EPIDERMAL GROWTH FACTOR: INTERNAL DUPLICATION AND PROBABLE RELATIONSHIP TO PANCREATIC SECRETORY TRYPSIN INHIBITOR

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

Received August 19,1974

SUMMARY: The sequence of the first half of epidermal growth factor is related to that of the second half, indicating a gene elongation through duplication; it is also similar to that of pancreatic secretory trypsin inhibitor, which suggests a common evolutionary origin for these two proteins. Statistical evaluation shows that it is very unlikely that either similarity in sequence could have occurred by chance. The two proteins may also be functionally related, as each forms a complex with an arginine esterase.

INTRODUCTION

Epidermal growth factor (EGF) is synthesized by the tubular cells of the submaxillary glands of adult male mice and is secreted into the blood stream. The secretion has been stimulated experimentally by administration of α-adrenergic compounds [1]. EGF stimulates growth and differentiation of various epithelial tissues *in vivo* and *in vitro*. Although its initial point of action is still unknown, EGF has been found to increase RNA, protein, and ribosome synthesis, and also to induce a transient increase in ornithine decarboxylase activity in cultured epidermal cells [1, 2]. The primary structure of mouse EGF has recently been determined [3]; it is a single polypeptide chain comprising 53 amino acid residues and has three disulfide bonds that are required for full biological activity [4]. No other modifications of the primary structure have been reported.

Pancreatic secretory trypsin inhibitor (PSTI) is found in mammalian pancreatic juice. The primary structures of bovine [5] and pig [6] PSTI have been reported; each contains 56 amino acid residues and has three disulfide bonds [7]. In the two sequences, 88% of the residues are identical [8]. The evolutionary rate of change of PSTI is slower than that of hemoglobin, indicating that proteins homologous to it found in any of the eukaryote kingdoms may be recognizable by statistical methods [8].

We investigated two genetic events in the evolutionary history of EGF. Intragenic duplication is the probable cause for the similarity between the two halves of its sequence. The statistical inference that EGF is much more similar in sequence to PSTI than one would expect by chance, supported by its physiological binding to an arginine esterase, indicates that these molecules may share a common evolutionary origin. If so, one would expect other similarities in chemical function and in biological expression and control that could be investigated experimentally.

METHODS

We study a new sequence by a variety of survey and statistical procedures to determine if it is related to any other sequenced proteins or if it has internal duplicated regions. Possible sequence similarities are inferred from chemical and biological properties, analysis of amino acid composition [9], and comparison by visual inspection of short regions of unusual sequence or of functional importance with the segments in the Protein Segment Dictionary [10] (a computerized listing of all 15-residue segments in the known sequence data [8, 11]). Next, we use our computer Search Program [12] to compare selected longer portions of the new sequence with all sequences on our most recently updated Protein Sequence Data Tape, which currently contains 725 sequences belonging to 176 protein families (sequences within a family are more than 50% identical and usually give similar scores in our comparisons). Possible relationships that are revealed by these procedures are statistically analyzed by means of our Alignment Score Program [13] or our Fragment Comparison Program [14] described below.

For these programs we need a matrix of scores for every possible comparison of a residue in one sequence with one in the other sequence. The simplest such matrix, the unitary matrix, gives a score of one for identities and zero for nonidentities. We have derived a scoring matrix from mutation data that has proved to be much more sensitive for detecting distant relationships between proteins [13, 14]. The elements of this matrix are the logarithms of the odds that two amino acids, found at corresponding positions in two proteins with average composition and about 20% identical residues, arose by common ancestry rather than by chance [13]. Therefore, the sum of scores of residue pairs in two segments gives the log of the odds that the segments are related by evolution. A probability that the observed similarity is a chance occurrence can be directly calculated from the odds.

The Alignment Score Program, based on the algorithm of Needleman and Wunsch [15], determines the highest possible score for any alignment (including gaps) of two sequences. This score is then compared with the highest possible scores obtained by aligning pairs of randomized sequences having the same amino acid composition as the two real sequences. The scores of the randomized sequences form a normal distribution for which the mean and standard deviation are calculated. The probability that the score for the real sequences is derived from this normal distribution can then be obtained [13].

