Motifs of Cadherin- and Fibronectin Type III-related Sequences and Evolution of the Receptor-Type-Protein Tyrosine Kinases: Sequence Similarity between Proto-Oncogene ret and Cadherin Family

Kei-ichi Kuma, *† Naoyuki Iwabe, *† and Takashi Miyata†

*Department of Biology, Kyushu University; and †Department of Biophysics, Kyoto University

Immunoglobulin (Ig)-, fibronectin type III (FN-III)-, and cadherin-related sequences are often found in multiple repeats in the extracellular regions of various cell adhesion molecules. The amino acid sequences of 82 different cadherin-like repeats from 13 known members of the cadherin superfamily were compared for six highly conserved regions, and a frequency matrix represented as amino acids versus position matrix was calculated based on the alignment. With the frequency matrix, further members of the cadherin superfamily were searched for in the protein data base. It was found that the ret protein, a receptor-type-protein tyrosine kinase, contains cadherin-like repeats in the extracellular region. A similar analysis was also carried out for the FN-III superfamily. Nine receptor-type-protein tyrosine kinases were shown to exhibit significant similarities, in terms of sequence, with known FN-III-like repeats. Several receptor-type-protein tyrosine kinases have already been reported to have Ig-like repeats in their extracellular regions. Thus these receptor-type-protein tyrosine kinases—together with the remaining receptors, whose structures are not yet characterized—may be classified into at least four distinct groups based on the structural differences in the extracellular domains. A molecular phylogenetic tree inferred from the shared kinase domains of these receptor-type-protein tyrosine kinases revealed a close relationship between the branching patterns and the grouping based on the structural differences of the extracellular region.

Introduction

Many protein sequences often share similarities with one another to form a large protein family. Among superfamilies now known, the immunoglobulin (Ig), fibronectin type III (FN-III), and cadherin superfamilies are particularly interesting in that (a) they comprise large family members from a wide range of species covering vertebrates and invertebrates and (b) the family members are often present in the extracellular domains of cell adhesion molecules and receptor molecules (Williams and Barclay 1988; Bazan 1990, 1991; Kuma et al. 1991a; Takeichi 1991). Although these members differ in function, the common role is, in most cases, molecular recognition at the cell surface (Ruoslahti 1988; Williams and Barclay 1988; Bazan 1990; Edelman and Crossin 1991; Takeichi 1991). Also, members of these families often have multiple repeats of a common structural unit of ~90–110 amino acids in length (Ruoslahti

1. Key words: cadherin, fibronectin type III, immunoglobulin, tyrosine kinase, superfamily, evolution.

Address for correspondence and reprints: Takashi Miyata. Department of Biophysics, Faculty of Science, Kyoto University, Kyoto 606, Japan.

© 1993 by The University of Chicago. All rights reserved.
0737-4038/93/1003-0004$02.00
1988; Williams and Barclay 1988; Bazan 1990, 1991; Edelman and Crossin 1991; Kuma et al. 1991a; Takeichi 1991). Within the repeats, there exist several highly conserved stretches of amino acids (motifs), which are hallmarks of each family.

These motifs are useful to find further family members in protein data bases. Even in these conserved segments, however, the number of unvaried amino acids is limited, and amino acid alternations are often observed. By counting the frequency of occurrence of amino acids at each position, on the basis of alignment of known family members, it is possible to evaluate the relative importance of alternative amino acids at each position and to construct a scoring system for searching for similar sequences (Staden 1984; Dodd and Egan 1987, 1990; Kuma et al. 1991b). For the Ig superfamily, we recently have reported the frequencies of amino acids at each position of four conserved segments, represented as amino acids versus position matrix (Kuma et al. 1991b). Using the frequency matrix, we have carried out computer-assisted searches for similar sequences in the protein data base and have found several novel members of the Ig family (Kuma et al. 1991b); to exclude a possibility of convergence, as well as to increase the detection power, several conserved motifs must be used.

We report here novel members of cadherin and FN-III superfamilies found in the extracellular domain of the receptor-type-protein tyrosine kinases. On the basis of the findings noted above, as well as on the basis of the phylogenetic tree of the kinase domains, we also propose that the receptor-type-protein tyrosine kinases be classified into at least four distinct groups.

