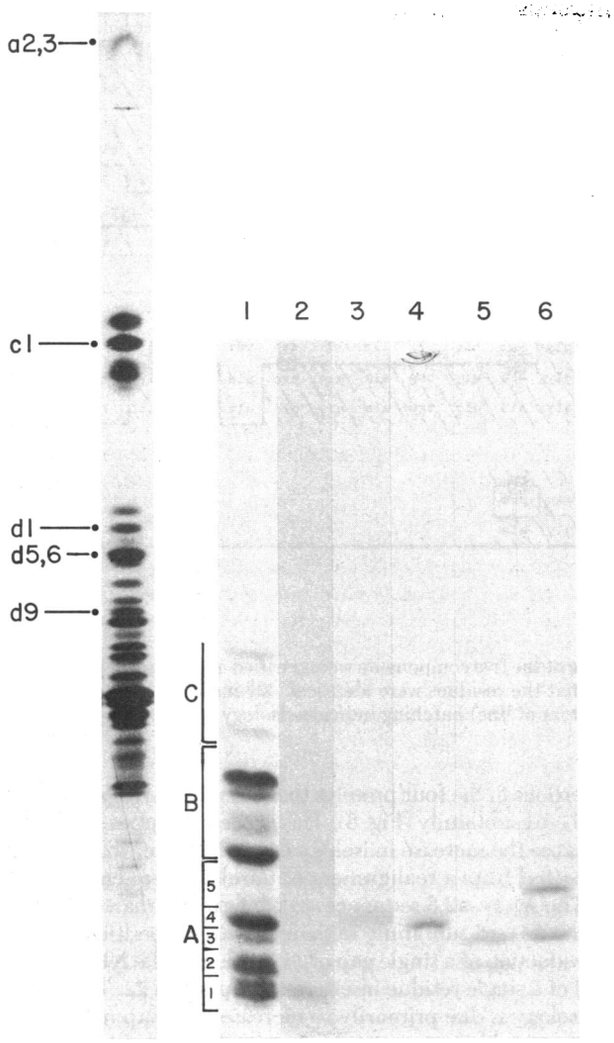


FIG. 1. Electropherograms of chorion proteins. (*Left*) Total chorion proteins fractionated on an isoelectric focusing gel. The five bands purified are indicated by dots. (*Right*) Electrophoresis on $NaDodSO_4$ gels. Lanes: 1, unfractionated proteins; 2, a2,3; 3, c1; 4, d1; 5, d5,6; 6, d9.

further identified by sequential numbering (e.g., band c1). Two of the purified fractions sometimes resolve into doublets (a2,3 and d5,6).

The chorion proteins of *A. polyphemus* have been resolved on $NaDodSO_4$ gels into three major classes—called A, B, and C, in order of increasing molecular weight (11). The molecular weights of these classes of proteins range from approximately 7000 to 18,000. These classes have been further divided into subclasses—A1, A2, . . ., also in order of increasing molecular weight. The five purified fractions were characterized by $NaDodSO_4$/polyacrylamide gel electrophoresis (Fig. 1 *right*) and they will be referred to henceforth by a combined code that indicates the subclasses of their major constituents on $NaDodSO_4$ gels followed by their identification on isoelectric focusing gels (e.g., A4- -c1).

Four of the five purified fractions were found to contain exclusively class A proteins after resolution on $NaDodSO_4$ gels (Fig 1 *right*). Two fractions (A4- -c1 and A4- -d1) appeared homogeneous both on isoelectric focusing and on $NaDodSO_4$ gels. The remaining fractions each contained two major components on $NaDodSO_4$ gels and one or two additional minor components. In one of these fractions (A2,3- -d5,6), a minor component was a class C protein. Fraction A1,1- -a2,3 contained two major components, both in subclass A1, but these were not resolved in the gel shown in Fig. 1 *right*. In total, the five purified fractions contained, as major constituents, eight A proteins, two each in subclasses A1, A3, and A4 and one each in subclasses A2 and A5.

