Vol. 95, No. 2, 1980 BIOCHEMICAL AND BIOPHYSICAL RESEARCH COMMUNICATIONS
July 31, 1980 Pages 864-871

A SURPRISING NEW PROTEIN SUPERFAMILY CONTAINING OVALBUMIN, ANTITHROMBIN-III, AND $ALPHA_1$-PROTEINASE INHIBITOR

Lois T. Hunt and Margaret O. Dayhoff
National Biomedical Research Foundation
Georgetown University Medical Center
3900 Reservoir Road, N.W.
Washington, D.C. 20007

Received June 20,1980

SUMMARY: A distant relationship between chicken ovalbumin and two human plasma protease inhibitors was revealed by computer analyses. We propose a new protein superfamily containing at least three families: ovalbumin (and probably gene X and gene Y proteins), antithrombin-III, and $alpha_1$-proteinase inhibitor. Although these families may have diverged from a common ancestor more than 500 million years ago, they may still share similarity in gene structure as well as in protein sequence.

The sequence of the mature messenger RNA, comprising eight exons, of chicken ovalbumin has now been determined, allowing the derivation of the protein sequence (1). In addition, the sequence of the 2368 nucleotides in a 2.4-kilobase fragment of the DNA, extending from within the first intron into the fifth, has recently been established (2). Other recent investigations of the DNA adjacent to the ovalbumin gene have revealed two apparently related genes (3,4).

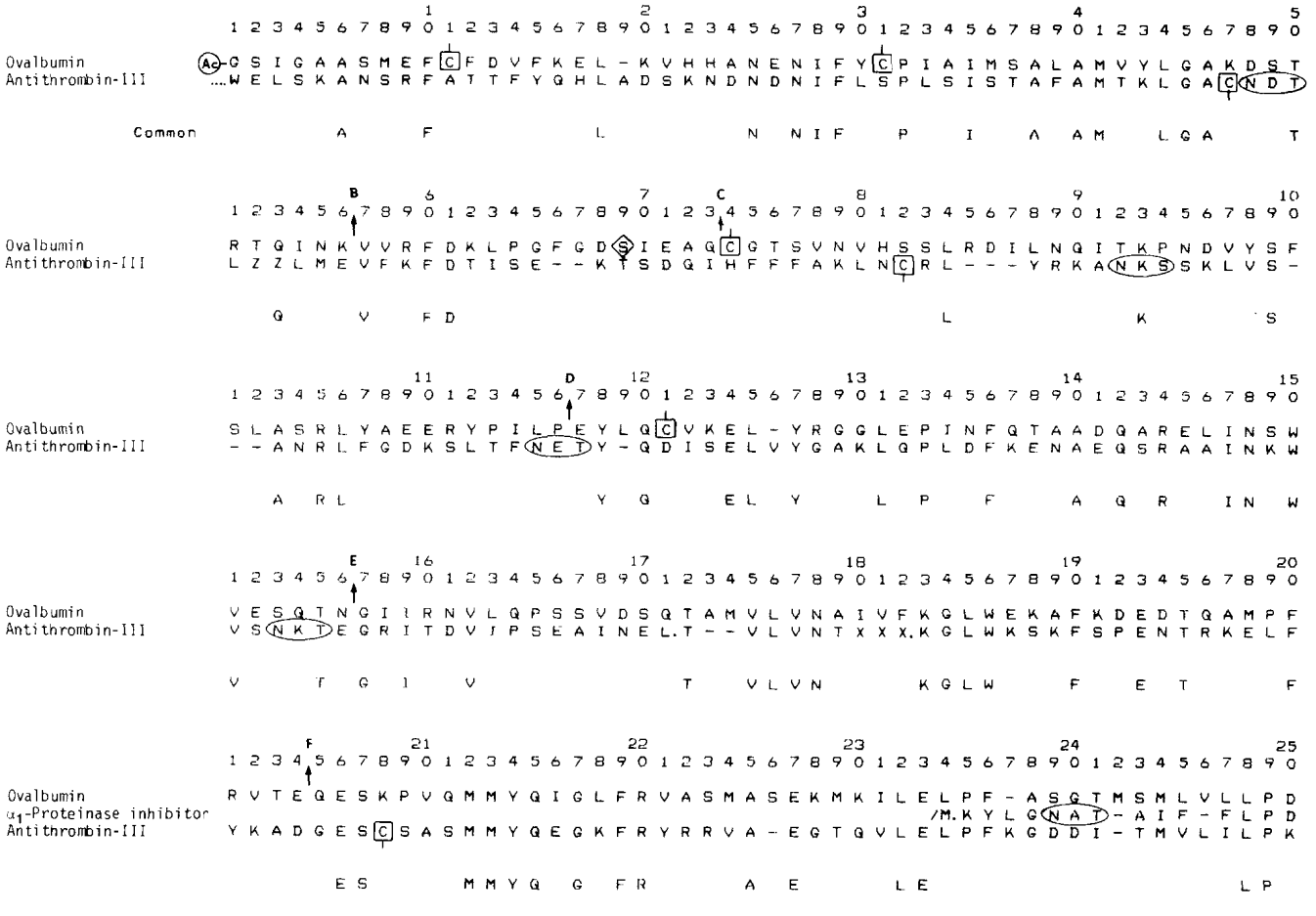
As part of our initial examination of a new protein sequence, we compared ovalbumin with all sequences in the Atlas Data Base by means of our computer program SEARCH (5). This analysis revealed an unexpected and definite relationship between the sequence of ovalbumin and that of human antithrombin-III, which is almost completely determined (6). The homology of the carboxyl-terminal 152 residues of human plasma $alpha_1$-proteinase inhibitor (7) to the carboxyl-terminal region of antithrombin-III has just been described (6,7). We find that chicken ovalbumin is clearly related to both proteins and propose that all three should be grouped in the same protein superfamily, which we provisionally name the ovalbumin-antithrombin superfamily.

