J Mol Evol (1995) 40:102–106

JOURNAL OF MOLECULAR EVOLUTION

Active Sites of Ligands and Their Receptors Are Made of Common Peptides That Are Also Found Elsewhere

Susumu Ohno

Department of Theoretical Biology, Beckman Research Institute of the City of Hope, 1450 E. Duarte Road, Duarte, CA 91010-0269, USA

Received: 4 October 1994

Abstract. The simultaneous emergence in evolution of a ligand and its receptor might have entailed their active sites being drawn from the pool of common oligopeptides. This was tested on the principal components of cell-matrix interaction: the RGD (Arg-Gly-Asp) site of matrix proteins and the EKKD (Gly-Lys-Lys-Asp) site of integrin cell-surface receptor. In the 32 diverse proteins scrutinized, which totalled 14,806 residues, there were 104 Arg-Gly dipeptides. Most common of the tripeptides beginning with Arg-Gly were Arg-Gly-Leu, Arg-Gly-Gly, and Arg-Gly-Asp; each was found in ten copies. RGD tripeptide was one of the commonest; the fortuitous presence of an RGD site was noted in two enzymes, fibrinogen, a pituitary hormone precursor, and a viral structural protein. The 32 proteins also contained 121 Lys-Lys dipeptides. Of the tetrapeptides centered by Lys-Lys, the commonest was Lys-Lys-Lys-Lys, in four copies. Second most common were Gly-Lys-Lys-Lys, Val-Lys-Lys-Leu, and Glu-Lys-Lys-Asp; each occurred in three copies. The fortuitous presence of an EKKD site was noted in three proteins—an intracellular transport protein, a pituitary hormone precursor and a protein of the cerebrospinal fluid. In most instances, protein–protein interaction between the fortuitously present active sites appears to bring about deleterious consequences. Occasionally, however, the fortuitous active site appears to confer a new function to a protein bearing it.

Key words: Active site — common oligopeptides — EKKD site — Epitheliomesenchymal interaction — Fibronectins — Integrins — RGD site

Introduction

One major unsolved problem in evolution is the mechanism by which a peptide ligand and its specific protein receptor came into being. Barring the simultaneous emergence of both components, one would drift away to become a senseless entity before the emergence of the other. It has been suggested that the translation of a complementary strand of the receptor coding sequence shall yield a peptide ligand of that receptor (Bost et al. 1985). This ingenious suggestion, however, leaves a number of problems unanswered. *A priori,* a complementary strand should be devoid of the transcription initiation signal, the appropriately placed initiation codon, the poly-A attachment signal, and other paraphernalia of transcription and translation. Furthermore, a serine codon, TCA, and two leucine codons, CTA and TTA, in coding strand become chain terminators in a complementary strand.

What if the active sites of both a ligand and its receptor are made of common oligopeptides that are invariably present in a number of totally unrelated proteins? The simultaneous presence of a ligand and its receptor is assured by their very commonness. The notion of common peptides evolved out of my previous observations that identical tetrapeptides, pentapeptides, and hexapeptides are far too often shared by a pair or more of unrelated proteins than expected by chance (Ohno 1991, 1992a,b).

RGD Ligand Site of Matrix Proteins and EKKD Receptor Site of Cell-Surface Integrins as the Targets of Choice

In view of the importance of epitheliomesenchymal interaction in organogenesis, the above possibility was

tested using an RGD (Arg-Gly-Asp) site of fibronectin and on other extracellular matrix proteins as ligands (Pierschbacher and Ruoslahti 1984; Rouslahti and Pierschbacher 1987) and using EKKD (Glu-Lys-Lys-Asp) sites of cell-surface integrins as their receptors (Argravos et al. 1987). The decisive role mesenchyme plays in organogenesis has been well established by numerous experiments. For example, judging from the rudimentary nature of their teeth in Jurassic *Archeopteryx lithographia* as well as in Cretaceous *Monychus olecranus* (Altangerel et al. 1993), the tooth formation was forfeited by members of the class *Aves* from the inception. Yet, the molar tooth was formed by chick embryos following the transplantation of mouse embryonic tooth mesenchyme (Kollar and Fisher 1980).

