LETTER TO THE EDITOR

Identification of an exceptionally short open reading frame in the genome of man, encoding a decapeptide, which regulates granulopoiesis by negative feedback

T. Rytömaa

Radiation and Nuclear Safety Authority, Helsinki, 00250, Finland

Received 30 January 2014; revision accepted 22 February 2014

An unsolved current puzzle with fundamental importance is the identification of a subset of functional short open reading frames (ORFs) in mammalian genomes (1–3). These peptide-encoding genes are buried in an enormous pile of meaningless short ORFs. Here, I describe a human gene with a functional short ORF, which encodes a decapeptide that regulates granulopoiesis by negative feedback.

In man, the vast majority of mature neutrophils exists in a large bone marrow reserve (4). These cells have been shown to release an inhibitory factor, termed granulocytic chalone (5–15), into their environment and hence maintain physiological homeostasis (16) between cell production and cell loss (mobilization of cells into the blood). Attempts to identify the exact chemical nature of the inhibitor initially did not yield conclusive results. Partially purified preparations made by *Weddel Pharmaceuticals Ltd.* from ox granulocytes suggested that the biologically active factor would be a small peptide. These substances were used in the trials to treat human leukaemia (11,12) and also in tests on a variety of leukaemic cell lines propagated *in vitro* (13). Later, Paukovits and Laerum isolated and synthesized a haemoregulatory peptide pGlu-Glu-Asp-Cys-Lys (14,15), which was shown to inhibit granulocyte progenitor cells *in vitro* and *in vivo*. It must be noted, however, that the sequence of the isolated peptide was not determined unambiguously (14,15,17).

Owing to fundamental importance of identifying the exact chemical nature of the feedback signal, I recently decided to re-study the problem by screening the human genome. Searches for a gene encoding the inhibitory peptide were performed using the basic local alignment search tool (Blast) available at <http://www.ncbi.nlm.nih.gov/blast/Blast.cgi>.

Amino acid sequence of the isolated sequence appeared to be pGlu-Glu/Asp-Glu/Asp-Cys-Lys (14,15,17). As pGlu can be formed from either Glu or Gln, and because

each of the amino acids of this pentapeptide has two alternative codons, total number of nucleotide sequences that translate to one of the putative pentapeptides is 258. Blast search of the human genome with a single pentapeptide sequence gives an E value of 5.1, that is, the expected total number of possible sequences in the human genome would be $5.1 \times 258 \approx 1316$. This number was far too large for a meaningful search for a gene encoding the pentapeptide. The conclusion was that if the pentapeptide should be a cleavage product of a protein encoded by a canonical gene, it could not be found by searching the human genome. However, if the small peptide happened to be encoded by a non-canonical gene with a short ORF, the gene could be identified.

Thus, to reach a sensible level for analysis, several assumptions were made. First, as spontaneous conversion of N-terminal Gln to pGlu appears to be more common than enzymatic conversion of Glu to pGlu, Gln codons were selected as the first codon. Second, judging from inhibitory effects of several synthetic pentapeptides (17), the amino acid sequence Glu-Asp was considered to be more likely than other alternatives (Asp-Glu, Glu-Glu, Asp-Asp). Finally, the third (critical) assumption was then made: the sequence Gln-Glu-Asp-Cys-Lys may be the C-terminal end of a small polypeptide encoded by a short ORF. It should be noted that this assumption arose from the fact that the N-terminal amino acid is pGlu and therefore the small peptide is necessarily cleaved at this site, but need not be cleaved elsewhere. E value for any single nucleotide sequence encoding the pentapeptide QEDCK, followed by a stop codon, is 0.083 in the human genome. This suggested that the total number of expected sequences would be as small as $0.083 \times 3 \times 32 \approx 8$.

Of the theoretically possible 96 (3×32) nucleotide sequences, 12 were actually present. Of these locations

AATGCAAGA**GGCTAATCT**GGCTG**AATAAG**CACAT**TTAAAG**TTTCTGCTCACATGT**GGCAAATCT**CATGTTT
 GTCCACATCCACTGTCCACAGCAGTCATGTGGCAACAT**GAGGGGAGGGGCCAGGAGGACTGCAAGTA**
GAGTAAAGAGTGTATCCTGCCAACTGGGAGCACTCAATGTCATGTGGCAAAGTTGTTGATGAGTG

■ ■ - ■ ■ ■ ■ ■ ■ ■ ■

Figure 1. Short open reading frame codons ATG and TAG (bold), CAAT boxes (red), possible TATA boxes (green), possible polyadenylation signals (blue) and the possible transcription start site A₊₁ (bold). Squares below apparent transcript 3'UTR show the possible base pairings with microRNA-223 (open squares indicate wobble pairs GU). Note that a similar base pairing, not shown in the figure, is also present in the first upstream CAAT box area of the gene.