0006-291X/80/140864-08$01.00/0

RESULTS OF COMPUTER ANALYSES

For these analyses we used three of our programs, SEARCH, ALIGN, and the alignment program DISP, which also generates a matrix of percent differences (5). In comparing a protein with the data base, we first search all consecutive 25-residue segments and examine the top scores. In the searches of the 16 segments from the ovalbumin sequence, a related region of antithrombin-III had the top score in 9 of the 16, and was among the top four scores in three more. In the searches of alpha$_1$-proteinase inhibitor, the top two scores in three of six searches were either ovalbumin or antithrombin-III.

Next we compared the three possible pairs of the sequences with program ALIGN, which calculates the best alignment between any pair of sequences, given a matrix of amino acid pair scores and a penalty for breaking a sequence (gap) (5). In this work we have used the Mutation Data Matrix with a bias of 6 and a gap penalty of 6. The maximum score that can be achieved by an alignment of a pair of real sequences is compared with the distribution of maximum scores for a large number (usually 100) of random permutations of the two sequences. The mean and standard deviation of this approximately normal distribution are calculated. The alignment score (AS) is the number of standard deviations by which the maximum score for the real sequences exceeds the average maximum score for the random permutations. The probability that a score as high as that from the real sequences could have been obtained in a comparison of randomized sequences can be determined from the cumulative standardized normal distribution table (5). The entire ovalbumin sequence (385 residues) compared with residues 49-423 of antithrombin-III gave an AS of 22.10 ($P < 10^{-100}$). Such a low probability is usually ascribed to a common evolutionary origin of the sequences. For comparison with the carboxyl-terminal 152-residue fragment of alpha$_1$-proteinase inhibitor, we used the corresponding regions (carboxyl-terminal 162 and 156 residues) of antithrombin-III and ovalbumin; the AS were 12.27 ($P < 10^{-34}$) and 13.49 ($P < 10^{-41}$), respectively.