The first attempt to understand the interaction between matrix proteins and cells as ligand–receptor binding was made in fibronectin and its cell-surface receptor. The active site on fibronectin was first thought to be tetrapeptide RGDS (Pierschbacher and Ruoslahti 1984). However, it was later shown that RGD tripeptide suffices as *conditio sine qua non* (Ruoslahti and Pierschbacher 1987). The cell-surface receptor for fibronectin, on which the active-site EKKD was first identified (Argravos et al. 1987), was later shown to be a member of the integrin family.

Oligopeptide Analysis Performed on 32 Proteins with a Combined Total of 14,806 Residues

In order to test whether RGD as well as EKKD belongs to the category of common peptides or rare ones, 32 proteins totally unrelated to each other, to the best of my knowledge, were assembled. Deliberately excluded from this group were fibronectin and other matrix proteins as well as members of the integrin family. However, mouse α1-chain of type IV collagen, 1,669 residues long, was included (Muthukumaran et al. 1989). The most important of the cell–matrix protein interactions is that between epithelial cells and basement membrane. Inasmuch as type IV is a collagen exclusively utilized for the construction of basement membrane, the presence of RGD sites on type IV collagen might be of more pertinence than that of RGD on fibronectin. Hence, the inclusion of mouse α1-chain of type IV collagen.

The sources of these 32 proteins were viral, bacterial, fungal, echinodermal, piscine, avian, and mammalian; two proteins of plant origin were also included. Nevertheless, 12 of them were of mammalian origin. The overall amino acid composition of the 32 proteins was very similar to the average deduced from 18,383 entries in DATAbase in which leucine, alanine, glycine, and serine, in that order, were dominant residues, together comprising 36.1% of the total (Seto 1989). Of the 32, as many as eight proteins contained either RGD and/or EKKD. These eight are identified in Figs. 1 and 2, including their reference sources.

RGD as the Commonest of the Related Tripeptides

One each of arginine, glycine, and aspartate yield six different tripeptides. Were the protein to represent a random assemblage of amino acid residues, all six of these tripeptides should occur at around the expected frequency of 6.04. As shown at the top of Fig. 1, such an expectation was rudely violated by the once-only appearance of Asp-Gly-Arg tripeptide at one end, while at the other end, the appearance in ten copies of Arg-Gly-Asp constituted the greatest excess. Of the six tripeptides, two contained Arg-Gly and both were in excess, while the other two that contained its reciprocal, Gly-Arg, were markedly less prominent. Indeed, Fig. 1 also shows that Arg-Gly dipeptide occurred 1.4 times more frequently than its reciprocal—Gly-Arg.

Of 20 different kinds of tripeptides starting with Arg-Gly, only Arg-Gly-Met was in a single copy, whereas three, Arg-Gly-Gly, Arg-Gly-Leu, and Arg-Gly-Asp, were in ten copies each. The above leaves little doubt that Arg-Gly-Asp is one of the commonest, whereas Asp-Gly-Arg is one of the rarest among the related tripeptides. Four of the ten Arg-Gly-Asp tripeptides were on mouse collagen IV α1-chain (Muthukumaran et al. 1989); three copies appeared as parts of triplets of the identical pentapeptide—Arg-Gly-Asp-Pro-Gly. No doubt either all or some of the four RGD sites residing in mouse collagen IV α1-chain contributed to the formation of the basement membrane adjacent to epithelial cells. The remaining six RGD sites were scattered over five proteins. Since none of the five were matrix proteins, their presence must be considered fortuitous. Yet, even here, the RGD site tended to be a part of the longer common oligopeptide. For example, the identical hexapeptide, Gly-Arg-Gly-Asp-Ser-Gly, was shared between Sindbis virus structural protein (Rice and Strauss 1981) and α-lytic protease of Myxobacter 495 (Olson et al. 1978).

