Glycoconjugate Journal (1996) **13:** 735–740

Alternative splicing of repetitive units is responsible for the polydispersities of integumentary mucin B.1 (FIM-B.1) from *Xenopus laevis*

WERNER JOBA[1,2] and WERNER HOFFMANN[2]*

[1]*Max-Planck-Institut für Psychiatrie, Abteilung Neurochemie, D-82152 Martinsried, Germany*
[2]*Institut für Molekularbiologie und Medizinische Chemie, Otto-von-Guericke-Universität, D-39120 Magdeburg, Germany*

Received 20 August 1995, revised 22 December 1995

Frog integumentary mucin B.1 (FIM-B.1) represents a polymorphic extracellular mosaic protein which contains tandemly arranged serine/threonine-rich modules as well as cysteine-rich domains. The latter are probably important for oligomerization of FIM-B.1 and have also been found in many proteins of the complement cascade as well as regions homologous to von Willebrand factor. The repetitive modules are targets for extensive O-glycosylation. Previous cDNA cloning experiments clearly established polydispersities within the same individual, which originate from deletions/insertions in the repetitive domain. Here, we analyse part of the corresponding genomic region. Each repetitive unit as well as the cysteine-rich domain is encoded by an individual class 1-1 exon typical of shuffled modules. Alternative splicing of these multiple cassettes creates the polydisperse FIM-B.1 transcripts.

Keywords: mucin, O-glycosylation, polydispersity, tandem repeats, exon shuffling, alternative splicing, short consensus repeat, shuffled module

Introduction

A layer of lubricant mucus protects many delicate epithelial surfaces, by forming a viscoelastic gel-like matrix (for review see [1]). Such gels mainly consist of a complex pattern of various mucins. Within the last few years, a number of mucins have been analysed on a molecular level ranging from amphibia to man. Thus far, more than seven human *MUC* genes have been localized and in *Xenopus laevis* at least three integumentary mucins are known, designated as FIM-A.1, FIM-B.1 and FIM-C.1 (for review see [2]). Generally, most mucins can be classified as typical extracellular mosaic proteins containing a mainly repetitive serine/threonine-rich O-glycosylated region flanked by cysteine-rich modules (for compilation see [3]). For example, in FIM-B.1 the central O-glycosylated part consists mainly of 11 amino acid residue long repetitive cassettes (type B repeats) which are interrupted by at least one cysteine-containing module, the so called 'short consensus repeat/SCR' (also known as 'complement control protein/CCP motif' or 'sushi domain'); an additional cysteine-rich module is found at the C-terminal end with homology to von Willebrand factor [4, 5].

The carbohydrate content of these typical glycoconjugates can make up to 80% of their mass. O-glycosylation of mucins is achieved by a variety of glycosyltransferases acting with distinct specificity on the peptide and the carbohydrate moiety [6–9]. Also the differentiation of the particular cell type, e.g. during tumourigenesis, plays a major role on the glycosylation pattern of mucins [10–12].

Determinants of the viscoelastic properties of mucins are their highly expanded conformation and their high molecular mass [13] resulting in an unusually large hydrodynamic radius [14]. Two basic strategies have been observed creating such long molecules. First, O-glycosylation of the serine/threonine-rich regions confers the typical stiff and rigid conformation. This classical hallmark is common to all mucins but is subject to

*To whom correspondence should be addressed.

great variations concerning the primary sequence. Second, many mucin subunits are able to aggregate to long linear polymers with no branch points [15]. This polymerization is probably achieved via the cysteine-rich modules at the ends of the molecules [3].

The repetitive nature of the O-glycosylated portion of mucins is responsible for the genetic polymorphism observed. This general feature has been documented manyfold, e.g. for MUC1 [16] and FIM-A.1 [2]. Here, the variable number of tandem repeats causes different lengths of the O-glycosylated domains within different individuals.