Ovalbumin
Antithrombin-III
Common
Ovalbumin
Antithrombin-III
Ovalbumin
Antithrombin-III
Ovalbumin
Antithrombin-III
Ovalbumin
α1-Proteinase inhibitor
Antithrombin-III

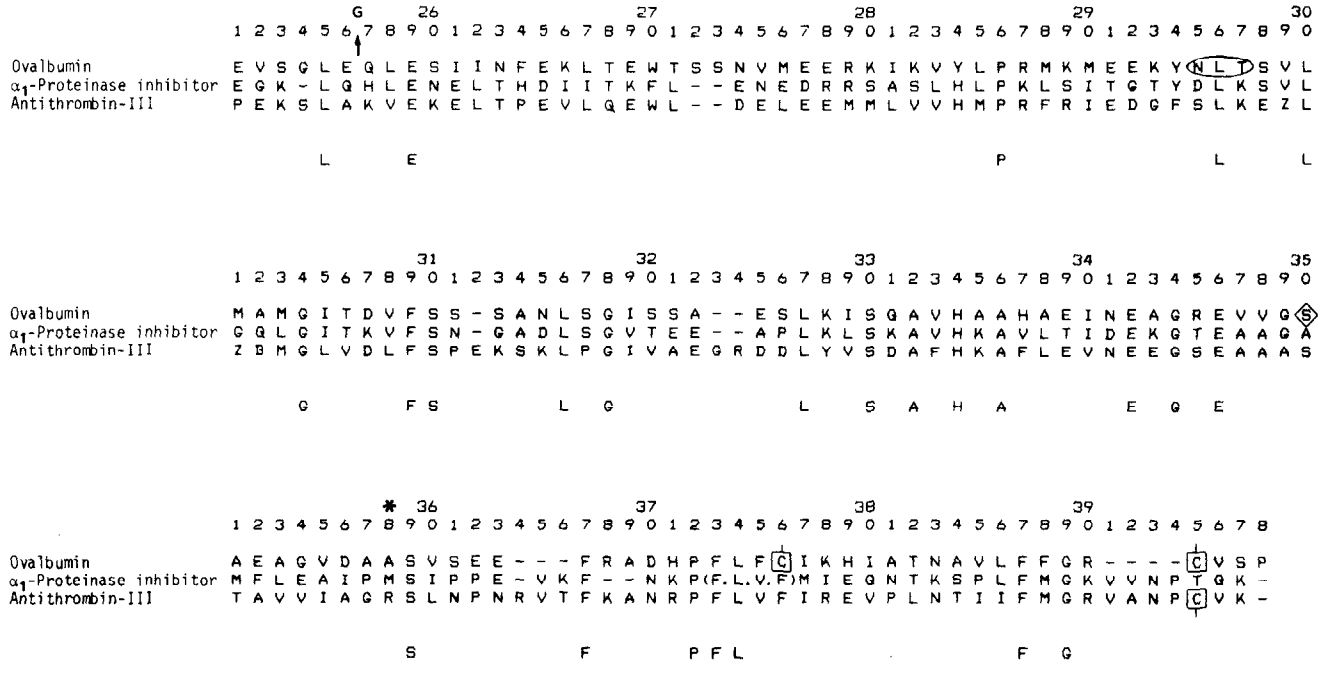


Figure 1. Alignment of chicken ovalbumin (entire sequence) with human alpha$_1$-proteinase inhibitor (carboxyl-terminal 153 residues) and human antithrombin-III (residues 49-423). Squares enclose cysteines, some of which may be involved in disulfide bonds; in antithrombin-III disulfide bonds link residues at positions 208-395 above and also cysteines at positions 47 and 82 above with those at sequence positions 21 and 8 (not shown), respectively. Diamonds enclose phosphorylated serines; ovals enclose Asn-X-Ser/Thr carbohydrate binding sites; and the asterisk indicates the position of the reactive site residues (Met and Arg). The points at which six of the seven introns (B-G) are spliced out of ovalbumin messenger RNA are indicated by the letters above the arrows. In ovalbumin, the residues at positions 236-256 constitute the proposed internal signal sequence.

The alignment of the three proteins, derived from the ALIGN matchings with gaps adjusted in a few places to improve the alignment, is shown in Figure 1. Based on this alignment, the three proteins are approximately 70% different from one another, placing them in different families (sequences within a family are <50% different) in the same superfamily (5). These values may change a little when the complete sequence (approximately 400 residues) of alpha$_1$-proteinase inhibitor can be compared.