Were DGR instead of RGD chosen as the active site of matrix proteins, the fortuitous presence of this active site in irrelevant proteins would surely have been avoided. Yet, being so rare, the presence in more than one copy of the active site on relevant collagen IV α1-chain would have also become very unlikely.

EKKD Receptor Active Site Also Is One of the Commonest of the Tetrapeptides Containing Lys-Lys in the Center

The cell-surface receptor of fibronectin first identified as such (Argravos et al. 1979) turned out to be a member of the integrin family; other members of the family also

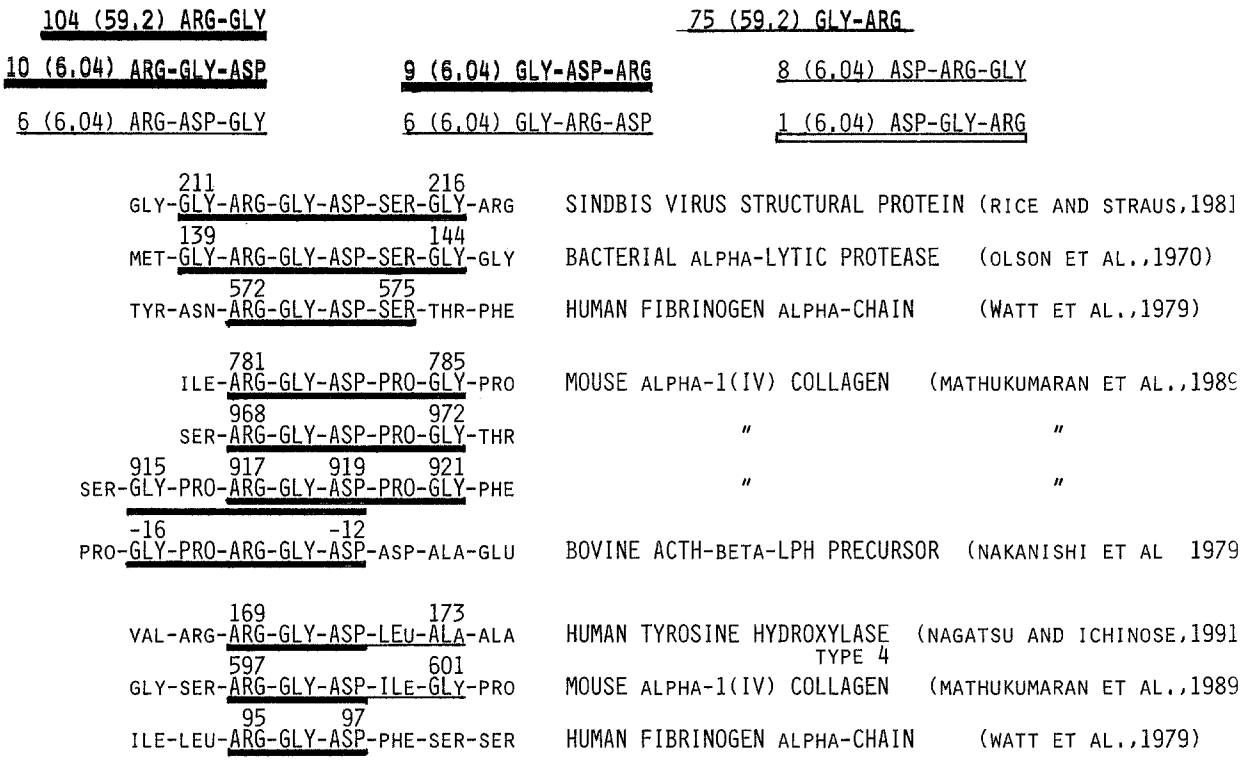

Fig. 1. At the top, contrasting incidences of Arg-gly vs Gly-Arg dipeptides are shown. Observed numbers are followed by expected numbers in parentheses. Shown immediately below are observed and expected incidences of six tripeptides containing one each of arginine, glycine, and aspartate. Shown next are ten RGD (Arg-Gly-Asp) sites found in six proteins. Mouse αl (IV) collagen contained four sites and human fibrinogen α-chain two. All except one of the ten RGD sites were parts of longer repeating units. For example, the RGD site of Sindbis virus structural protein as well as bacterial α-lytic protease was a part of the identical hexapeptide shared by the two proteins. Similarly, three of the four RGD sites within mouse αl (IV) collagen were parts of identical pentapeptides. When observed frequencies of di- and tripeptides significantly exceeded the expected frequencies deduced from the overall amino acid composition, they are *underlined by solid bars.* Conversely, Asp-Gly-Arg tripeptide that made only one appearance was *underlined by a thick blank bar.* Arg-Gly-Asp tripeptide and longer repeating units containing this tripeptide are also *thickly underlined by solid bars.*

