Structural analyses of the polymorphic area in type II collagen gene

Miikka Vikkula and Leena Peltonen

Laboratory of Molecular Genetics, National Public Health Institute, Mannerheimintie 166, SF-00300 Helsinki, Finland

Received 21 April 1989

The structure of type II collagen gene is extremely well conserved but contains a cluster of high frequency polymorphisms in a 2.2 kb area. Here we report the nucleotide sequence of this DNA area, essential for the PCR-facilitated RFLP-analyses of this gene. In the structural analyses we found four differences in the deduced human amino acid sequence when compared to the published bovine amino acid sequence. The donor and acceptor signals and branch point signals required for the splicing events were in agreement with mammalian consensus sequences. The frequency of inverted repeats which could provoke the DNA strand to loop formation and consequently to deletion mutations did not differ from that found in other sequenced genes coding for fibrillar collagens.

DNA, recombinant; Collagen, α1(II); Polymorphic cluster; Splicing signal; Inverted repeat

1. INTRODUCTION

The studies on the polymorphisms of type II collagen gene have revealed that high-frequency polymorphisms are situated close to each other, separated only by about 2200 bp [1]. This polymorphic area of the well conserved type II collagen gene is located in the middle of the coding area for the triple helical part of the molecule and serves as a good candidate for a mutation sensitive area of this gene. We sequenced this area for two reasons: firstly to provide information for scientists interested in the non-radioactive detection of the polymorphisms of type II collagen gene using the polymerase chain reaction and secondly to search for the possible structural elements that could make the DNA strand more vulnerable to genetic rearrangements.

Correspondence address: M. Vikkula, Laboratory of Molecular Genetics, National Public Health Institute, Mannerheimintie 166, SF-00300 Helsinki, Finland

Abbreviations: bp, basepair; RFLP, restriction fragment length polymorphism

2. MATERIALS AND METHODS

We constructed two subclones from the full-length genomic clone of type II collagen gene [2] into pGem 3 blue vector (Promega). The sequencing was performed on two strands by Sanger's dideoxynucleotide method using T7 DNA polymerase (United States Biochemical) (see fig.1) and the DNA sequences were analysed with the PC GENE program (GENOFIT).

3. RESULTS AND DISCUSSION

3.1. *Exon sequences*

Collagen gene sequences demonstrate extremely high evolutionary conservation from species to species. Especially the gene regions coding for the triple helical parts of the collagens have obviously been exposed to an evolutionary pressure preserving the pattern of the exonic domains [3]. E.g. the sequences of human, bovine and even chicken type I collagen genes demonstrate a high degree of homology at the amino acid level [4]. The well conserved type II collagen gene contains, however, a 2.2 kb area with two high frequency polymorphisms [1]. For the sequence analyses of this area, the *Pvu*II (polymorphic)-*Pvu*II and *Kpn*I-*Bam*HI fragments were subcloned (see fig.1). This subcloning strategy was based on the published

Published by Elsevier Science Publishers B.V. (Biomedical Division)
00145793/89/$3.50 © 1989 Federation of European Biochemical Societies