Antithrombin-III is longer than ovalbumin by 48 residues at the amino end. This 48-residue segment does not significantly resemble, using ALIGN, any of the three sequences translated from the frameshifts of ovalbumin intron A-exon 2 (preceding Gly-1) (2). Nor did a search of the 48-residue segment reveal any significantly related sequences in the data base. A search of the amino-terminal 33-residue fragment of alpha$_1$-proteinase inhibitor (8), as well as comparisons, using ALIGN, with the three frameshift sequences, gave no significant results. A comparison with ALIGN of the two amino-terminal inhibitor sequences gave a negative score; thus the amino ends of these two proteins do not appear to be related.

DISCUSSION

The sequences of ovalbumin and the two plasma inhibitors are clearly related; indeed, the greatest similarity is between ovalbumin and antithrombin-III. However, the relationship among them is distant enough that, by our criteria, they should be placed in different protein families within one superfamily (5).

Although ovalbumin and the two inhibitors possess some of the same structural modifications, their positions do not usually correspond when the sequences are aligned. As is common among secreted proteins, all three are glycoproteins. Ovalbumin has one functional Asn-X-Ser/Thr carbohydrate binding site (1,9), antithrombin-III has four (6), and alpha$_1$-proteinase inhibitor has at least one (7). None of these sites are at corresponding positions in the alignment. In ovalbumin alone, two serines bind phosphate

and the amino end is acetylated after removal of the initiator methionine (1,10,11). Of the six cysteines in both ovalbumin and antithrombin-III, only the last in each sequence can be aligned to correspond (1,6). Antithrombin-III has three disulfide bonds (6), but ovalbumin has only one, whose location is not definitely established (1,12,13). Ovalbumin is synthesized from its messenger (14) without the amino-terminal hydrophobic signal peptide (11) found in the sequences of the other major egg white proteins. However, a 20-residue region, at alignment positions 236-256, apparently forms an internal signal sequence that functions in translocation across the endoplasmic reticulum membrane (15). No information about the signal sequences of the two inhibitors is available for comparison.

Antithrombin-III and alpha$_1$-proteinase inhibitor, together with the general endopeptidase inhibitor alpha$_2$-macroglobulin (most of whose sequence is not yet known), account for all of the normal plasma antithrombin activity (75-80%, <5%, and 20-25%, respectively) (16). However, alpha$_1$-proteinase inhibitor may primarily inhibit elastase activity, especially in the lungs (7). Antithrombin-III has one reactive site, Arg-Ser at alignment positions 358-359; the disulfide bond 208-395 joins the two fragments after thrombin cleavage (17). Alpha$_1$-proteinase inhibitor has a Met-Ser reactive site near the carboxyl end; it may also have a second site, nearly identical with the other over a distance of 25 residues, near the amino end (7). The reactive sites of the two inhibitors, at alignment positions 358-359, are homologous (7,17,18), and the short regions from position 356 to 363 are similar to the reactive site regions of other serine protease inhibitors (5,18). Positions 358-359 are Ala-Ser in the ovalbumin sequence; the region 356-363 also resembles reactive site regions of some serine protease inhibitors, for example, elastase inhibitors. A protease inhibitory function for ovalbumin has not yet been proposed, but considering the sequence similarities, it should be investigated. No other specific function for ovalbumin has been experimentally demonstrated, although several have been suggested. It appears

to be structurally unrelated to the other three major egg white proteins, ovomucoid (protease inhibitor), ovotransferrin (iron-binding, bacteriostatic), and lysozyme (bactericide).

At present we do not have enough information to calculate the rates of change among these proteins or to deduce an evolutionary tree distinguishing the protein and the species divergences. If we assume that the rates of change for all three proteins lie somewhere between 9.8 PAMs (accepted point mutations per 100 residues) and 18 PAMs per 100 million years, the rates for animal lysozyme and for pancreatic secretory trypsin inhibitor (5), then their divergence times could have been at least 500 million years ago, before or during early vertebrate evolution. Based on such an estimate, genes for all three proteins would have been present in the mammal-bird ancestral line and could still be present and expressed in birds; in mammals the ovalbumin gene may have been lost or be present but not expressed by any type of cell. The mammal-bird ancestral gene most likely coded for an inhibitory protein. If ovalbumin should also prove to be a protease inhibitor, then this function would have been conserved in all three resulting families of proteins.