Table 1. Amino acid compositions of *A. polyphemus* chorion proteins

	Residues per 100 residues					
	Unfractionated	A1,1- -a2,3	A4- -c1	A4- -d1	A2,3- -d5,6	A3,5- -d9
CM-Cys	6.4	9.8	9.4	10.0	8.4	10.3
Asp*	3.7	1.3	3.6	2.1	3.3	2.9
Thr	3.0	3.4	2.7	3.1	3.0	3.8
Ser	3.7	1.2	3.6	3.5	4.7	3.4
Glu*	4.5	4.0	3.6	4.5	4.3	3.1
Pro	4.4	1.9	3.6	4.0	4.4	4.8
Gly	32.6	39.4	31.5	30.9	28.6	29.7
Ala	12.1	10.8	12.9	14.1	14.7	15.3
Val	6.5	7.3	7.1	7.3	7.3	6.9
Met	0.4	0.0	0.0	0.0	0.2	0.0
Ile	3.8	4.4	3.5	4.0	4.4	3.8
Leu	7.6	5.6	7.4	6.8	6.3	6.1
Tyr	6.4	6.9	6.1	5.4	5.5	6.2
Phe	1.4	1.0	0.9	0.9	1.2	1.0
His	0.0	0.0	0.0	0.0	0.1	0.0
Lys	0.5	0.9	0.9	1.0	0.7	0.9
Trp	1.0	0.0	0.8	0.8	0.8	0.8
Arg	2.3	2.1	2.5	1.7	2.0	1.0
$GluNH_2$	0.0					
$GalNH_2$	0.0	0.0	0.0	0.0	0.0	0.0

* Includes the amidic forms.

Amino acid analyses of the five purified fractions are summarized in Table 1, along with an analysis of unfractionated chorion proteins for comparison. Both the unfractionated and the purified chorion proteins were unusually rich in cysteine, glycine, alanine, and tyrosine (7, 12). Relative to the unfractionated sample, all of the purified fractions were enriched in cysteine and deficient in methionine.

NH_2-terminal sequences

The NH_2-terminal sequences of eight chorion proteins included in the five purified fractions are shown in Fig. 2. Comparison of the two homogeneous components (A4- -c1 and A4- -d1) reveals striking sequence similarities (isologies, as defined in ref. 13): the first 27 residues are identical, and there are only four differences in the first 51 residues, three of them conservative.

The three fractions containing more than one chorion protein (Fig. 1 *right*) also contained multiple residues at several positions. In almost all cases, no more than two residues could be identified definitely; this is consistent with the electropherograms which indicated that each of these fractions only contained two major bands. In only a few positions was a third residue identifiable, and invariably it was minor compared to the two other residues. These minor residues probably reflect the minor protein contaminants and were not considered further. In positions where only one residue could be identified, it was assumed that the two sequences were identical. We think the possibility of failing to identify a second residue over the first 25 positions is extremely unlikely because of the independent methods of residue identification used and the multiple sequencer runs performed on each component. The possibility of not detecting a second residue becomes more likely as the sequencer run continues, depending on the particular residue.