10 had an upstream in-frame stop codon and thus the potential pentapeptide was not in an open reading frame. Of the remaining two sequences, one had two in-frame ATG codons before an in-frame stop codon. However, neither of the start codons appeared to be a good candidate for initiating a functional ORF (for example, no similarity with consensus Kozak sequence). The conclusion was that this ORF would be a member of the enormous pile of meaningless short reading frames (1).

The last sequence had also two in-frame ATG codons before an in-frame stop codon. The first upstream ATG was in a context almost identical to the Kozak consensus sequence (GCCA/GCCATGG/A), while the other ATG, 15 codons upstream, had no similarity with Kozak. Furthermore, the second ATG was one nucleotide downstream of a CAAT box and part of the binding site of the transcription factor p53. Thus, it was highly unlikely that this ATG would be a start codon of a functional ORF.

The genomic sequence of the short ORF with 10 codons is in chromosome 3 in the human genome, an identical sequence is also present in chromosome 3 of all the great apes. The 210 nucleotide long genomic sequence is shown in Fig. 1.

The short open reading frame encodes a decapeptide Met-Arg-Gly-Gly-Ala-Gln-Glu-Asp-Cys- Lys. I suggest that neutrophil elastase cleaves the decapeptide between Ala and Gln, and that N-terminal Gln is spontaneously converted to pGlu. Therefore, the N-terminal of the decapeptide would be a short signal peptide and the C-terminal, the inhibitory pentapeptide, identical to the haemoregulatory peptide synthesized by Paukovits and Laerum (14,15).

The upstream sequence shown in Fig. 1 has binding sites for several transcription factors, including AML-1, CEBP and p53. These sites are not marked in the sequence, but they may be identified by the TFINFO search tool available at <http://tfbind.hgc.jp/> (18).

It is also important to note that microRNA-223 is a myeloid-specific microRNA participating in regulatory signalling networks involving lineage-specific transcription factors (19,20). As indicated in the figure, microRNA-223 has complementary homology (10/11) over a stretch of the apparent 3'UTR of the transcript. Consid-

ering suggested function of the decapeptide, apparent binding of microRNA-223 with the transcript (and possibly also with the promoter region of the genomic sequence) is consistent with the observation that overexpression of microRNA-223 increases granulopoiesis, and its knockdown, meanwhile, impairs it (19,20).

The conclusion that the small gene encodes an inhibitor of granulopoiesis is supported by the following considerations: under normal physiological conditions, granulopoiesis in the bone marrow remains at a constant level. In the case of acute demand for neutrophils in the blood periphery (e.g. during bacterial infection), cells from the large bone marrow reserve are discharged into the blood and transported to the site of bacterial invasion. Consequently, bone marrow content of the inhibitory factor is reduced and rate of cell proliferation is increased. Quiescent stem cells and mitotically more active myeloid precursor cells are stimulated. In addition to the sudden reduction in inhibitor in the bone marrow, and to the role played later by granulocyte colony stimulating factor (21,22), another stimulator seems to be formed from the inhibitor at the site of the inflammatory focus (23,24).

Mature neutrophils migrating to a site of bacterial invasion use different means of destroying and neutralizing the intruders. In these processes, inhibitory pentapeptide pEEDCK is oxidized to a disulphide-bonded homodimer (pEEDCK)₂ (23). This dimer is a stimulator of granulopoiesis and leads to entry of quiescent haemopoietic stem cells into the cell cycle (24). It is likely that the dimer acts by inducing expression of granulocyte colony-stimulating factor (21,22), which is the direct stimulator of the appropriate stem cells.

It is intellectually appealing that granulopoiesis is controlled by negative feedback. However, acceptance of this general principle requires identification of the mechanism mediating the feedback signal. Above, I have shown that the signal in man appears to be a small peptide encoded by a short ORF, and that expression of the gene may be regulated by microRNA-223. Identification of a peptide-encoding short ORF shows that the feedback signal is not, as might have been expected (17), a cleavage product of a protein encoded by a canonical gene. This finding provides a hint that also

other chalone-type inhibitory oligopeptides (25) may be encoded by short ORFs and may thus represent “beautiful needles in the haystack” (1).