The Fragment Comparison Program, similar to one developed independently by Fitch [16], compares all fragments of a given length from one sequence with all possible fragments of the same length from the second sequence. A score is calculated for each comparison. The score consists of the sum of scores for each individual pair of amino acids occupying corresponding positions within the two fragments. If the proteins are related, there will be two populations of scores: a large population from unrelated fragments and a very small population from related fragments. The scores from the latter will be higher than those from the unrelated fragments and may be detectable as a hump or an elongation of the upper tail of the distribution.

We calculate the expected number of scores in this second population and store that number of top scores from the upper tail of the distribution. The average of these top scores is compared with similar averages derived from comparisons in which one or both of the two sequences have been randomized. The mean and standard deviation are calculated for the comparisons of randomized sequences. The probability that the score from real sequences is part of this distribution is then obtained.

EVIDENCE FOR INTERNAL DUPLICATION

A visual inspection of the EGF sequence indicated a possible duplication, as there are several regions of similarity between the two halves of the chain. The amino-terminal 26 residues and the carboxyl-terminal 27 residues can be aligned, with the insertion of three gaps, to give seven identities (including two cysteines and two tyrosines in each half-chain) and 13 additional positions at which the codons for the residues differ by only one nucleotide (see Figure 1). The relationship of the two halves was evaluated statistically with the Alignment Score Program; the score obtained is 2.9 standard deviations from the mean, and the probability is <0.002 that the two halves would be so similar in sequence by chance.

Intragenic duplication is a recognized evolutionary mechanism. A number of proteins containing duplications have been sequenced [11, 17]; among these are two types of protease inhibitors, the Bowman-Birk type from lima bean [18] and soybean [19], and the chymotrypsin inhibitor from potato [20]. The internal duplication of EGF should be detectable in its tertiary structure, as, for example, are the duplications in ferredoxin and parvalbumin [21, 22, 23].

EVIDENCE FOR RELATEDNESS OF EGF TO PSTI

Those protein segments that the Search Program selected as most similar to each half-chain of EGF are presented in Table 1. The primary structure of

TABLE 1. Segments of protein sequences selected by the Search Program as most similar to the halves of the EGF sequence

EGF Residue Nos.		Residue Nos.	No. of Identities	Mutation Data Matrix Score (Log Odds)	Probability
1-26	EGF Mouse	1-26	26	14.7	2.0×10^{-15}
1-26	PSTI Bovine	11-36	10	4.9	1.3×10^{-5}
1-26	PSTI Pig	11-36	7	4.6	2.5×10^{-5}
1-26	Ig κ chain Rabbit 2717	86-111	6	3.9	1.3×10^{-4}
1-26	Ig κ chain Mouse MOPC21	53-78	7	3.3	5.0×10^{-4}
1-26	7 Segments		4-7	2.5-3.1	
1-26	25 Segments		1-8	2.0-2.4	
27-53	EGF Mouse	27-53	27	18.3	5.0×10^{-19}
27-53	Phospholipase A Honey Bee	101-127	6	4.9	1.3×10^{-5}
27-53	17 Segments		2-8	2.0-3.2	

the first half of EGF is very similar to the middle portion (residues 11-36) of the sequences of bovine and pig PSTI; not only is it 38.5% identical with that region of the bovine sequence, but it also shares more identities with bovine PSTI than with any other segment in the data collection. Furthermore, when the bovine PSTI segment is searched, the two most similar segments are the corresponding pig PSTI segment and the first half of EGF. The segments of the two immunoglobulin chains listed in Table 1 are not homologous to each other and probably lie in the upper tail of the distribution of scores from unrelated sequences. The highest-scoring segment for the second half of EGF is the carboxyl-terminal region of honey bee phospholipase A, but a closer examination reveals that this score is derived almost entirely from the last five residues and does not reflect similarity over the whole segment. The relationships of the second half of EGF to its first half and to PSTI are not found by the program, because three gaps are required for proper alignment (see Figure 1).

```
                                 1                   2                   3                   4                   5
               1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8

PSTI (PIG)     T S P Q R E A T C T S E V S G - C P K I Y N P V - C G T D G I T Y S N E C V L C S E N K K R Q T P V L I Q K S G P C
PSTI (BOVINE)  N I L G R E A K C T N E V N G - C P R I Y N P V - C G T D G V T Y S N E C L L C M E N K E R Q T P V L I Q K S G P C
EGF  (1-26)    - - - - - - - - - - N S Y P G - C P S S Y D G Y - C L N G G V C M H I E S L
EGF  (27-53)                       D S Y T C N C V I G Y S G D R C Q T R D L R W W - E L R - - - - - - - - - - - - - - - - - - - -

COMMON                                    (G)  C(P)    Y         C  (T) (G)          E
```