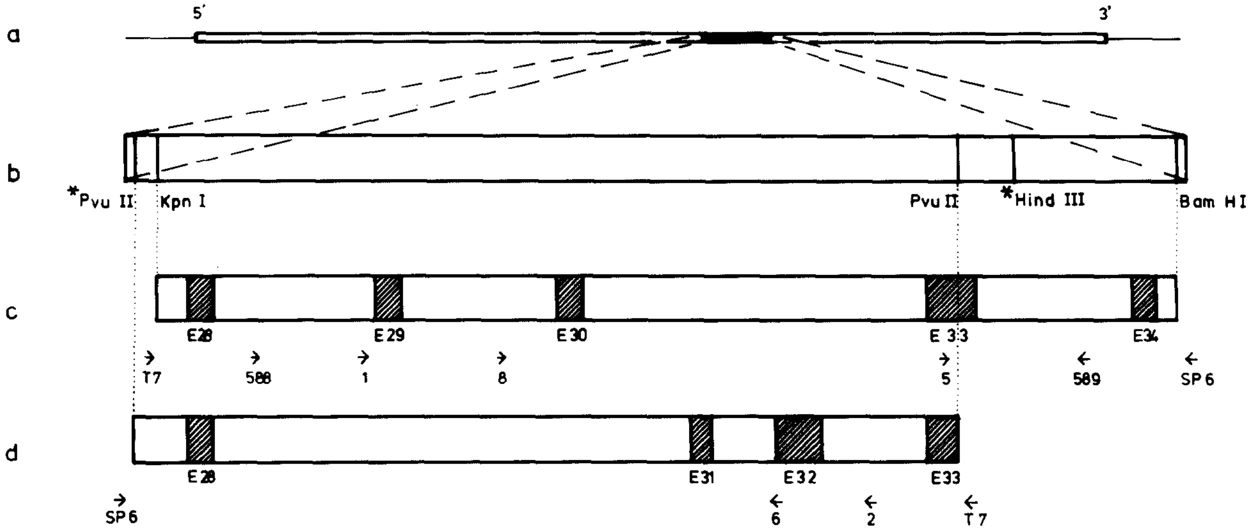

Fig.1. (a) Collagen α1(II) gene, (b) the sequenced area, (c) the *Kpn*I-*Bam*HI subclone with the exons found, (d) the *Pvu*II-*Pvu*II subclone with the exons found. The primers with their names and the primary sequencing directions are indicated by arrows. The polymorphic restriction sites are marked with an asterisk.

restriction maps locating the polymorphic *Pvu*II and constant *Kpn*I sites which are distinctly apart from each other [5,6]. However, we established the location of these sites to be only 9 bp apart. The sequenced individual did not have the polymorphic *Hin*dIII site but with a computerized search allowing one point mutation in the restriction site we found four potential *Hin*dIII sites within 120 bp and they are given in fig.2.

The sequenced DNA strand accommodated seven exons, numbers 28–34, coding for 156 amino acids of the triple helical region of type II collagen. The detected high frequency of exons is typical to triple helical regions also in other, better known collagen genes [7]. The predicted amino acid sequence of human exons differed from the reported bovine amino acid sequence [8] by four residues (fig.2). Two of these variations occurred at X and two at the Y positions in the Gly-X-Y triplet essential for triple helix formation.

3.2. *Exon/intron boundaries and splicing signals*

When the intron-exon ratio was compared to the published values for chicken type II collagen, the ratio of 4.7 versus that of 2.4 demonstrates the larger average size of human introns in the triple helical region leading to the larger size of the human gene [2,9]. The sequences for donor and acceptor splicing signals in the sequenced introns (table 1) were in good agreement with the generalized mammalian splicing signals [10,11]. The branch point signals, considered essential for the lariat formation in splicing, are obligatorily located about 18–37 bp from the 3′-end of each intron [10,12]. We found these signals in all sequenced introns less than 31 nucleotides from the 3′-end and intron 32 had this signal only 12 bp from the 3′-end. Intron 29 contained two potential branch point signals, 31 and 21 bp from the 3′-end (fig.2). All the branch point signals were in agreement with the general consensus sequence reported for mammalian introns: C/T T N A Pu/T [10–12].

3.3. *Structural features providing a hypothetical predisposition to rearrangements*