In recent investigations of the DNA region in the 5' direction from the chicken ovalbumin gene (3,4), two more genes, named X and Y, were found that have the same exon-intron pattern and approximate sizes. Genes X and Y are about 8500 and 6500 nucleotides long and the mature messengers are about 2400 and 2000 nucleotides long (3,4). Regions of all three genes cross-hybridize, indicating varying degrees of structural similarity, and the structure of the gene X eighth exon fragment is fairly similar (57% identity) to the homologous region in ovalbumin (3,4). Most likely all three protein sequences will prove to be less than 50% different and, therefore, in the same family, analogous to the alpha-type and the beta-type hemoglobin families (5). We also consider it likely that, as a result of possessing a common ancestral gene, the genes for the two inhibitors will be found to have an exon-intron pattern similar to that of the ovalbumin family. Indeed, this newly proposed superfamily

resembles several others now known, such as those of the globins, several hormones, and several protease inhibitors, that demonstrate distant duplication events preceding most of vertebrate evolution (5).

ACKNOWLEDGMENTS: This research was supported by NIH grant GM-08710 from the National Institute of General Medical Sciences.

REFERENCES

1. McReynolds, L., O'Malley, B.W., Nisbet, A.D., Fothergill, J.E., Givol, D., Fields, S., Robertson, M., and Brownlee, G.G., Nature 273: 723-728, 1978
2. Robertson, M.A., Staden, R., Tanaka, Y., Catterall, J.F., O'Malley, B.W., and Brownlee, G.G., Nature 278: 370-372, 1979
3. Royal, A., Garapin, A., Cami, B., Perrin, F., Mandel, J.L., LeMeur, M., Brégégègre, F., Gannon, F., LePennec, J.P., Chambon, P., and Kourilsky, P., Nature 279: 125-132, 1979
4. Chambon, P., Perrin, F., O'Hare, K., Mandel, J.L., LePennec, J.P., LeMeur, M., Krust, A., Heilig, R., Gerlinger, P., Gannon, F., Cochet, M., Breathnach, R., and Benoist, C., in Eucaryotic Gene Regulation, Axel, R., Maniatis, T., and Fox, C.F., eds., pp.259-278, Academic Press, New York, 1979
5. Dayhoff, M.O., ed., Atlas of Protein Sequence and Structure, vol.5, suppl. 3, National Biomedical Research Foundation, Washington, D.C., 1979
6. Petersen, T.E., Dudek-Wojciechowska, G., Sottrup-Jensen, L., and Magnusson, S., in The Physiological Inhibitors of Blood Coagulation and Fibrinolysis, Collen, D., Wiman, B., and Verstraete, M., eds., pp.43-54, Elsevier/North-Holland Biomedical Press, Amsterdam, 1979
7. Carrell, R., Owen, M., Brennan, S., and Vaughan, L., Biochem. Biophys. Res. Commun. 91: 1032-1037, 1979
8. Morii, M., Odani, S., Koide, T., and Ikenaka, T., J. Biochem. 83: 269-277, 1978
9. Lee, Y.C., and Montgomery, R., Arch. Biochem. Biophys. 97: 9-17, 1962
10. Thompson, E.O.P., and Fisher, W.K., Aust. J. Biol. Sci. 31: 443-446, 1978
11. Palmiter, R.D., Gagnon, J., and Walsh, K.A., Proc. Nat. Acad. Sci. USA 75: 94-98, 1978
12. Thompson, E.O.P., and Fisher, W.K., Aust. J. Biol. Sci. 31: 433-442, 1978
13. Fothergill, L.A., and Fothergill, J.E., Biochem. J. 116: 555-561, 1970
14. Gagnon, J., Palmiter, R.D., and Walsh, K.A., J. Biol. Chem. 253: 7464-7468, 1978
15. Lingappa, V.R., Lingappa, J.R., and Blobel, G., Nature 281: 117-121, 1979
16. Davie, E.W., and Hanahan, D.J., in The Plasma Proteins, 2nd ed., vol.3, Putnam, F.W., ed., pp.421-544, Academic Press, New York, 1977
17. Jörnvall, H., Fish, W.W., and Björk, I., FEBS Lett. 106: 358-362, 1979
18. Carrell, R.W., Boswell, D.R., Brennan, S.O., and Owen, M.C., Biochem. Biophys. Res. Commun. 93: 399-402, 1980