Furthermore, many mucin mRNAs – even when isolated from a single individual – show polydispersities after Northern blot analysis. Thus far, it is not entirely clear if this is the result of a particular instability of mucin mRNAs and/or genetically based variations. Only for frog skin mucins FIM-B.1 and FIM-C.1 the polydispersities have been unambiguously analysed at the cDNA level [5, 17]. As a hallmark, individual cDNA clones differ within the repetitive serine/threonine-rich region by insertions and deletions of various cassettes; this led to the hypothesis that the polydispersities originate by alternative splicing [5]. Here, we analyse the corresponding part of the *FIM-B.1* gene in order to bring evidence for this cassette model.

Materials and methods

Two $\times 10^5$ recombinant phages of a genomic *X. laevis* library kindly provided by Dr G. Spohr (Genève) were screened with a FIM-B.1 specific probe. This library was constructed by insertion of partially *Hae* III-digested DNA from *X. laevis* livers into Charon 4A λ phages via *Eco* RI linkers. As an FIM-B.1 specific probe the radioactively labelled insert of cDNA clone pFIM-5′-21 [5] was used. After rescreening of 24 positive phages with the FIM-B.1 specific oligonucleotide SCR2 d(GGGAGGTTGTGCTGATCCAGGG), phage λFIM-B-17 was chosen for further analysis containing more than 20 kb genomic *X. laevis* DNA. Recombinant phages were grown in *Escherichia coli* LE392 [18], and preparative amounts of DNA were purified using Qiagen-tip 20 or 100 (Diagen).

DNA from phage λFIM-B-17 was cut with *Eco* RI and a Southern blot analysis was performed using oligonucleotide SCR1 d(CACAGCTTGGTGTATTTC) as a probe. This oligonucleotide recognizes the short consensus repeat in FIM-B.1. After subcloning the positively hybridizing 6.5 kb *Eco* RI restriction fragment into pBluescript-II/SK$^-$ (Stratagene), plasmid pGFIM-B-17.1 was obtained.

Furthermore, using λFIM-B-17 as a template, a DNA fragment was amplified with the help of the polymerase chain reaction (PCR) and the synthetic oligonucleotides XGL12 d(CCCGGATCCGCTACACTTAAATCTACA) and REP7 d(CCCTCGAGAATTCGGATCCTGCTACCGTTCCGTTT). The underlined regions represent parts of FIM-B.1 cDNA sequence described previously [5]. XGL12 was deduced from a region encoding the ATLKST sequence whereas REP7 recognized part of the short consensus repeat. After restriction with *Bam* HI, the amplified fragment was subcloned into pBluescript-II/SK$^-$ yielding plasmid pG17-X12R7.1.

Plasmid DNA was purified with Quiagen-tip 20 (Diagen) and sequencing of double-stranded DNA was accomplished with a Sequenase kit (version 2.0, US Biochemicals) using [α-^{35}S]dATP for labelling. The full sequence of pGFIM-B-17.1 was determined after subcloning various restriction fragments into pBluescript-II/SK$^-$ or pBluescript-II/KS$^-$. Computerized analysis have been described previously [19].

Results

Figure 1 represents schematically the 7.0 kb long genomic region analysed by sequencing the combined inserts of pG17-X12R7.1 and pGFIM-B-17.1. Here, pG17-X12R7.1 overlaps with the upstream part of pGFIM-B-17.1 and elongates this sequence 5′ towards the *Eco* RI site.

The two clones encode genomic sequences which correspond to FIM-B.1 mRNA sequences analysed previously [5]. Here, the 'short consensus repeat/SCR' and mainly downstrean serine/threonine-rich sequences are encoded. As a hallmark, these sequences are not contiguous at the genomic level but are arranged in at least 11 exons (Fig. 2). Each of these exon sequences is surrounded by consensus intron sequences [20] implicating potential splice junctions. There is only one region hypothetically encoding the tripeptide VPL (positions 3001–3009), which is also highly similar in its flanking sequences to the functional VPS-exon (positions 6510–6518). However, the VPL-sequence has never been recognized at the mRNA level implicating that the splice junctions are probably not functional. The reason is not completely understood; but it is noteworthy that the VPL-encoding potential exon sequence ends with TAA, which is not a preferred proto-splice site [20, 21], whereas the VPS-exon ends with a clearly favoured CAG.