References

- Basrai MA, Hieter P, Boeke JD (1997) Small open reading frames: beautiful needles in the haystack. *Genome Res.* **7**, 768–771.
- Ladoukakis E, Pereira V, Magay EG, Eyre-Walker A, Couso JP (2011) Hundreds of putatively functional small open reading frames in *Drosophila*. *Genome Biol.* **12**, R118.
- Slavoff SA, Mitchell AJ, Schwald AG, Cabili MN, Ma J, Levin JZ *et al.* (2013) Peptidomic discovery of short open reading frame-encoded peptides in human cells. *Nat. Chem. Biol.* **9**, 59–64.
- Osgood EE (1954) Number and distribution of human hemic cells. *Blood* **9**, 777–787.
- Rytömaa T, Kiviniemi K (1968) Control of granulocyte production. I Chalone and antichalone, two specific humoral regulators. *Cell Prolif.* **1**, 329–340.
- Rytömaa T, Kiviniemi K (1969) Chloroma regression induced by the granulocytic chalone. *Nature* **222**, 995–996.
- Paukovits WR (1971) Control of granulocyte production: separation and identification of a specific inhibitor (chalone). *Cell Prolif.* **4**, 539–547.
- Benestad HB, Rytömaa T, Kiviniemi K (1973) The cell specific effect of the granulocytic chalone demonstrated with the diffusion chamber technique. *Cell Prolif.* **6**, 147–154.
- Bateman AE (1974) Cell specificity of chalone-type inhibitors of DNA synthesis released by blood leucocytes and erythrocytes. *Cell Prolif.* **7**, 451–461.
- Lord BI, Cercek L, Cercek B, Shah GP, Dexter TM, Lajtha LG (1974) Inhibitors of cell proliferation?: specificity of action within the haemopoietic system. *Br. J. Cancer* **29**, 168–175.
- Rytömaa T, Vilpo JA, Levanto A, Jones WA (1976) Effect of granulocytic chalone on acute and chronic granulocytic leukaemia in man. Report of seven cases. *Scand. J. Haematol.* **17**(suppl. 27), 5–28.
- Rytömaa T, Vilpo JA, Levanto A, Jones WA (1977) Effect of granulocytic chalone on acute myeloid leukaemia in man. *The Lancet* **309**, 771–774.
- Foa P, Maiolo TA, Lombardi I, Rytömaa T, Polli EE (1982) Effect of granulocytic chalone on the growth rate of continuous cell lines propagated *in vitro*. A new assay system. *Scand. J. Haematol.* **29**, 257–264.
- Paukovits WR, Laerum OD (1982) Isolation and synthesis of a hemoregulatory peptide. *Z. Naturforsch. C.* **37**, 1297–1300.
- Paukovits WR, Laerum OD (1984) Structural investigations on a peptide regulating hemopoiesis *in vitro* and *in vivo*. *Hoppe Seylers Z. Physiol. Chem.* **365**, 303–311.
- Bullough WS, Rytömaa T (1965) Mitotic homeostasis. *Nature* **205**, 573–578.
- Laerum OD, Frostaad S, Tøn HI, Kamp D (1990) The sequence of the hemoregulatory peptide is present in G_{iα} proteins. *FEBS Lett.* **269**, 11–14.
- Tsunoda T, Takagi T (1999) Estimating transcription factor probability on DNA. *Bioinformatics* **15**, 622–630.
- Johannidis B, Harris MH, Wheeler RT, Stehling-Sun S, Lam MH, Kirak O *et al.* (2008) Regulation of progenitor cell proliferation and granulocyte Tfunction by microRNA-223. *Nature* **451**, 1125–1129.
- Vian I, Di Carlo M, Pelosi E, Fazi F, Santoro S, Cerio AM *et al.* (2013) Transcriptional fine-tuning of microRNA-223 levels directs lineage choice of human hematopoiesis progenitors. *Cell Death Differ.* **21**, 290–301.
- Metcalfe D (1985) The granulocyte-macrophage colony-stimulating factors. *Science* **229**, 6–22. TC
- Souza IM, Boone TC, Gabrilove J, Lai PH, Zsebo KM, Murdock DC *et al.* (1986) Recombinant human granulocyte colony-stimulating factor: effects on normal and leukemic myeloid cells. *Science* **232**, 61–65.
- Paukovits WR, Paukovits JB, Mose MH, Konstantinov S, Schulte-Hermann R (1998) Activated granulocytes oxidize the endogenous stem cell inhibitory peptide pGlu-Glu-Asp-Cys-Lys (pEEDCK) to the stimulatory dimer: a redox-mediated mechanism for demand-induced hematopoietic regulation. *Exp. Hematol.* **26**, 351–356.
- Paukovits JB, Paukovits WR (1995) Stem cell stimulation *in vitro* by the pentapeptide (pEEDCK)₂: a single-factor alternative for multifactor cocktails. *Leukemia* (suppl. 1), S48–S52.
- Elgjo K, Reichelt KL (2004) Chalone: from aqueous extracts to oligopeptides. *Cell Cycle* **3**, 1208–1211.