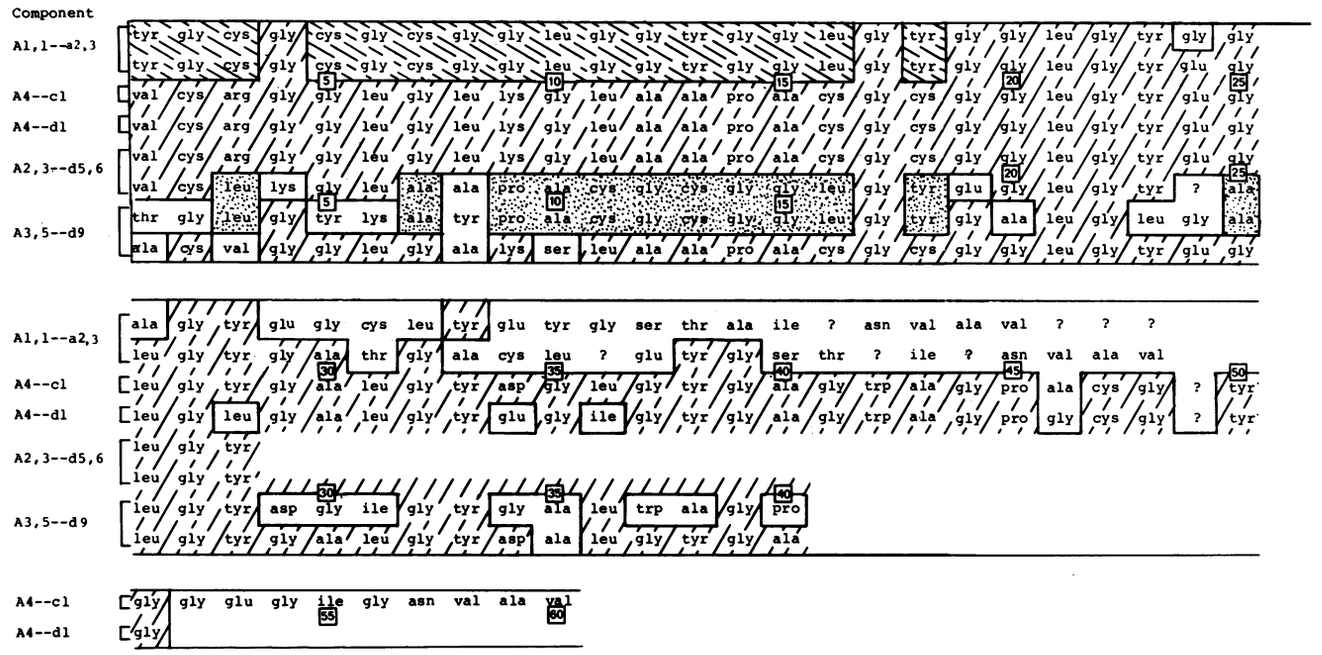


FIG. 2. NH_2-Terminal sequences of *A. polyphemus* chorion proteins. Three of the five components were purified and sequenced as mixtures. Failure to detect two residues at individual positions was assumed to mean that the residues were identical. "?" indicates uncertainty in the identification of a residue. Right-leaning (top of line farther to right than bottom of line) hatching indicates isology with A4- -c1 and A4- -d1; left-leaning hatching and stippling indicate additional isologies.

In all cases we sequenced beyond the positions given in Fig. 2 but report here only unambiguously identified residues.

Accurate separation of mixtures into discrete sequences requires careful quantitation of recoveries at each position where there are multiple residues, particularly if the proteins are present in similar proportions. For some amino acids, quantitation is inherently unreliable due to variable recoveries. For consistency, we have not used quantitative analyses to separate the sequences of mixtures. Instead, the separation shown in Fig. 2 corresponds to that which best maximizes isology with A4- -c1 for at least one of the two components. Although this separation is somewhat arbitrary, in many cases it is supported by quantitation of recoveries. Furthermore, it is interesting that this separation also reveals isology between the two remaining sequences of A2,3- -d5,6 and A3,5- -d9.

Significant isologies of all eight sequences are clearly evident in Fig. 2. Of the first 28 residues, four positions are identical in all eight sequences, and six positions vary in only one of the sequences. Protein A4- -c1 contains several short repeated segments (7): Gly-Leu-Gly-Tyr is present three times and the closely related tetrapeptide Ala-Leu-Gly-Tyr occurs once between residues 20 and 38. Of the 10 invariant or nearly invariant positions, 8 are within these tetrapeptide repeats.

Similarities are more pronounced among certain components. A4- -c1 and A4- -d1 only differ in 4 of 51 positions. One of the components of A2,3- -d5,6 is identical to both of these proteins in the first 27 positions and one of the components of A3,5- -d9 only differs in 5 of 40 residues. We will designate these four proteins the A4- -c1 subfamily. The two components of A1,1- -a2,3 are quite different from this subfamily and much more similar to each other (identical in the first 23 positions and largely different thereafter). The remaining two sequences (one each from A2,3- -d5,6 and A3,5- -d9) bear some resemblance to the A4- -c1 subfamily but even more to each other. Thus, three "subfamilies" can be distinguished among the eight partial sequences.