Discussion

The genomic regions defined as exons in Fig. 2 correspond precisely to part of the published mRNA sequence [5]. There are only two point mutations at positions 2540 and 6072, which are changed in the corresponding cDNA sequence (clone pFIM-6.2-15). A comparison of polydisperse cDNA sequences and the genomic sequence presented here clearly indicates that the polydispersities observed at the mRNA level are the result of alternative splicing events between different exons. Consequently, in

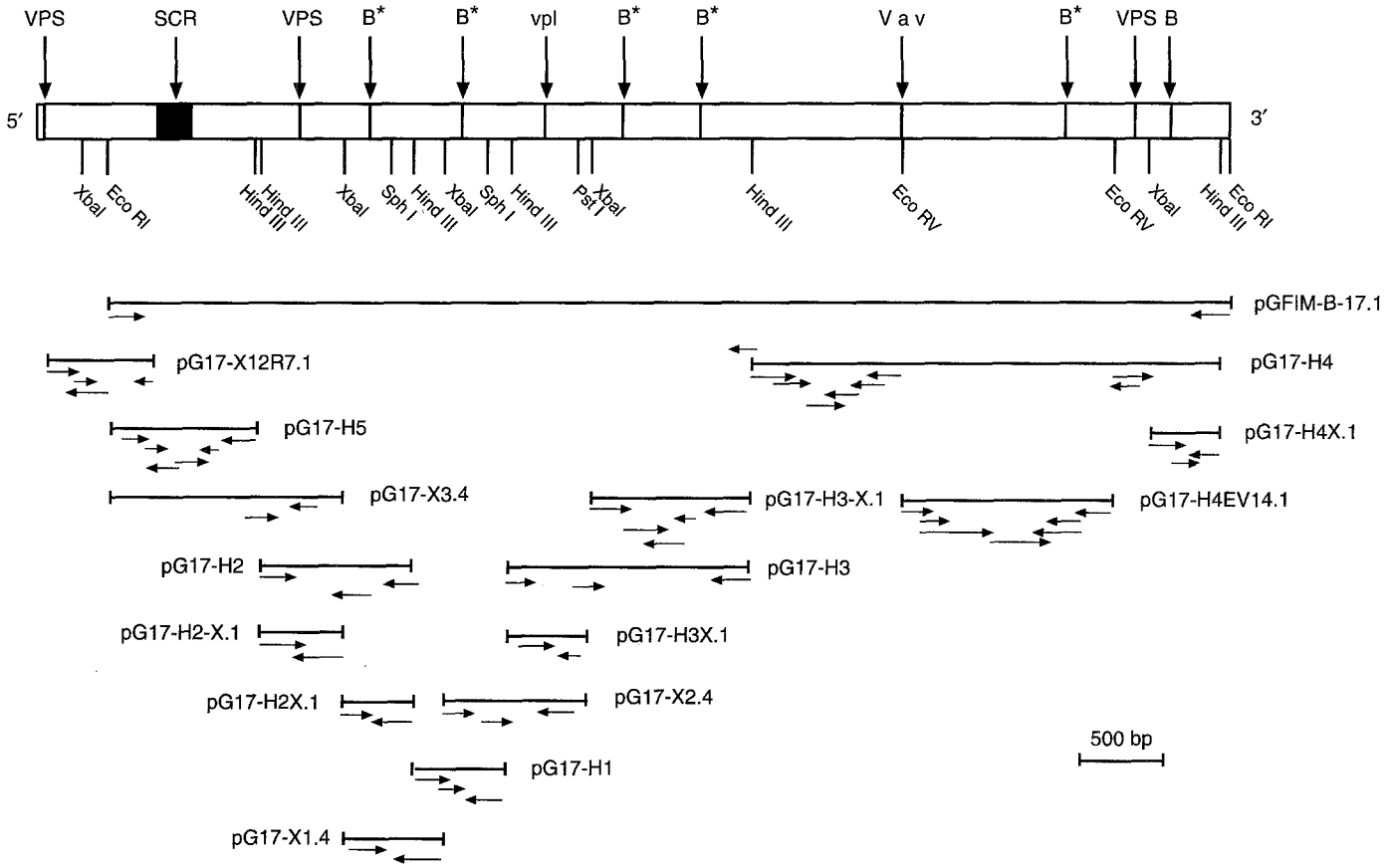

Figure 1. Schematic representation of the analysed portion of the *FIM-B.1* gene. Potential exons (VPS, SCR, B*, B, vpl, Vav) and restriction sites in the 7.0 kb spanning fragment are marked. Also shown are the subclones generated for sequencing. Arrows herein indicate sequenced regions.

order to maintain the reading frame, all exon-intron boundaries are of the same phase creating class 1-1 modules [22]. This feature is typical of shuffled modules particularly found in extracellular mosaic proteins [22]. Thus, the (semi)repetitive region of FIM-B.1 should be an excellent target for exon shuffling; this hypothesis is in agreement with detection of a typical shuffled module in FIM-B.1, i.e. the 'short consensus repeat/SCR' found in many proteins of the complement cascade as well as in certain receptors and cell adhesion molecules [23]. Remarkably, the 3′ splice junctions of the four SCR-encoding exons in decay-accelerating factor [24] are located at the same position as in FIM-B.1. Furthermore, a similar case of exon shuffling has been observed for FIM-C.1, where P-domains have been introduced between tandem repeats [17].