In the well conserved collagen genes, some mutations lead to various inherited connective tissue diseases of man [13], and two human diseases have so far been linked to type II collagen gene: Stickler's syndrome [14] and a familial form of osteoarthrosis [15]. The more detailed studies of mutations of this particular gene are severely hampered by the practical unavailability of mRNA and protein from cartilage tissues of diseased individuals, because the levels of type II collagen mRNA are very low or non-existent in the mature

```
 PvuII  KpnI
cagctggggtaccgtggaggtctggaaactctgggccagaagtacctttgcccaatccta
                                                          60

gggggactgcaagtgggaagaaaagcgtgtcattgactttctttttctcttctgtctag
                                                         120
G  L  P  G  K  D  G  E  T  G  A  A  G  P  P  G  P  A
GGTCTTCCTGGCAAAGATGGTGAGACAGGTGCTGCAGGACCCCCTGGCCCTGCTgtaagt
E28                                                      180

acctgcccagcctccccaggtggccctgggcaggggctgggaggggtgggggtgggagag
                                                         240

cccatccattaatggagctgacagatgtgaatgtgggctgagctgatacaccagactcac
                                                         300

tctgagctgaggcagggtgtcccaggaggctgtgtggacccacattggtggagaggagtg
                                                         360
                 1
tgggtggctgatgggagtgcagggaggcatgcatgcactgtctgagtggtgcaggaagac
                                                         420

gcctgtgctgcccaccctgctgacctcccctgggccctactagctgtggctctcagggtc
                                                         480
                                          G  P  A  G  E  R
tctggaacactggctcagctcaccttttcttttacactgcagGGACCTGCTGGTGAACGA
                                          E29            540
G  E  Q  G  A  P  G  P  S  G  F  Q
GGCGAGCAGGGTGCTCCTGGGCCATCTGGGTTCCAGgtaggtggctggaccaggctctct
                                                         600

gtgtcagtcctttgccatacccagggctccctggacagcagcaggcactatcggtggagg
                                                         660

gcccacacctcttgcagtgtccaggcatcgagccttccctgcacccctggctgtcactgc
                                                         720
                                         1
tgctgcttcctttctttgggtctgccctatactgtgcctccctggggccagggcagcaa
                                                         780
                                       G  L  P  G  P  P  G
actcactcctttgctaacgcttgtcacttcggcttctagGGACTTCCTGGCCCTCCTGGT
             (P)                       E30               840
P  P  G  E  G  G  K  P  G  D  Q
CCCCCAGGTGAAGGTGGAAAACCAGGTGACCAGgtgagtatggggctctttggacctgca
                                                         900

acctgtttagatgggaaggtcttttctgatgcctaggaggcaagggcaagagggcatgag
                                                         960

gagcctgtgaggcctgggaatgtctggacccatgtcccagcctccacagatgacacaatc
                                                        1020

ccatggaggagtgatattcagccctgctgtggagaattgttcaggggtctgtgatatgaa
                                                        1080
```

```
                                   G  V  P  G  E  A  G  A  P
gccttcactctcacacatctttctttgttctccagGGTGTTCCCGGTGAAGCTGGAGCCC
                                   E31                  1140
 G  L  V  G  P  R
CTGGCCTCGTGGGTCCCAGGgtgagtatcctgttggccaatgcgggctgcctccttgggc
                                                        1200

ctgccctgggtcctatgctcctgctcctttccccacctccctgcttctccctggaccttc
                                                        1260
                                              G  E  R  G  F
tcccccactgctgttggttgatcacttcttggtgtctctgccgcagGGTGAACGAGGTTT
                        (S)                (A) E32      1320
 P  G  E  R  G  S  P  G  A  Q  G  L  Q  G  P  R  G  L  P  G
CCCAGGTGAACGTGGCTCTCCCGGTGCCCAGGGCCTCCAGGGTCCCCGTGGCCTCCCCGG
                                                        1380
 T  P  G  T  D  G  P  K
CACTCCTGGCACTGATGGTCCCAAAgtaagtgaggctgcatccagtaggggtcttcgtgg
                                                        1440

tagcctggagtcccactgagcaggagagaggagcgggctcaggaggaatgaagaacagaa
                                                        1500
            2
gtgggggagctggaaaggaggtctacatgggaggaagggaaggaagaggggtttggggc
                                                        1560

ctggttacccaggctccatgaacatgggttcagggagaggtgctgtccactacagactcc
                        (A)                             1620
                   G  A  S  G  P  A  G  P  P  G  A  Q  G  P
ctcttacctccctccccagGGTGCATCTGGCCCAGCAGGCCCCCCTGGGGCTCAGGCCCC
                   E33                                  1780
 P  G  L  Q  G  M  P  G  E  R  G  A  A  G  I  A  G  P  K  G
TCCAGGTCTTCAGGGAATGCCTGGCGAGAGGGGAGCAGCTGGTATCGCTGGGCCCAAAGG
                                                        1740
 D  R                                 *HIII
GGACAGGgtaagtactgaggttacagcctcctcaccaaagctgtggctttgccaatgtcc
                                                        1800
                                               *HIII
tgccccttgtgatcgcttccgttcccttatggcacctggtgatgaaggtttctgttagcc
                                                        1860
         *HIII               2 *HIII              3
ctttttgaggagcttaaagactcctttccaaagctccctgcctttagtgacatcctttc
                                                        1920

gcctgttccttcatctcaccccctgctgctcctcacccaccctgagaccacagcaaattcc
                                                        1980

tcttgggcagggactgggctttccctagcaccccagcctgggtggactgagcaaaccat
                                                        2040
                                              G  D  V  G
ggggtcctggggtgcctggctgaggcggctggtttctcttccctcagGGTGACGTTGG
                                   3          E34       2100
 E  K  G  P  E  G  A  P  G  K  D  G  G  R
TGAGAAAGGCCCTGAGGGAGCCCCTGGAAAGGATGGTGGACGAgtaagtgaatgcgggct
                                                        2160
                     BamHI
gctggactgctgggcattaggatcc
                     2185
```

Fig.2. The genomic DNA sequence of collagen α1(II) from the 5′-polymorphic *Pvu*II site to a 3′-*Bam*HI site. The boxed sequences marked 1 to 3 are the three inverted repeats and the underlined sequences are the branch point signals, intron 29 having two of them (see section 3). Above the numbered exonic sequences (capital letters) are marked the corresponding amino acids. The four bovine amino acids differing from this human amino acid sequence are marked above. The four possible *Hin*dIII sites are indicated by *HIII.

cartilage and it is practically impossible to obtain enough protein from living individuals. This forces the research to the DNA level. Potential candidate areas for rearrangements in large, well conserved collagen genes are regions demonstrating variations between individuals. In type II collagen gene the known high frequency polymorphisms are located on the 2.2 kb area close to the middle of the gene [1]. After determining the nucleotide sequence of this area we performed the computerized search for structural features which could make this particular area more susceptible to rearrangements e.g. during meiotic divisions. We found 9 inverted repeats of 10 or more nucleotides and three of them were located 200–500 bp apart, the distance favoring small deletions during meiotic divisions (see fig.2). The distance between the repeats and the start of the exons varied from 65 to 209 bp. The inverted repeats did not carry a high degree of internal homology nor significant homology to inverted repeats we have found in other sequenced collagen genes [6,8]. The majority of the repeats represented GC-rich DNA areas, these nucleotides counting for over 70% of the nucleotide sequences.