The similarity of the eight sequences in Fig. 2 can be dramatically increased by introducing a small number of gaps or insertions in the four proteins that are most divergent from the A47- -c1 subfamily (Fig. 3). This procedure appears to be valid because the increase in isology is much greater than would be expected from a realignment of unrelated sequences.

The A1,1- -a2,3 sequences exhibit greater than 70% isology to the A4- -c1 subfamily in their first 28–30 positions after the introduction of a single gap of 11 residues at the NH_2-terminus and of a single residue insertion after position 22. The increase of isology is due primarily to increased overlap in the region containing the tetrapeptide repeats. Beyond this region, introduction of a single gap in each of the two A1,1- -a2 sequences (five and three residues, respectively) makes them more similar to the A4- -c1 subfamily. Without the gap, virtually no isology is apparent (Fig. 2). Preliminary data suggest that this staggered isology continues even beyond the 70th position.

The two other sequences (one each from A2,3- -d5,6 and A3,5- -d9) which appear to be quite different from the A4- -c1 subfamily can be made similar by the introduction of one or two gaps, totaling five residues in each case (Fig. 3). Again, these rearrangements bring the repeating internal peptides into register with the corresponding regions of the A4- -c1 subfamily.

In summary, all eight chorion sequences are very similar. They can be divided into three subfamilies, that are internally even more similar, and are distinguished by the existence of characteristic gaps and insertions (Fig. 3).

DISCUSSION

Homology

Chorion proteins from the silkmoth *A. polyphemus* are a biochemically complex group and can be resolved into at least 33 components on $NaDodSO_4$ gels and 41 on isoelectric focusing gels (7). A subset of these proteins (called s/s), identified by similar solubility, molecular weight, and amino acid composition, appear by these criteria to be more similar to each other than to other chorion proteins (7).