Interestingly, the repetitive regions of the human mucins MUC1 and MUC2 are encoded by an uninterrupted array of individual units not allowing alternative splicing [25, 26]. Also, tandem repeats in polymorphic human glycoprotein Ib, which resemble type B repeats, are not interrupted by introns [27]. Thus, in analogy to the situation of LDL-receptor-related proteins or thrombospondin [21], the interrupted structure of the FIM-B.1 locus consisting of a series of shuffled class 1-1 exons may be the indication for a younger gene than *MUC1* and *MUC2*, where original introns may have been lost during evolution. Alternatively, *MUC* genes may have evolved differently than *FIM-B.1*, e.g. by unequal crossing over.

Introns appeared relatively late during evolution, i.e. a time when multicellularity developed. Interestingly, most extracellular mosaic proteins are associated with multicellularity and enable cell-cell or cell-matrix interactions [21]. Intron sequences may represent relics of the original assembly process. Thus, intron sequences do normally not belong to the highly conserved regions within a gene. In contrast, introns in the *FIM-B.1* gene separating VPS-encoding exons or type B-like repeats show an unusual high degree of similarity (Fig. 3). Only the introns flanking the 'short consensus repeat/SCR' and the VAVTALQISPT-encoding exon represent unique sequences. This may indicate that repeated duplication of genomic DNA including intron sequences occurred quite recently. Alternatively, genetic mechanisms like unequal crossing over or gene conversion could prevent drifting of