Material and Methods

For each of the cadherin and FN-III superfamilies, the amino acid sequences of known members (master set) were compared for highly conserved motifs, where alignments are possible without introducing gaps; closely related sequences were excluded from comparison. On the basis of the alignment, frequency $f(X, S_i)$ of occurrence of amino acid $X$ at position $i$ of motif $S$ (frequency matrix) was calculated. We further introduced the weight matrix $W(X, S_i)$, defined as the frequency of occurrence of amino acid $X$ at position $S_i$ relative to the expected frequency in randomized sequences: $W(X, S_i) = \ln[f(X, S_i)/np(X)] \times 100$, where $n$ is the total number of sequences of the master set, and $p(X)$ is the average content of amino acid $X$ in the master set; for $f(X, S_i) = 0$, $W(X, S_i)$ was set to $\ln[1/(n+1)p(X)]$ (Dodd and Egan 1987, 1990; Kuma et al. 1991b). The weight matrix was used as a probe to search similar sequences in the NBRF data base by the method described by Kuma et al. (1991b). To exclude a possibility of convergence, as well as to increase detection power, the weight matrix consisting of multiple motifs was used (Kuma et al. 1991b).

To assess the statistical significance of observed similarity, the alignment score $A$ of a test sequence defined as $A = (M - \langle m \rangle)/SD$ was calculated, where $M$ is the maximum score of a test sequence, which was evaluated from the weight matrix, and $\langle m \rangle$ and SD are, respectively, the mean and the standard deviation of the scores of randomized sequences (Kuma et al. 1991b).

Results and Discussion

Motifs of Cadherin Repeats

Cadherins are a family of Ca$^{2+}$-dependent cell-cell adhesion molecules that comprise an extracellular region, a transmembrane region, and a cytoplasmic region. The extracellular region has four repeats with characteristic amino acid sequences (cadherin repeats), each consisting of ~110 amino acids, in all vertebrate cadherins identified...
to date (Takeichi 1991). There are marked sequence similarities between different cadherin repeats, suggesting that they have derived from a single precursor sequence by internal duplications (Hatta et al. 1988). The recently cloned fat locus of Drosophila has been shown to encode a protein containing 34 cadherin-like repeats in the extracellular region (Mahoney et al. 1991). Thus members of the cadherin superfamily are now extending over a wide range of species covering vertebrates and invertebrates.

At present, 12 different cadherin species from vertebrates, each with four repeats (Takeichi 1991), as well as the fat protein from Drosophila, have been reported, and thus amino acid sequences of 82 different cadherin-like repeats in total (master set) are now available for comparison. From a comparison of these sequences for six highly conserved motifs (I–VI), we have determined the frequency matrix of the cadherin-like repeats (fig. 1), which has been used to search for further members of the cadherin superfamily in the NBRF data base (release 30.0) by a recently developed method [Kuma et al. 1991b; a preliminary report has been published (Kuma et al. 1992)]. A cadherin-related sequence was found in the extracellular region of the ret protein, a receptor-type-protein tyrosine kinase whose ligand has not been determined (Takahashi et al. 1988, 1989); the ret protein has an alignment score of 5.0, which is highly significant. Of the 82 sequences of the master set, 33 sequences have score values <5.0.

Dot-matrix comparisons (Toh et al. 1983) between the ret protein and members of the cadherin superfamily revealed marked sequence similarities (data not shown). The amino acid sequence of a part of the ret protein was aligned with those of 12 vertebrate cadherins, as well as with those of four repeats of the Drosophila fat protein (fig. 2). The ret protein shares 20%–31% identity with cadherins compared. These similarities are statistically significant; as shown in figure 2, the probabilites that the observed similarities are realized by chance (Toh et al. 1983) are significantly small, except for two cases (i.e., the ret/desmoglein pair and the ret/M-cadherin pair). The similarity with chicken T-cadherin is 31% identity, and $P = 10^{-11}$. In the aligned region shown in figure 2, 14 positions are identical among cadherins and fat, of which 11 are preserved when ret is included in the comparison. Furthermore, the two motifs DXND and DXD, which are thought to be putative Ca$^{2+}$ binding sites (Ringwald et al. 1987; Ozawa et al. 1990), are conserved in ret. From these results, we conclude that the ret protein is a novel member of the cadherin superfamily. The ret protein is also likely to have another cadherin-like repeat in the C-terminal region, immediately followed by the aligned region shown in figure 2, although the similarity between the two repeats is very weak.