Figure 1. Alignment of pig PSTI, bovine PSTI, and the halves of mouse EGF (residues 1-26 and 27-53). Residues common to three (in parentheses) or to all sequences are indicated below the alignment. Disulfide bonds link residues in the two PSTI sequences at alignment positions 9-40, 17-37, and 26-58, and in EGF at alignment positions 17-32 (first half), 26 (first half)-15 (second half), and 17-26 (second half). The Arg or Lys at position 19 in PSTI interacts with trypsin.

When we analyzed the complete sequences of EGF and bovine PSTI with the Fragment Comparison Program, we obtained a low probability (<0.0001) that the similarity between the two sequences is a chance one. On the other hand, similar comparisons of EGF with honey bee phospholipase A and with rabbit 2717 immunoglobulin kappa chain gave much higher probabilities (<0.08 and <0.05, respectively) of the sequence similarities being due to chance.

EGF and PSTI have other overall resemblances. They have similar amino acid composition and length. Each has three disulfide bonds, although the pairs of cysteines involved are different (see Figure 1); any original correspondences may have been altered by subsequent additions, deletions, duplications, or changes in hydrolysis of a precursor. Shifting of disulfide bonds between cysteines that are distant in the primary structure but close together in the tertiary structure is found in the immunoglobulins [24].

These two proteins may also share a functional relationship. PSTI binds to trypsin and is reversibly cleaved by it at the Arg or Lys residue at position 19 in Figure 1 [25]. However, the enzyme-product complex thus formed is stable and trypsin is inhibited from further proteolytic activity. EGF occurs in a readily reversible complex with a glycoprotein, EGF-binding protein, in homogenates of the submaxillary gland; as isolated, the complex is composed of two molecules of each protein [26, 27]. Like trypsin, EGF-binding protein is an arginine esterase [26, 27]. Its composition is similar, in proportions of the different chemical groups of the amino acids, to that of trypsinogen, including 12 cysteines that form six disulfide bonds [26]. It has been postulated that EGF-binding protein catalyzes the conversion of inactive to active EGF by proteolytic cleavage at the carboxyl-terminal Arg in EGF; that is, the substrate of the esterase is the inactive precursor of EGF and the complex is actually an enzyme-product complex [27]. However, the sequence preceding the carboxyl-terminal Arg is not noticeably similar to that of the PSTI reactive site. Furthermore, in the alignment EGF has a Ser or an Ile corresponding to the Arg/Lys of the PSTI reactive site. It is possible that one of these regions of EGF binds near the arginine esterase active site, although no subsequent hydrolysis of EGF occurs. In any case, PSTI and EGF seem to be an example of two proteins with a common ancestor that have evolved sequence differences leading to a marked alteration in functional capabilities; such a relationship exists between mammalian lysozyme and lactalbumin [8] and between troponin C and myosin light chains [28].

There are several experimental approaches for investigating the postulated relationships between EGF and PSTI. Both EGF and its precursor could be tested for protease inhibitor activity. Further sequence analysis of the EGF precursor might reveal a more extensive region of homology with PSTI, including additional internal duplications. Finally, X-ray crystallographic analysis might substantiate both the internal duplication of EGF and the similarity of its tertiary structure to that of PSTI.

ACKNOWLEDGMENTS: This investigation was supported by NIH Grant GM-08710 from the National Institute of General Medical Sciences and by Contract NASW2546 from the National Aeronautics and Space Administration.