In this paper we show that the NH_2-terminal sequences of

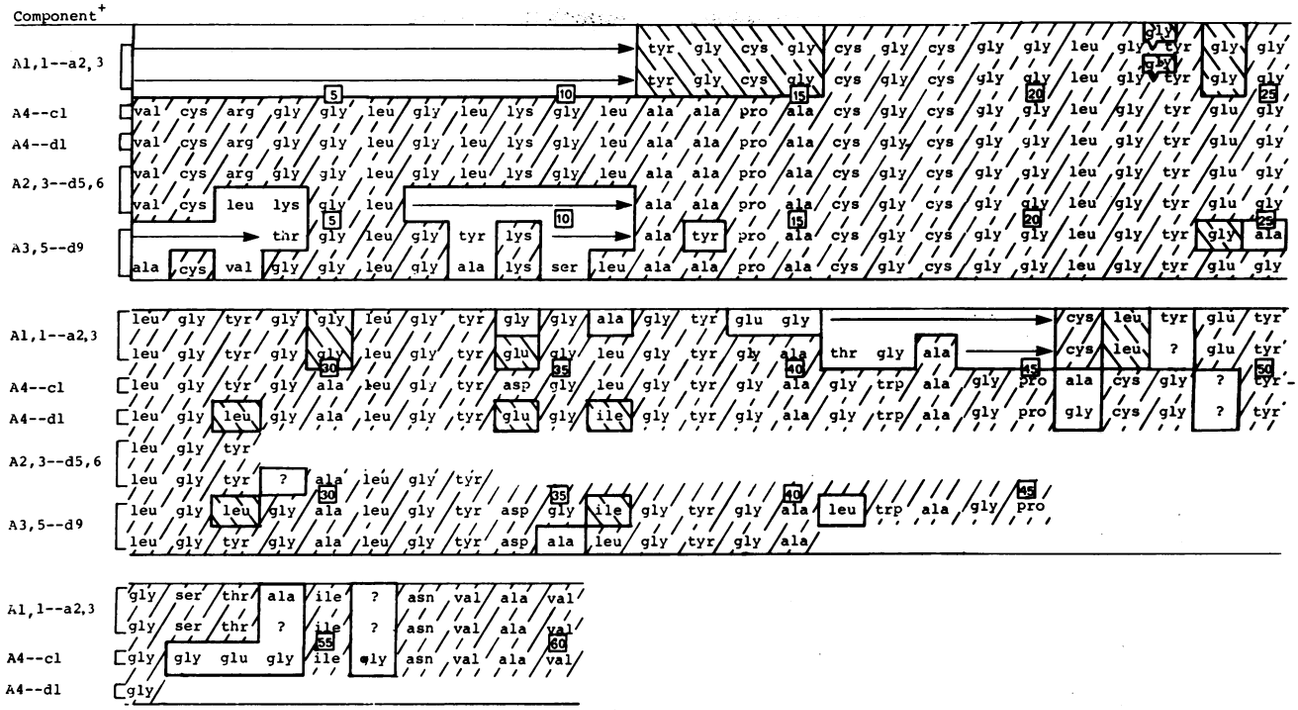


FIG. 3. Aligned NH_2-terminal sequences of *A. polyphemus* chorion proteins. Gaps (→) or insertions (↷) were introduced into the sequences in Fig. 2 to maximize isology with A4- -c1. Hatching as in Fig. 2.

several major components of the s/s fraction also appear to be very similar to each other. There can be little doubt that these proteins are derived from a common evolutionary origin: the isologies are so extensive that they must be interpreted as homologies. The proteins belong to the A class, defined by size upon electrophoresis in $NaDodSO_4$/polyacrylamide gels (11); although A2,3- -d5,6 as purified contains a minor amount of a C protein, almost certainly this impurity is not represented in the derived sequences. It is notable that, despite the existence of some polymorphism in the chorion, the highly isologous proteins A4- -c1 and A4- -d1 are major and constant chorion constituents (7). Thus, the results indicate that many (and possibly all) A proteins in A. *polyphemus* are homologous.

The differences that do exist in the eight sequences in Fig. 3 are of two kinds: amino acid substitutions and gaps. It is premature to suggest whether the gaps represent deletions or insertions (e.g., whether the 11 NH_2-terminal amino acids of the A4- -c1 subfamily were present in the ancestral protein and were lost during the evolution of A1,1- -a2,3 or whether these residues were absent in the ancestral protein and were inserted during the evolution of the remaining six proteins). Deducing the probable sequence and length of the ancestral protein will require complete sequence analysis of as many present-day relatives as possible and the construction of an evolutionary dendrogram (14). In any case, the evolution of A proteins must have been based on the existence of multiple gene copies, which were diversified by nucleotide substitutions and deletions or insertions.

The gaps introduced in Fig. 3 account for sizable portions of the molecular weight differences between the respective proteins and the A4- -c1 subfamily. For example, each A1,1- -a2,3 protein has a mass of approximately 7000 daltons and a length of approximately 80 residues; A4- -c1 has a mass of approximately 10,000 daltons and, from preliminary sequence data, a length of 112 residues. Thus, approximately one-third of the difference in length is already accounted for by the gaps shown in Fig. 3. In the two other cases, the observed gaps are less extensive, and correspondingly the proteins have higher molecular weights than A1,1- -a2,3.

Structure-function

Important features (7) of the sequence of protein A4- -c1 are the presence of a cysteine-rich tail followed by a cysteine-poor interior, and the presence of internal repeats, including the dipeptide Cys-Gly and the tetrapeptides Gly-Leu-Gly-Tyr and Pro-Ala-Cys-Gly. Both of these features are present in all chorion proteins sequenced, suggesting that they are functionally important and that the proteins in which they are found have similar functions. The most constant region includes the first Gly-Leu-Gly-Tyr repeat (residues 20–23 in Fig. 3). preceded by two or three repeats of Cys-Gly. The first Pro-Ala-Cys-Gly repeat (residues 14–17) is also highly conserved (six of eight proteins), but the second one appears to be variable.