Figure 3 shows a phylogenetic tree of the members of the cadherin superfamily, inferred by the neighbor-joining method (Saitou and Nei 1987) and based on the alignment in figure 2. The inferred unrooted tree revealed three distinct clusters corresponding to vertebrate cadherins, Drosophila fat, and human ret. This classification correlates well with their structural differences: vertebrate cadherins are transmembrane glycoproteins with four cadherin repeats in the extracellular region and are involved in cell-cell adhesion (Takeichi 1991). The fat protein, on the other hand, is a very large transmembrane protein of >5,000 amino acids with 34 cadherin-like repeats in the extracellular region and a unique cytoplasmic region and is required for the control of cell proliferation in the imaginal disks (Mahoney et al. 1991). The ret protein differs from the above two subclasses in structure, except for the presence of shared cadherin-like repeats, although its function is still unclear: the ret protein is a receptor-type molecule with possibly two cadherin-like repeats in the extracellular region and
FIG. 1.—Frequency matrix derived from the master set of the cadherin-like repeats. a), Schematic representation of the location of six conserved motifs (I–VI) on the cadherin repeats. The first repeat of the mouse E-cadherin (Nagafuchi et al. 1987) is shown as an example. The start and end positions of the repeat are shown. Motifs I (at amino acid positions 190–194), II (196–198), III (222–225), IV (230–234), V (252–259), and VI (262–264) are shown. The amino acid sequence of each motif is also shown.

b), Frequency matrix shown as an amino-acid-vs.-position matrix. The amino acid length between different motifs differs for different sequences, and the range is shown in parentheses; e.g., the length between I and II varies from 0 to 2 amino acids.
FIG. 2.—Comparison of ret and members of the cadherin superfamily from vertebrates and invertebrates. Position numbers of the first and last amino acids of the aligned regions are shown in parentheses. Amino acids that are identical with that of ret are boxed. % = Percent identity between ret and each of the family members; Pr. = probability that the observed similarity is realized by chance (e.g., 1.0E-11 represents 1.0 x 10^-11); ● and ○ = positions that are occupied by identical and chemically similar amino acids among all the sequences compared, respectively; and = gap. The six conserved motifs are denoted by the underscoring at the bottom of the columns. Abbreviations and sequence data sources (data not referred to are from NBRF data base release 30.0) are as follows: ret = human ret; cT = chicken T-cadherin (Ranscht and Dours-Zimmerman 1991); CN = chicken N-cadherin; cLCAM = chicken L-CAM; cR = chicken R-cadherin; cB = chicken B-cadherin (Napolitano et al. 1991); mP = mouse P-cadherin; mE = mouse E-cadherin; mM = mouse M-cadherin (Donalies et al. 1991); bDC = bovine desmocollins (Mechanic et al. 1991); hDG3 = human desmosomal glycoprotein III (Parker et al. 1991); hPVA = human pemphigus vulgaris antigen (Amagai et al. 1991); hDG = bovine desmoglein; and Df = Drosophila fat (Mahoney et al. 1991).
Motifs of FN-III Repeats

Fibronectin, an extracellular matrix and plasma protein that plays a critical role in cell adhesion, contains three kinds of repeats: types I–III (Ruoslahti 1988). Amino acid sequences homologous to those of FN-III repeats were identified in a variety of protein species from vertebrates and invertebrates. These include protein tyrosine phosphatases (for review, see Kuma et al. 1991a), cell adhesion molecules (for review, see Kuma et al. 1991a), cytokine-receptor family (for review, see Bazan 1990, 1991), tissue factor (Mackman et al. 1989), C protein (Einheber and Fischman 1990), cytotactin (Jones et al. 1989), and twitchin (Benian et al. 1989). Together, these molecules constitute an FN-III superfamily. Most of these molecules have FN-III-related sequences in multiple repeats (FN-III-like repeats). In an invertebrate cytoplasmic protein twitchin, an interleukin receptor IL6R, and several cell adhesion molecules—including F3, contactin, TAG-1, neuroglian, L1, fasciclin, and NCAM—the FN-III-like repeats are present together with Ig-like repeats (Kuma et al. 1991b; for review, see Kuma et al. 1991a).