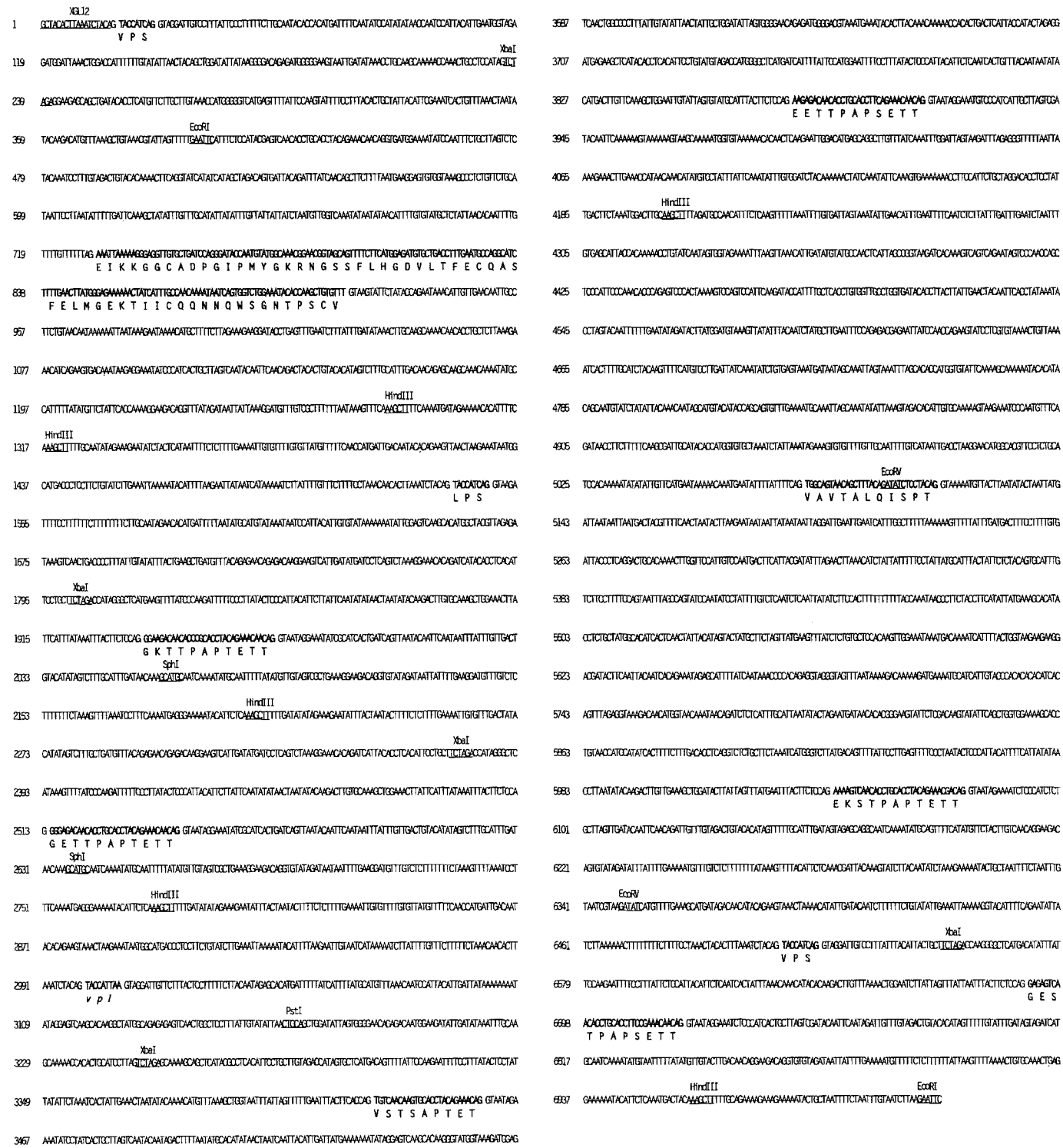

Figure 2. Genomic *FIM-B.1* sequence as obtained from clones pG17-X12R7.1 and pGFIM-B-17.1. Potential exon sequences as well as deduced amino acid sequences are shown in bold type. Also marked are restriction sites. This sequence has been submitted to the EMBL/ GenBank Data Base with the accession number X95549.

intron sequences [28, 29]. In particular, unequal crossing over has also been discussed explaining the sequence homogeneity in *MUC1* [25].

Currently, it is not clear why FIM-B.1 is still in a plastic state allowing size variations of its O-glycosylated portion. This is obviously in contrast to human mucins but has also been found in FIM-C.1 [17]. However, the viscoelastic properties of mucins are highly size dependent [14]. Thus, by alternative splicing the rheological properties of *X. laevis* integumentary mucus may be