Closer analysis of the sequences suggests that a substantial portion of them can be arranged into tandem pentapeptide segments. Omitting the A1,1--a2,3 sequences, which appear to be more divergent than the rest, the segments are as follows:

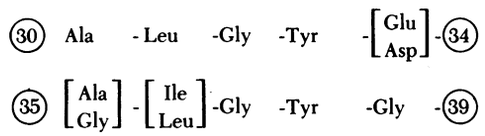
(5) Gly - Leu -[—, Gly]-[Ala, Tyr, Leu]-[—, Lys]-(9)

(10) [—, Ser, Gly]-[—, Leu]-Ala -[Tyr, Ala] -Pro -(14)

(15) Ala - Cys -Gly -Cys -Gly -(19)

(20) Gly - Leu -Gly -Tyr -[Gly, Glu]-(24)

(25) [Ala, Gly] - Leu -Gly -[Leu, Tyr] -Gly -(29)



Thus, the Gly-Leu-Gly-Tyr repeat may be part of a slightly longer, more variable segment. As the sequences are written, the last four repeats of the pentapeptide segment are seen to be highly conservative, the first two are variable but probably homologous, and the third may represent a distinct inserted sequence. Whatever the mode of origin of these tandem segments, they must figure prominently in the specification of the three-dimensional structure and hence in the function of the proteins.

Multigene families

The occurrence of multiple, closely linked identical or related genes within a single species is a common phenomenon. Examples include the genes for rRNA (15), tRNA (16), histones (17), antibodies (1), and β-like hemoglobins (18). Other potential examples include keratins such as wool (19), bovine dental enamel proteins (20), and actins (21). Unlike a single gene, a family of related genes can provide either very high rates of synthesis of a single product (such as rRNA or histones) or products with slight variations in their structural information (such as antibodies). When evolutionarily related genes code for products with identical, similar, or overlapping functions and when they are closely linked on the chromosome, they are said to form a multigene family (1).

Multigene families appear to have certain characteristic properties. Natural selection appears to act on the family as a whole rather than on individual genes because of their identical or similar functions. New genes that appear through mutation are selected for or against after they have expanded—i.e., after their multiplicity within a single genome has increased. If the new gene is favorable, it becomes established; if it is not, it disappears, resulting in a contraction of the gene pool for a given population. This gene expansion and contraction is species-specific, and thus two closely related species may differ both quantitatively and qualitatively in their multigene families. An example of this phenomenon is shown by the kappa and lambda families of immunoglobulins in mouse and man (1).

Extensive sequencing of the kappa and lambda families of immunoglobulins has demonstrated the existence of subfamilies (22). Subfamilies share certain structural features (specific residues, insertions, deletions) not found in other members of the family. The differences are minor, however, compared with similarities they share with all members of the family.

Chorion proteins of class A form a potential multigene family. They are clearly homologous and functionally related. We do not yet know whether the genes are closely linked, although gene linkage studies in the silkmoth *Bombyx mori* indicate that many chorion genes are located on the same chromosome (M. R. Goldsmith, personal communication). Likewise, we do not yet know the functional importance of the multiplicity of sequences—i.e., whether the high number of distinct sequences is a consequence of the mechanism of chorion evolution or is dictated by a multiplicity of subtly different needs in the construction of the chorion.

Chorion proteins also appear to belong to subfamilies. The A1,1- -a2,3 proteins are clearly a subfamily (Fig. 3). A4- -c1, A4- -d1, and one component each of A2,3- -d5,6 and A3,5- -d9 are another. The remaining two sequences may belong to yet a third subfamily or to two different subfamilies.

The multigene family concept may also be useful for understanding the evolution of other classes of chorion proteins. Preliminary sequencing studies on class B proteins show them to be homologous to each other and quite different from the class A proteins described here (G. Rodakis, J. C. Regier, and F. C. Kafatos, unpublished data). Thus, it is probable that chorion genes are composed of more than one multigene family, just as are immunoglobulin genes.