We have accumulated 24 members of the FN-III superfamily, with 122 repeats
in total as a master set; only distantly related sequences with percent identities to each other that are <40% were included. On the basis of the alignment of these FN-III-related sequences for four highly conserved motifs, I–IV, we have calculated the frequency matrix (fig. 4), as well as weight matrix, by which further members of the FN-III superfamily have been searched in the NBRF data base (release 30.0). We found FN-III-like repeats in several receptor-type-protein tyrosine kinases with significant alignment score; these are sevenless (score A = 4.3 SD) and its vertebrate homologue c-ros (A = 3.5 SD), insulin receptor IR (A = 4.2 SD), insulin receptor-related protein IRR (A = 3.0 SD), insulin-like growth factor I receptor IGFl R (A = 2.7 SD), elk (A = 4.1 SD), eck (A = 4.0 SD), eph (A = 2.9 SD), and torso (A = 2.3 SD). There are 31 sequences in the master set whose score values are <2.3 SD.

The amino acid sequences of these receptors were aligned with those of four known members of the FN-III superfamily (fig. 5); the aligned region was extended,

---

**Fig. 4.—**Frequency matrix derived from the master set of the FN-III-like repeats. a), Schematic representation of the location of four conserved motifs (I–IV) on the FN-III-like repeats. The third repeat of FN-III (Kornblihtt et al. 1985) is shown as an example. The start and end positions of the repeat are shown. Motifs I (at amino acid positions 798–803), II (810–812), III (836–841), and IV (845–851) are shown. The amino acid sequence of each motif is also shown. b), Frequency matrix shown as an amino-acid-vs-position matrix. The amino acid length between different motifs differs for different sequences, and the range is shown in parentheses; e.g., the length between I and II varies from 4 to 14 amino acids.
FIG. 5.—Alignment of nine receptor-type-protein tyrosine kinases with known members of the FN-III superfamily. Alignment is shown only for four conserved motifs; for divergent regions, only amino acid lengths are shown. 0 and O = Positions that are occupied by identical and chemically similar amino acids among all the sequences compared, respectively; and − = gap. On the right-hand side of the alignment, percent identity (%) and probability (Pr.) that the similarity is realized by chance are shown as a matrix, of the nine receptors vs. four known members; e.g., in the comparison between sev and FN, identity = 21 %, and Pr. = 1E-03 (= 1 \times 10^{-3}). Abbreviations and amino acid positions of the aligned region are as follows: FN = human fibronectin 1417-1501; DLAR = a Drosophila receptor phosphatase 709-802; L1 = mouse cell adhesion molecule L1 917-1004; Twitchin = Caenorhabditis elegans Twitchin 2356-2440; sev = Drosophila sevenless 1304-1388; torso = Drosophila torso 186-270. Sequence data were taken from NBRF data base release 30.0, except for elk (Lhoták et al. 1991).
to both sides, from the regions shown in figure 4. The percent identity and probability that the observed sequence similarity is realized by chance were also calculated, by a standard method (Toh et al. 1983). As judged from the estimated chance probabilities, the observed similarities are statistically significant, and thus these receptors are novel members of the FN-III superfamily. O’bryan et al. (1991) also noted similarities of these sequences with FN-III-like repeats except for torso, although no significance test has been done.

Classification of Receptor-Type-Protein Tyrosine Kinases

The receptor-type-protein tyrosine kinases may now be classified into at least four distinct groups based on the structural differences of the extracellular region. Growth factor receptors including fibroblast growth factor receptors (FGFRs), platelet-derived growth factor receptors (PDGFRs), colony-stimulating factor I receptor (CSF1R) and its close homologue kit, as well as flk2 and flt, all contain Ig-like repeats in the extracellular region (Williams and Barclay 1988; for review, see Kuma et al. 1991a). Recently we identified the Ig-like repeats in the extracellular region of neurotrophic factor receptors trk and its homologues (Kuma et al. 1991b). These receptors are classified as group I.

Group II is a group of receptors with FN-III-like repeats: these are the eck, elk, and eph subgroup; the IR, IGF1R, and IRR subgroup; the ros and sevenless subgroup; and torso. The ret protein has cadherin-like repeats in the extracellular region and forms a unique group, group III. The remaining group(s) include epidermal growth factor receptor (EGFR), hepatocyte growth factor receptor (HGFR), and ltk, whose structures are not yet characterized. Although both EGFR and HGFR have cysteine-rich regions, no apparent sequence similarity was detected in the extracellular regions.

Because these receptors share kinase domains, it is possible to reconstruct the phylogenetic tree of the receptors, which allows one to classify the receptors from an evolutionary viewpoint. Figure 6 shows a phylogenetic tree inferred from a comparison of the amino acid sequences of the kinase domains. As this tree shows, the branching pattern of these receptors roughly coincides with the classification based on the structural differences of their extracellular region, except for the case of torso. From the structural and evolutionary considerations, we suggest that the known receptor-type-protein tyrosine kinases can be classified into at least four distinct groups.