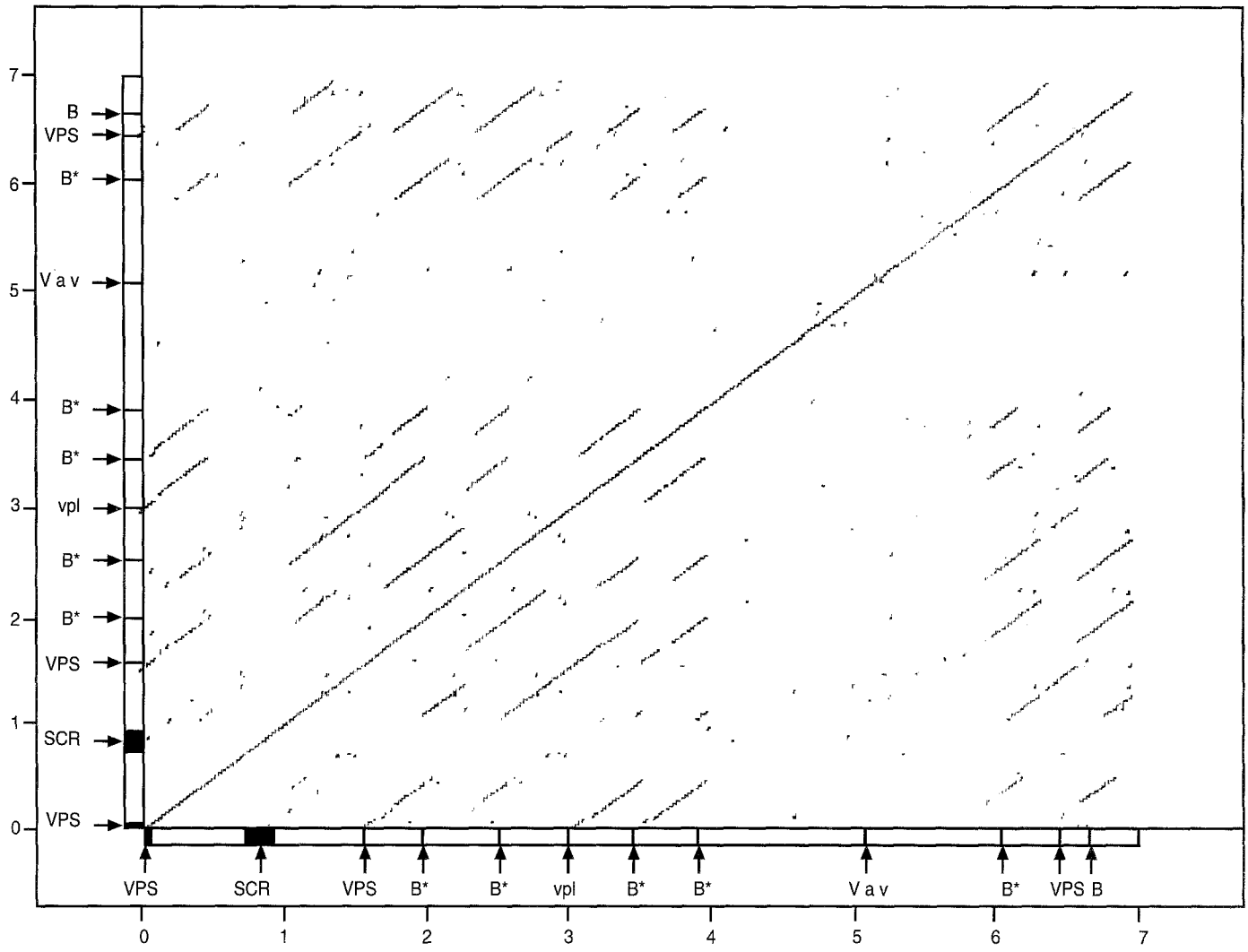

Figure 3. Dot matrix plot analysis [30] of the genomic *FIM-B.1* sequence presented in Fig. 2 versus itself. Shown are internal similarities within the *FIM-B.1* gene. Segments of 15 nucleotides were compared sequentially and a dot was plotted for a match of at least 13 nucleotides.

variable to a certain extent which might be of advantage for the individuals when adapting to varying environmental conditions.

Acknowledgements

We thank C. Roeben for valuable technical assistance, U. Schimanko for oligonucleotide synthesis and the 'Fonds der Chemischen Industrie' for financial support.