We thank M. Koehler for patient help with the figures and M. Randell for expert secretarial assistance. This work was supported by grants from the National Science Foundation and the National Institutes of Health to F.C.K. and to L.H.

1. Hood, L., Campbell, J. H. & Elgin, S. C. R. (1975) *Annu. Rev. Genet.* **9**, 305–353.
2. Kafatos, F. C., Regier, J. C., Mazur, G. D., Nadel, M. R., Blau, H. M., Petri, W. H., Wyman, A. R., Gelinas, R. E., Moore, P. B., Paul, M., Efstratiadis, A., Vournakis, J. N., Goldsmith, M. R., Hunsley, J. R., Baker, B., Nardi, J. & Koehler, M. (1977) "The eggshell of insects: Differentiation-specific proteins and the control of their synthesis and accumulation during development," in *Biochemical Differentiation of Insect Glands*, Results and Problems in Cell Differentiation, ed. Beerman, W. (Springer-Verlag, Berlin), Vol. 8, pp. 45–146.
3. Paul, M. & Kafatos, F. C. (1975) *Dev. Biol.* **42**, 141–159.
4. Regier, J. C. (1975) Ph.D Dissertation, Harvard University, Cambridge, MA.
5. Sim, G.-K., Efstratiadis, A., Jones, W., Kafatos, F. C., Koehler, M., Kronenberg, H., Maniatis, T., Regier, J. C., Roberts, B. & Rosenthal, N. (1977) *Cold Spring Harbor Symp. Quant. Biol.* **42**, in press.
6. Efstratiadis, A., Kafatos, F. C. & Maniatis, T. (1977) *Cell* **10**, 571–585.
7. Regier, J. C., Kafatos, F. C., Kramer, K. J., Heinrikson, R. L. & Keim, P. S. (1978) *J. Biol. Chem.*, in press.
8. Brauer, A. W., Margolies, M. N. & Haberr, E. (1975) *Biochemistry* **4**, 3029–3035.
9. Terhorst, C., Parham, P., Mann, D. L. & Strominger, J. L. (1976) *Proc. Natl. Acad. Sci. USA* **73**, 910–914.
10. Hermodson, M. A., Ericsson, L. H., Titani, K., Neurath, H. & Walsh, K. A. (1972) *Biochemistry* **11**, 4493–4502.
11. Paul, M., Goldsmith, M. R., Hunsley, J. R. & Kafatos, F. C. (1972) *J. Cell Biol.* **55**, 653–680.
12. Kawasaki, H., Sato, H. & Suzuki, M. (1971) *Insect Biochem.* **1**, 130–148.
13. Florkin, M. (1960) *Unity and Diversity in Biochemistry* (Pergamon Press, New York), p. 333.
14. Dayhoff, M. O. (1972) *Atlas of Protein Sequence and Structure* (National Biomedical Research Foundation, Washington, DC), Vol. 5.
15. Brown, D. D. & Sugimoto, K. (1973) *Cold Spring Harbor Symp. Quant. Biol.* **38**, 501–505.
16. Clarkson, S. G. & Birnstiel, M. L. (1973) *Cold Spring Harbor Symp. Quant. Biol.* **38**, 451–459.
17. Weinberg, E. S., Birnstiel, M. L., Purdom, I. F. & Williamson, R. (1972) *Nature* **240**, 225–228.
18. Kabat, D. (1972) *Science* **175**, 134–140.
19. Swart, L. S. (1973) *Nature New Biol.* **243**, 27–29.
20. Eggert, F. M., Allen, G. A. & Burgess, R. C. (1973) *Biochem. J.* **131**, 471–484.
21. Bray, D. (1972) *Cold Spring Harbor Symp. Quant. Biol.* **37**, 567–571.
22. Hood, L. (1973) *Stadler Symp.* **5**, 73–142.