Available online at www.sciencedirect.com
ScienceDirect
Genomics 88 (2006) 513–520
GENOMICS
www.elsevier.com/locate/ygeno

IGFL: A secreted family with conserved cysteine residues and similarities to the IGF superfamily

Peter Emtage, Paolo Vatta, Matthew Arterburn, Matthew W. Muller, Emily Park, Bryan Boyle, Sophie Hazell, Renee Polizotto, Walter D. Funk *, Y. Tom Tang

Nuvelo, San Carlos, CA 94070, USA

Received 29 January 2005; accepted 31 May 2006
Available online 4 August 2006

Abstract

We have discovered a family of small secreted proteins in *Homo sapiens* and *Mus musculus*. The IGF-like (IGFL) genes encode proteins of approximately 100 amino acids that contain 11 conserved cysteine residues at fixed positions, including two CC motifs. In *H. sapiens*, the family is composed of four genes and two pseudogenes that are referred as IGFL1 to IGFL4 and IGFL1P1 and IGFL1P2, respectively. Human IGFL genes are clustered together on chromosome 19 within a 35-kb interval. *M. musculus* has a single IGFL family member that is located on chromosome 7. Further, evolutionary analysis shows a lack of direct orthology between any of the four human members and the mouse gene. This relationship between the mouse and the human family members suggests that the multiple members in the human complement have arisen from recent duplication events that appear limited to the primate lineage. Structural considerations and sequence comparisons would suggest that IGFL proteins are distantly related to the IGF superfamily of growth factors. IGFL mRNAs display specific expression patterns; they are expressed in fetal tissues, breast, and prostate, and in many cancers as well, and this pattern is consistent with that of the IGF family members.

Keywords: Signal peptide; Intercellular signaling proteins; IGF superfamily; Conserved cysteine residues; Growth factors; *Homo sapiens*; *Mus musculus*; Molecular cloning; Protein family informatics; Gene duplication

The completion of the human and mouse genomes now allows the mapping of entire gene families and their orthologous relationships. Such work has been done recently for important protein families, such as the tumor necrosis factor superfamily [1], the insulin growth factor (IGF) family [2], the receptor tyrosine kinase family [3], and the carcinoembryonic antigen family [4] and will be gradually extended to all protein families within the human genome. We have recently finished the characterization of two protein families, the C1q-domain-containing protein family and the progestin and adipoQ receptor family [5,6]. The objective of these studies was to define the complete complement in the human genome, including transcript variants and pseudogenes.

In a previous study [7], we identified the TAFA family as a novel secreted family marked by conserved cysteine residues. With the aim of identifying potential protein therapeutics, we have continued a broad review of potential secreted protein families to distill the complete complement of each family in the human genome.

The IGF family includes a large proportion of potentially valuable therapeutic proteins; a total of 10 active peptides have been identified, including insulin, insulin-like growth factors I and II (IGF-1, IGF-2), relaxin 1–3, Leydig cell-specific insulin-like peptide (LeyIL), early placenta insulin-like peptide (EPIL), and insulin-like peptides 5 and 6 [2,8]. Within the mature protein, all members have six cysteines split into two chains (A chain, B chain). The cysteine patterns are $CGX_{10}C$ in B chain and CCX_3CX_8C in A chain. Structurally, the first A-chain cysteine is linked to the third cysteine within the A chain by a disulfide bond. The second and fourth cysteines in A chain are linked by interchain disulfide bonds to the two cysteines in B chain.

The IGF signaling system plays a critical role in cellular energy metabolism and in growth and development, especially prenatal growth [9,10]. Insulin is involved in the regulation of

* Corresponding author. Fax: +1 650 517 8041.
E-mail address: wfunk@nuvelo.com (W.D. Funk).

0888-7543/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2006.05.012

normal glucose homeostasis, as well as other specific physiological functions [8]. In humans, IGF-1 and IGF-2 maintain distinct expression patterns [11]. IGF-1 is the primary protein involved in the cellular response to growth hormone, particularly in bone growth and cartilage metabolism [12,13]. Relaxins play a critical role in the development of the mammary gland as well as in many of the physiological processes involved in pregnancy and labor, including growth and softening of the cervix. EPIL is expressed during the "invasive" phase of placental development. In males, LeyIL has been shown to be involved in testes descent. In addition, IGFs have powerful antiapoptotic effects [9]. The IGF system has also been implicated in various pathological conditions and is thought to play a prominent role in tumorigenesis [10]. High levels of circulating IGF-1 constitute a risk factor for the development of breast, prostate, lung, and colon cancer. Therefore, IGF proteins are important not only for growth, but also for the development of male and female reproductive systems and cancer development and progression.