Strong Conservation of FGFR

The group I receptors have a common architecture, Ig-like repeats, and a kinase domain. Thus it is expected that, in the respective regions, the rates of evolution are similar for different receptors. For the four group I receptors, the nucleotide sequences are available for comparison between human and mouse (or rat). Table 1 shows the number of synonymous (silent) substitutions per site \(k_S\) and that of nonsynonymous (replacement) substitutions per site \(k_A\) in each of the Ig-like repeats and in the kinase domain. FGFR has small \(k_A\) values in both the Ig-like repeats and the kinase domain, when compared with the corresponding values of the other three receptors. It is unlikely that reduced mutation rate is responsible for the strong conservation of FGFR, because the \(k_S\) values are similar among the four receptors. The \(k_S\) values between different units of Ig-like repeats in a single chain are approximately constant within statistical fluctuation (data not shown). A possibility of paralogous comparison is also excluded, because, in table 1, genes with large \(k_A\) values are not always divergent at synonymous sites. Alternative interpretations include the following two: (1) Different configurations
of repeat units within a chain and/or different modes of interaction between chains would result in different $k_A$'s. (2) In a complex biochemical network, a molecule placed in a central position of the network would be constrained against amino acid alternations much more strongly than would a molecule in a peripheral position. Because both the Ig-like repeats and the kinase domain are conserved strongly in FGFR, the latter interpretation is plausible. A similar difference is also found in the FN-III superfamily; in the FN-III-like repeats, the insulin receptor evolves at a rate $\sim$10 times slower than that of the ros protein.
Table 1

$k_s$ and $k_A$ Values of Extracellular and Cytoplasmic Regions of Several Receptor-Type-Protein Tyrosine Kinases

<table>
<thead>
<tr>
<th>GENE</th>
<th>EXTRACELLULAR-REGION VALUE</th>
<th>CYTOPLASMIC-REGION (Kinase-Domain) VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$k_s$</td>
<td>$k_A$</td>
</tr>
<tr>
<td>Ig superfamily:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CSFIR</td>
<td>0.90</td>
<td>0.27</td>
</tr>
<tr>
<td>kit</td>
<td>0.90</td>
<td>0.15</td>
</tr>
<tr>
<td>PDGFR</td>
<td>0.79</td>
<td>0.11</td>
</tr>
<tr>
<td>bFGFR</td>
<td>0.64</td>
<td>0.0034</td>
</tr>
<tr>
<td>FN-III superfamily:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ros</td>
<td>0.68</td>
<td>0.11</td>
</tr>
<tr>
<td>Insulin R</td>
<td>0.99</td>
<td>0.010</td>
</tr>
<tr>
<td>Cadherin superfamily:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N-cadherin</td>
<td>0.72</td>
<td>0.011</td>
</tr>
</tbody>
</table>

NOTE.—For each gene except ros, the nucleotide sequences were compared between human and mouse; for ros, human and rat sequences were compared. $k_s$ and $k_A$ were calculated by Miyata and Yasunaga's (1980) method; multiple substitutions were corrected for (e.g., see Kimura 1983, pp. 55–97). Only the Ig-like repeats and the FN-III-like repeats in the extracellular region and the tyrosine kinase domain in the cytoplasmic region were compared. Sequence data were taken from the GenBank data base (release 70.0).

Acknowledgments

We thank Prof. M. Takeichi for discussion and comments. This work was supported in part by grants from the Ministry of Education, Science and Culture of Japan.

LITERATURE CITED


cell adhesion molecule uvomorulin: insights into the molecular mechanism of Ca^{2+}-dependent
Res. 12:505–519.
Takahashi, M., Y. Buma, and H. Hiai. 1989. Isolation of ret proto-oncogene cDNA with an
and expression of the ret proto-oncogene encoding a tyrosine kinase with two potential
251:1451–1455.
Toh, H., H. Hayashida, and T. Miyata. 1983. Sequence homology between retroviral reverse
transcriptase and putative polymerases of hepatitis B virus and cauliflower mosaic virus.
Williams, A. F., and A. N. Barclay. 1988. The immunoglobulin superfamily—domains for

TAKASHI GOJOBORI, reviewing editor

Received August 31, 1992; revision received January 5, 1993

Accepted January 5, 1993