carried out this function (Rusolahti and Pierschbacher 1987). Nevertheless, the receptor active site has been identified as an EKKD (Glu-Lys-Lys-Asp) site. As shown near the top of Fig. 2, both Lys-Lys and Lys-Lys-Lys homodi- and homotripeptides occurred twice as frequently as expected. Since oligopeptide formation follows the principle of like attracting like (Ohno 1992b), all homo-oligopeptides occur in greater frequency than expected. Figure 2 shows five pairs of Lys-Lys-containing tripeptides; the identical residue placed either in the front (left column) or the rear (right column) of Lys-Lys constitutes each pair of tripeptides. With regard to every pair, observed frequencies of members are very divergent—e.g., 11 × Glu-Lys-Lys vs 7 × Lys-Lys-Glu as well as 6 × Asp-Lys-Lys vs 10 × Lys-Lys-Asp. The above makes it again clear that even among oligopeptides made of the same set of amino acid residues, there are common ones and rare ones. In view of the abundance of Glu-Lys-Lys as well as Lys-Lys-Asp, it was rather expected that the EKKD (Glu-Lys-Lys-Asp) site would rank among the most common of the tetrapeptides centered by Lys-Lys. Figure 2 shows that this was indeed the case as the EKKD site occurred in three proteins. Inasmuch as glutamate and aspartate are acidic residues, Glu-Lys-Lys-Asp is a pseudopalindrome, and palindromic oligopeptides tend to occur in greater frequency than expected (Ohno 1992b). Shown together with 3 × Glu-Lys-Lys-Asp in Fig. 2 are observed incidences of three other related palindromic and pseudopalindromic tetrapeptides: Asp-Lys-Lys-Asp, Glu-Lys-Lys-Glu, and Asp-Lys-Lys-Glu. Glu-Lys-Lys-Asp is by far the most favored of this group of four.

Various Consequences Brought About by the Fortuitous Presence of Active Sites on Irrelevant Proteins

In Figs. 1 and 2, we have seen that both the RGD site of ligands and the EKKD site of its receptors are common enough oligopeptides to the extent that they are found in fortuitous positions on irrelevant proteins. For these inadvertently present active sites to bring about meaningful consequences, deleterious or otherwise, each should occupy the exposed position on a protein, thus becoming accessible. At any rate, the RGD site present on bacterial α-lytic protease (Olson et al. 1978) as well as on human