We have discovered a novel secreted family of human proteins that we have named the IGF-like (IGFL) family. The family is composed of four expressed genes and two pseudogenes marked by 11 conserved cysteine residues, including two CC motifs, and IGFL members are predicted to share structural homology to the IGF family.

Results

IGFL family genes encode small secreted proteins

We discovered the IGFL family by applying a strategy that sorts available EST and cDNA sequences into contigs [7]. The algorithm is based on TBLASTN similarities of the assembled sequences within themselves [14,15]. Namely, if there is a novel protein family within the assembled contigs, family members will share statistically significant protein-level sequence similarities. Gene clusters are generated based on TBLASTN *S* score greater than 150 and putative protein sequences are extracted based on TBLASTN alignments. We then selected specific clusters for further review based on the presence of at least one putative protein containing a signal peptide [16] and the absence of putative proteins containing a predicted transmembrane domain. In this way, we are able to narrow down to a manageable set of clusters that are likely to encode novel secreted protein families.

The identified IGFL cluster initially contained three members and these contigs were then extended to complete coding sequences employing additional information from genomic sequences. We have named these IGFL1, IGFL2, and IGFL3. Upon mapping these sequences to the human genome, we observed that the three members clustered together on chromosome 19. A more detailed look at this genomic interval revealed one additional potential IGFL gene and two likely pseudogenes, based on interrupted coding regions. We designed primers to each of the predicted family members and were able to confirm the transcription of coding mRNAs for IGFL1 to IGFL4, while the two putative pseudogenes failed to produce PCR products from any of the cDNA sources tested, suggesting that these are not transcribed genes (data not shown).

Mapping the human IGFL genes to the mouse genome identified only a single homologous sequence that cannot be assigned direct orthology with any single human IGFL sequence. Fig. 1 shows the ClustalW alignment [17] of the five predicted amino acid sequences of the IGFL family from human and mouse. The divergent leading signal peptide sequence is followed by a more conserved region representing

```
IGFL-1      M--APRGCIVAVFAIFCISRLLCSHG APVAPMTPYLMLCQPHKRCGDKFYDPLQHCCYDD
IGFL-2      M-RTDYPRSVLAPAYVSVCLLLLCPREVIA PAGSEPWLCQPAPRCGDKIYNPLEQCCYND
IGFL-3      M-RPR-CCILALVCWITVFLLQCSK GTTDAPVGSGLWLCQPTPRCGNKIYNPSEQCCYDD
IGFL-4      MVPRISAAIFIFELLGSNS EGVTDLR---------LWLCQPAPRCGEWTYNPLEQCCDDG
IGFL_Mm     M-KIRNACAVLIEVLLFILEGVT GARKISTFSGPGSWPCN--PKCDGRTYNPSEECCVHD
            *        .                                *:   :*.   *:* :.** ..

IGFL-1      AVVPLARTQTCG-NCTFRVCFEQCCPWTF-------MV-KLINQN--CDSARTSDDRLCRSV
IGFL-2      AIVSLSETRQCGPPCTFWPCFELCCLDSFGLTNDFVVKLKVQGVNSQCHSSPIS--SKCESR
IGFL-3      AILSLKETRRCGSTCTFWPCFELCCPESFGPQQKFLVKLRVLGMKSQCHLSPIS--RSCTRN
IGFL-4      VILDLNQTRLCGSSCTFWPCFQHCCLESLGSQNQTVVRFKVPGMKPDCKSSPIT--RICAQE
IGFL_Mm     TILPFKRINLCGPSCTYRPCFELCCPESYSPKKKFIVKLKVHGERSHCSSSPIS--RNCKSN
            .:: : . . **  **:  **: **  :          .:: . .  *  :.      *

IGFL-1      S----------------------     87aa
IGFL-2      RRFP-------------------     95aa
IGFL-3      RRHVLYP----------------    102aa
IGFL-4      YHPKSPVSRSDLI----------    106aa
IGFL_Mm     KIFHGEDIEDNQLSLRKKSGDQP    118aa
```

Fig. 1. ClustalW alignment of IGFL family members in human and mouse. Signal peptides are in italic and underlined. Conserved cysteine residues are in bold. Asterisks (*) indicate identical residues, colons (:) conserved residues, and periods (.) semiconserved residues. The overall pair-wise identity is 27–55%; in the region representing the predicted mature protein, pair-wise identity is 26–63%. The amino acid lengths of the mature proteins are indicated.