In the previously published genomic sequence from the triple helical area of α1(I) collagen containing 4300 bp and 19 exons, we found four inverted repeats predisposed to small deletions [6]. The analyzed polymorphic area of the human type II collagen gene thus only contains a slightly higher frequency of such inverted repeats in proportion to

Table 1

(A) 5′ splice junction

	A	G	:	G	T	A	A	G	T
A	3	2	:	0	0	5	6	0	0
C	1	0	:	0	0	0	0	0	0
G	3	4	:	7	0	2	1	7	0
T	0	1	:	0	7	0	0	0	7

(B) 3′ splice junction

	← Py × 10 →										C	C	A	G:G	
A	2	0	0	0	0	1	0	1	0	0	0	0	7	0	0
C	1	4	2	2	3	2	3	2	5	3	4	5	0	0	0
G	0	1	0	0	0	2	1	1	1	0	2	0	0	7	7
T	4	2	5	5	4	2	3	3	1	4	1	2	0	0	0

The different nucleotides found in specific positions in the 5′ and the 3′ splice junctions. The consensus sequences are marked under the line

the length of the sequence. Further, we found five such inverted repeats in the published 2450 bp long sequence including 8 exons sequenced from the human gene coding for the triple helical part of type II collagen [8]. Consequently, these comparisons suggest that this polymorphic DNA area is structurally not exceptionally vulnerable to meiotic disturbances.

Acknowledgements: The authors are grateful to Ms Ritva Timonen for her intelligent help in sequencing and to Ms Helena Rantanen for typing the manuscript and Dr Taina Pihlajaniemi for valuable comments. This work was supported by grants from Paulo Foundation, Finnish Medical Association Duodecim and Sigrid Juselius Foundation.

REFERENCES

[1] Väisänen, P., Elima, K., Palotie, A., Peltonen, L. and Vuorio, E. (1988) Human Heredity 38, 65–71.

[2] Cheah, K.S.E., Stoker, N.G., Griffin, J.R., Grosveld, F.G. and Solomon, E. (1985) Proc. Natl. Acad. Sci. USA 82, 2555–2559.

[3] Yamada, Y., Avvedimento, V.E., Mudryj, M., Ohkubo, H., Vogeli, G., Irani, M., Pastar, I. and De Crombrugghe, B. (1980) Cell 22, 887–892.

[4] Chu, M.L., De Wet, W., Bernard, M., Ding, I.F., Morabito, M., Myers, I., Williams, C. and Ramirez, F. (1984) Nature 310, 337–340.

[5] Sangiorgi, O.F., Benson-Chanda, V., De Wet, W.I., Sobel, M.E., Tsipuras, P. and Ramirez, F. (1985) Nucleic Acids Res. 13, 2207–2225.

[6] Sykes, B., Smith, R., Vipond, S., Paterson, C., Cheah, K. and Solomon, E. (1985) Med. Gen. 22, 187–191.

[7] D'Alessio, M., Bernard, M., Pretorious, P.J., De Wet, W. and Ramirez, F. (1988) Gene 67, 105–115.

[8] Piez, K. (1976) in: Biochemistry of Collagen (Ramachandran, G.N. and Reddi, A.H. eds) pp.1–44, Plenum, New York.

[9] Upholt, W.B. and Sandell, L.J. (1986) Proc. Natl. Acad. Sci. USA 83, 2325–2329.

[10] Green, M.R. (1986) Annu. Rev. Genet. 20, 671–708.

[11] Keller, E.B. and Noon, W.A. (1984) Proc. Natl. Acad. Sci. USA 81, 7417–7420.

[12] Krainer, A.R. and Maniatis, T. (1988) in: Transcription and Splicing (Hames, B.D. and Glover, D.M. eds) pp.131–172, IRL Press, Oxford.

[13] Prockop, D.J. and Kivirikko, K.I. (1984) New England J. Med. 311, 376–386.

[14] Francomano, C.A., Liberfarb, R.M., Hirose, T., Maumenee, I.H., Streeten, E.A., Meyers, D.A. and Pyeritz, R.E. (1987) Genomics 1, 293–296.

[15] Palotie, A., Väisänen, P., Ott, J., Ryhänen, L., Elima, K., Vikkula, M., Cheah, K., Vuorio, E. and Peltonen, L. (1989) Predisposition to familial osteoarthrosis is linked to type II collagen gene, Lancet, in press.