REFERENCES

1. Cohen, S. (1972) *J. Invest. Dermatol.* 59, 13-16.
2. Taylor, J.M., Mitchell, W.M. and Cohen, S. (1972) *J. Biol. Chem.* 247, 5928-5934.
3. Savage, C.R. Jr., Inagami, T. and Cohen, S. (1972) *J. Biol. Chem.* 247, 7612-7621.
4. Savage, C.R. Jr., Hash, J.H. and Cohen, S. (1973) *J. Biol. Chem.* 248, 7669-7672.
5. Greene, L.J. and Bartelt, D.C. (1969) *J. Biol. Chem.* 244, 2646-2657.
6. Tschesche, H., Wachter, E., Kupfer, S. and Niedermeier, K. (1969) *Hoppe-Seyler's Z. Physiol. Chem.* 350, 1247-1256.
7. Guy, O., Shapanka, R. and Greene, L.J. (1971) *J. Biol. Chem.* 246, 7740-7747.
8. Dayhoff, M.O., ed. (1972) *Atlas of Protein Sequence and Structure 1972,* 5, Nat. Biomed. Res. Found., Washington, D.C.
9. Dayhoff, M.O. and Hunt, L.T. (1972) in *Atlas of Protein Sequence and Structure 1972,* 5, ed. Dayhoff, M.O., D-355-D-359, Nat. Biomed. Res. Found., Washington, D.C.
10. Dayhoff, M.O., Hunt, L.T., Barker, W.C., McLaughlin, P.J. and Shiu, M.R. (1973) *Protein Segment Dictionary 73,* Nat. Biomed. Res. Found., Washington, D.C.
11. Dayhoff, M.O., ed. (1973) *Atlas of Protein Sequence and Structure 1973,* 5, Suppl. 1, Nat. Biomed. Res. Found., Washington, D.C.

12. Hunt, L.T. and Dayhoff, M.O. (1972) *Biochem. Biophys. Res. Commun.* 47, 699-704.
13. Barker, W.C. and Dayhoff, M.O. (1972) in *Atlas of Protein Sequence and Structure 1972,* 5, ed. Dayhoff, M.O., 101-110, Nat. Biomed. Res. Found., Washington, D.C.
14. Barker, W.C. and Dayhoff, M.O. (1970) *Biophys. Soc. Abstracts* 10, 152a.
15. Needleman, S.B. and Wunsch, C.D. (1970) *J. Mol. Biol.* 48, 443-453.
16. Fitch, W.M. (1966) *J. Mol. Biol.* 16, 9-16.
17. Dayhoff, M.O. and Barker, W.C. (1972) in *Atlas of Protein Sequence and Structure 1972,* 5, ed. Dayhoff, M.O., 41-45, Nat. Biomed. Res. Found., Washington, D.C.
18. Tan, C.G.L. and Stevens, F.C. (1971) *Eur. J. Biochem.* 18, 515-523.
19. Odani, S. and Ikenaka, T. (1972) *J. Biochem.* 71, 839-848.
20. Richardson, M. (1974) *Biochem. J.* 137, 101-112.
21. Eck, R.V. and Dayhoff, M.O. (1966) *Science* 152, 363-366.
22. Matsubara, M., Jukes, T.H. and Cantor, C.R. (1969) *Brookhaven Symp. Biol.* 21(I), 201-216.
23. Kretsinger, R.H. (1972) *Nature New Biol.* 240, 85-88.
24. Poljak, R.J., Amzel, L.M., Avey, H.P., Chen, B.L., Phizackerley, R.P. and Saul, F. (1973) *Proc. Nat. Acad. Sci. USA* 70, 3305-3310.
25. Laskowski, M. Jr. (1970) in *Structure-Function Relationships of Proteolytic Enzymes,* ed. Desnuelle, P., Neurath, H. and Ottesen, M., 89-101, Munksgaard, Copenhagen.
26. Taylor, J.M., Mitchell, W.M. and Cohen, S. (1974) *J. Biol. Chem.* 249, 2188-2194.
27. Taylor, J.M., Mitchell, W.M. and Cohen, S. (1974) *J. Biol. Chem.* 249, 3198-3203.
28. Collins, J.H. (1974) *Biochem. Biophys. Res. Commun.* 58, 301-308.