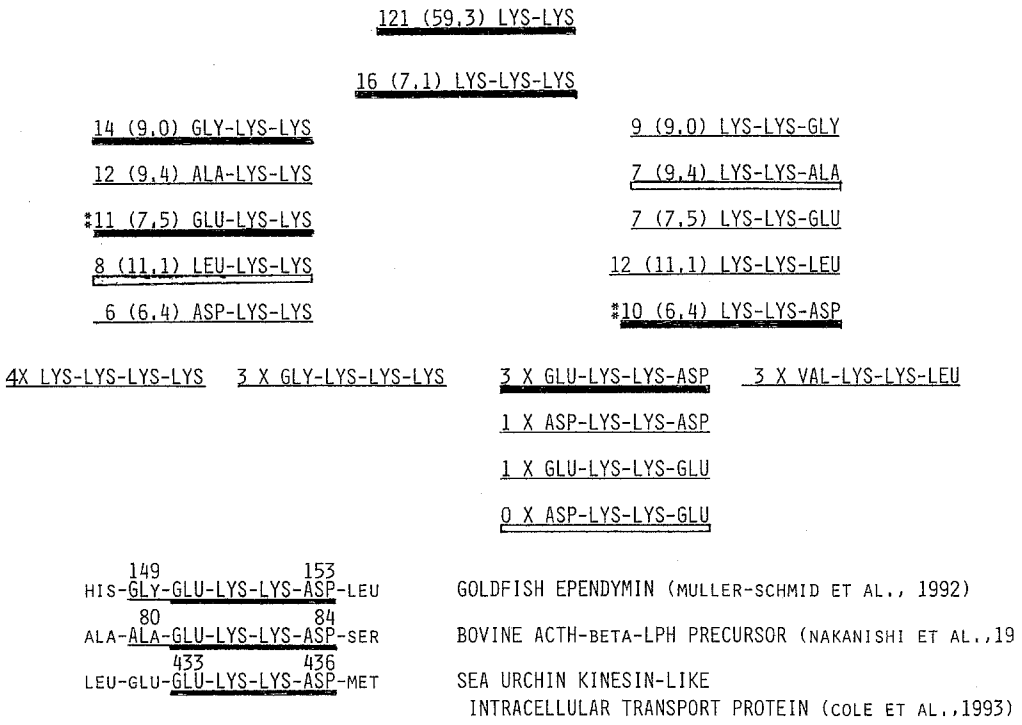

Fig. 2. At the **top,** observed vs expected frequencies of Lys-Lys dipeptide and Lys-Lys-Lys tripeptide are shown, followed by five pairs of Lys-Lys-containing tripeptides. The identical residue placed either in the front *(left column)* or the rear *(right column)* of Lys-Lys constituted each pair of tripeptides. Shown next in line are four tetrapeptides centered by Lys-Lys. Aligned below 3 × Glu-Lys-Lys-Asp are three related palindromic and pseudopalindromic tetrapeptides and their frequencies. Identified at the bottom are positions of three EKKD sites found in three proteins.

tyrosine hydroxylase (Nagatsu and Ichinose 1991) appeared to be of no consequence for obvious reasons. In the case of bovine ACTH-β-LPH precursor (Nakanishi et al. 1979), on the other hand, RGD and EKKD sites coexisted on the same precursor protein. It would follow that within relevant bovine pituitary cells, the tendency for autopolymerization is inherent in this precursor. Over the years, such tendency for autopolymerization might manifest itself in the formation of intracellular precipitates which would surely impair the function of relevant pituitary cells. In such an irrelevant interaction between ligand and receptor active sites, one might seek the reason for the species-specific pathology of the aging process.

The consequence of having two RGD sites on human fibrinogen α-chain (Watt et al. 1979) is also likely to be deleterious. For the possession places fibrinogen α-chain in the competitive position against legitimate matrix proteins in search of cell-surface EKKD sites.

What about the RGD site on one of the structural proteins of Sindbis virus (Rice and Strauss 1981)? Provided that this RGD site is exposed on the surface of infectious viral particles, this possession enables Sindbis virus to utilize integrins of the host cell as its receptor, thus gaining an entrance to the host cell. In fact, various viruses might change their preferred host cell types by such fortuitous possession of ligand active sites.

Among the teleost fish, the presence of Gly-Glu-Lys-Lys-Asp pentapeptide in ependymin of the goldfish was apparently fortuitous, since the corresponding pentapeptide was not seen in this protein of the rainbow trout (Müller-Schmid et al. 1992). Whatever the assigned function of this cerebrospinal fluid protein might be, it can be carried out without this pentapeptidic sequence. Yet through this pentapeptidic sequence, goldfish ependymin does form aggregates by interacting with fibronectin. The point of interest here is that in this particular instance, the formation of aggregates with fibronectin appeared to have led to the acquisition of a new function by goldfish ependymin (Shashoua 1991).