References

1. Forstner JF, Forstner GG (1994) In *Physiology of the Gastrointestinal Tract*, Vol. 2, 3rd ed. (Johnson LR, Alpers DH, Christensen J, Jacobson ED, Walsh JH, eds) pp. 1255–83. New York: Raven Press.
2. Hoffmann W, Hauser F (1993) *Comp Biochem Physiol* **105B**: 465–72.
3. Hoffmann W, Joba W (1995) *Biochem Soc Trans* **23**: 805–10.
4. Probst JC, Gertzen E-M, Hoffmann W (1990) *Biochemistry* **29**: 6240–44.
5. Probst JC, Hauser F, Joba W, Hoffmann W (1992) *J Biol Chem* **267**: 6310–16.
6. O'Connell B, Tabak LA, Ramasubbu N (1991) *Biochem Biophys Res Commun* **180**: 1024–30.
7. Lamblin G, Lhermitte M, Klein A, Houdret N, Scharfman A, Ramphal R, Roussel P (1991) *Am Rev Respir Dis* **144**: S19–S24.
8. Pollex-Kruger A, Meyer B, Stuike-Prill R, Sinnwell V, Matta KL, Brockhausen I (1993) *Glycoconj J* **10**: 365–80.
9. Granovsky M, Bielfeldt T, Peters S, Paulsen H, Meldal M, Brockhausen J, Brockhausen I (1994) *Eur J Biochem* **221**: 1039–46.
10. Taylor-Papadimitriou J, Epenetos AA (1994) *Trends Biotech* **12**: 227–33.

11. Lesuffleur T, Zweibaum A, Real FX (1994) *Crit Rev Oncol Hematol* **17**: 153–80.
12. Vavasseur F, Dole K, Yang J, Matta KL, Myerscough N, Corfield A, Paraskeva C, Brockhausen I (1994) *Eur J Biochem* **222**: 415–24.
13. Carlstedt I, Sheehan JK, Corfield AP, Gallagher JT (1985) *Essays Biochem* **20**: 40–76.
14. Jentoft N (1990) *Trends Biochem Sci* **15**; 291–94.
15. Carlstedt I, Sheehan JK (1984) In *Ciba Foundation Symposium 109 on Mucus and Mucosa* (Nugent J, O'Connor M, eds) pp. 157–66. London: Pitman.
16. Swallow DM, Gendler S, Griffiths B, Taylor-Papadimitriou J, Bramwell ME (1987) *Nature* **328**: 82–84.
17. Hauser F, Hoffmann W (1992) *J Biol Chem* **267**: 24620–24.
18. Sambrook J, Fritsch EF, Maniatis T (1989) *Molecular Cloning. A Laboratory Manual*, Vol. 1, 2nd ed. Cold Spring Harbor: Cold Spring Harbor Laboratory Press.
19. Hoffmann W (1988) *J Biol Chem* **263**: 7686–90.
20. Breathnach R, Chambon P (1981) *Annu Rev Biochem* **50**: 349–83.
21. Patthy L (1994) *Curr Opin Struct Biol* **4**: 383–92.
22. Patthy L (1991) *Curr Opin Struct Biol* **1**: 351–61.
23. Bork P (1991) *FEBS Lett* **286**: 47–54.
24. Post TW, Arce MA, Liszewski MK, Thompson ES, Atkinson JP, Lublin DM (1990) *J Immunol* **144**: 740–44.
25. Lancaster CA, Peat N, Duhig T, Wilson D, Taylor-Papadimitriou J, Gendler SJ (1990) *Biochem Biophys Res Commun* **173**: 1019–29.
26. Toribara NW, Gum JR, Culhane PJ, Lagace RE, Hicks JW, Petersen GM, Kim YS (1991) *J Clin Invest* **88**: 1005–13.
27. López JA, Ludwig EH, McCarthy BJ (1992) *J Biol Chem* **267**: 10055–61.
28. Baltimore D (1981) *Cell* **24**: 592–94.
29. Wysocki LJ, Gefter ML (1989) *Annu Rev Biochem* **58**: 509–31.
30. Maizel JV, Lenk RP (1981) *Proc Natl Acad Sci USA* **78**: 7665–69.