It has been shown that aggregates formed extracellularly between goldfish ependymin and fibronectin in the vicinity of synapses move into the neuronal cytoplasm. Inasmuch as cytoplasmic aggregates tend to concentrate at the base of a spine, it was thought that the strategic placement of these aggregates contributes to the conversion of a weak synapse to a strong synapse, thereby contributing to consolidation of the long-term memory (Shashoua 1991). Since this purportedly new function of ependymin is dependent upon its ability to form aggregates with fibronectin via its EKKD site, this newly acquired function of ependymin was created by the fortuitous acquisition of this pentapeptide. Once created, however, this new function appears to have become indispensable since the EKKD site persists in mouse as well as a human ependymin (Shashoua 1991).

References

Argravos WS, Suzuki S, Arai H, Thompson K, Pierschblacher MD (1987) Amino acid sequence of the human fibronectin receptor. J Cell Biol 105:1183–1190

Altangerel P, Norell MA, Chiappe LM, Clark JM (1993) Flightless bird from the cretaceous of Mongolia. Nature 362:623–626

Bost KL, Smith EM, Blalock JE (1985) Similarity between the corticotropin (ACTH) receptor and a peptide encoded by an RNA that is complementary to ACTH mRNA. Proc Natl Acad Sci USA 82: 1372–1375

Cole DG, Chinn SW, Wedaman KP, Hall K, Vuong T, Scholey JM (1993) Novel heterotrimeric kines in related protein purified from sea urchin eggs. Nature 366:268–270

Kollar EJ, Fisher C (1980) Tooth induction in chicken epithelium: expression of quiescent genes for enamel synthesis. Science 207: 993–995

Müller-Schmid A, Rinder H, Lottspeich F, Gertzen E-V, Hoffmann W (1992) Ependymins from the cerebrospinal fluid of salmonid fish: gene structure and molecular characterization. Gene 118:189–196

Muthukumaran G, Blumberg B, Kurkinen M (1989) The complete primary structure for the αl-chain of mouse collagen IV. J Biol Chem 264:6310–6317

Nagatsu T, Ichinose H (1991) Comparative studies on the structure of human tyrosine hydroxylase with those of the enzyme of various mammals. Comp Biochem Physiol 98C:203–210

Nakanishi S, Inoue A, Kita T, Nakamura M, Chang ACY, Cohen SN, Nathans J, Thomas K, Hogness DS, Numa S (1979) Nucleotide sequence of cloned cDNA for bovine corticotropin-β-lipotropin precursor. Nature 278:423–427

Ohno S (1991) To be or not to be a responder in T-cell responses: ubiquitous oligopeptides in all proteins. Immunogenetics 34:215–221

Ohno S (1992a) How cytotoxic T cells manage to discriminate nonself from self at the nonapeptide level. Proc Natl Acad Sci USA 89: 4643–4647

Ohno S (1992b) Of palindromes and peptides. Hum Genet 90:342–34

Olson MOJ, Nagabhushan N, Dzwiniel M, Smillie LB, Whitaker DR (1978) Primary structure of α-lytic protease: a bacterial homologue of the pancreatic serine proteases. Nature 228:438–442

Pierschbacher MD, Ruoslahti E (1984) Cell attachment activity of fibronectin can be duplicated by small synthetic fragments of the molecule. Nature 309:30–33

Rice CM, Strauss JH (1981) Nucelotide sequence of the 26S mRNA of sindbis virus and deduced sequence of the encoded virus structural proteins. Proc Natl Acad Sci USA 78:2062–2066

Ruoslahti E, Pierschbacher MD (1987) New perspectives in cell adhesion: RGD and integrins. Science 238:491–497

Seto Y (1989) Formation of proteins on the primitive earth. Evidence for the oligoglycine hypothesis. Viva Origino 17:153–163

Shashoua VE (1991) Ependymin, a brain extracellular glycoprotein, and CNS plasticity. Ann NY Acad Sci 627:94–114

Watt KWK, Cottrell BA, Strong DD, Doolittle RF (1979) Amino-acid sequence studies on the α-chain of human fibrinogen; overlapping sequences proving the complete sequence. Biochemistry 18:5410–5416