Volume 244, number 2, 296–300 FEB 06821 February 1989

Primary structure of human proacrosin deduced from its cDNA sequence

Tadashi Baba, Ken Watanabe, Shin-ichi Kashiwabara and Yuji Arai

Institute of Applied Biochemistry, University of Tsukuba, Ibaraki 305, Japan

Received 26 October 1988; revised version received 23 December 1988

cDNA clones encoding proacrosin, the zymogen of acrosin, were isolated from a human testis cDNA library by using a fragment of boar acrosin cDNA as a probe. Nucleotide sequencing of the longest cDNA clone has predicted that human proacrosin is synthesized with a 19-amino-acid signal peptide at the N-terminus. The cleavable signal sequence is followed by a 23-residue segment corresponding to the light chain and then by a 379-residue stretch that constitutes the heavy chain containing the catalytic site of the mature protease. The C-terminal portion of the deduced sequence for the heavy chain is very rich in proline residues, most of which are encoded by a unique repeat of CCCCCA. The active-site residues including histidine, aspartic acid, and serine are also predicted to be located at residues 69, 123, and 221, respectively.

Acrosin; Serine protease; Amino acid sequence; cDNA; (Human testis)

1. INTRODUCTION

Acrosin is an endoprotease occurring in the acrosomes that cover the anterior part of the sperm head [1]. On the basis of the substrate specificity [2,3], inhibition profile [3], and N-terminal amino acid sequence [4,5], acrosin is thought to belong to the family of serine proteases. Like other members of this family, acrosin is synthesized as an enzymatically inactive precursor, proacrosin, and converted to the mature form by limited proteolysis [1].

Most of the biochemical data on (pro)acrosin have been obtained by using boar sperms [6,7]. Boar acrosin and its precursor (zymogen) have been reported to possess apparent molecular masses of 34–40 and 53–55 kDa, respectively [5–9]. The mature enzyme is a two-chain protein consisting of a 23-residue light chain and a heavy chain containing the functionally active site, the two chains being covalently linked by two disulfide bridges [5]. On the other hand, our recent results indicate that the zymogen is a single-chain polypeptide containing a segment corresponding to the light chain at its N-terminus [10]. It is, therefore, conceivable that boar proacrosin is converted to the mature enzyme by two activation processes, i.e. the liberation of proenzyme segments from the C-terminus and cleavage of a peptide bond producing the light and heavy chains. In contrast to the case for the boar (pro)enzyme, biochemical information concerning human (pro)acrosin is still very limited. For instance, considerably variable values have been reported for the molecular masses of human proacrosin and acrosin [11–18].

To facilitate further studies of human (pro)acrosin, it is necessary to establish the structures of its protein and gene. Here, we report the isolation of proacrosin cDNA clones from a human testis cDNA library. The primary structure of human (pro)acrosin deduced from the nucleotide sequence of the cloned cDNA is also reported.

Correspondence address: T. Baba, Institute of Applied Biochemistry, University of Tsukuba, Tsukuba, Ibaraki 305, Japan

The nucleotide sequence presented here has been submitted to the EMBL/GenBank database under the accession no. Y00970

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

2. EXPERIMENTAL

A cDNA library in λgt11 consisting of 1.4×10^6 independent clones (average insert size, 1.2 kbp) was prepared from boar testis poly(A)$^+$ RNA [19]. Approx. 7×10^5 phages were screened with polyclonal rabbit anti-boar acrosin antibodies that had been purified by chromatography on an acrosin-conjugated Sepharose 4B column. Horseradish peroxidase-conjugated goat anti-rabbit IgG antibodies (Jackson Immunoresearch Labs) were used to detect antibody-binding clones. Positive clones were selected and plaque-purified. The DNA was isolated and digested with *Eco*RI, and the cDNA inserts were subcloned into pUC19 for restriction mapping and nucleotide sequencing. Another λgt11 library from human testis (Clontech) were screened by the plaque hybridization method [20] using a cDNA fragment for boar acrosin as a probe. The probe was labelled by the random-primer technique using a DNA labelling kit (Nippon Gene) and [α-^{32}P]dCTP (6000 Ci/mmol, Amersham) according to the supplier's protocol. Positive clones were plaque-purified, and the insert DNA was isolated and digested with restriction enzymes. The DNA fragments were subcloned into M13mp18 or 19, or pUC19 at appropriate restriction sites. The nucleotide sequences of the subcloned inserts were determined by the dideoxy chain-termination method [21] using an M13 sequencing kit (Toyobo), a 7-deaza sequencing kit (Takara), or a Sequenase™ (US Biochemical Corp.).

3. RESULTS

Screening of the boar testis cDNA library with affinity-purified antibodies to boar acrosin yielded four immunopositive clones, λPA1, λPA2, λPA3, and λPA4, having cDNA inserts of 1.7, 1.4, 0.6, and 0.7 kbp, respectively. The λPA2 insert had an internal *Eco*RI site and formed two fragments 695 (PA3) and 646 bp long (PA4). Sequence analysis of PA4 indicated that it contained an open reading frame encoding 215 amino acid residues including the N-terminal 75 residues previously determined for purified boar (pro)acrosin [4,5,10]. We thus concluded that PA4 was an authentic fragment of boar acrosin cDNA (details will be reported elsewhere).

When 8×10^4 λgt11 recombinant phages from a human testis cDNA library were screened using PA4 as a probe, three positive clones, termed H4, H8, and H9, were obtained. Since the inserts of these clones had similar restriction maps, the nucleotide sequence of the insert of the longest clone, H4, was determined according to the strategy shown in fig.1. Partial DNA sequences of the H8 and H9 inserts showed that they were related to the H4 insert (not shown). The H4 cDNA was 1388 bp in length and contained an open reading frame of 1263 nucleotides (421 amino acids) starting from an ATG initiation codon at nucleotides 17–19 (fig.2). The open reading frame was followed by a 3′-noncoding region 94 nucleotides long. Three overlapping polyadenylation signals (AATAAATAAATAAA) were located 17 nucleotides upstream from the poly(A)$^+$ addition site.

The deduced amino acid sequence of human proacrosin along with the N-terminal 75-residue sequence covering the light chain and a part of the heavy chain of boar (pro)acrosin [4,5,10] is shown in fig.2. In this 75-residue region, the sequence similarity between human and boar proacrosins was almost 75%. The high similarity indicated that the H4 insert was actually a cDNA encoding human proacrosin consisting of the light and heavy chains (residues 1–23 and 24–402, respectively). The amino acid sequence at residues −19 to −1 was strongly hydrophobic. It seemed to be a cleavable signal peptide required for translocation of the newly synthesized protein across a membrane. Thus, the cDNA-derived sequence indicated that human proacrosin contained 402 amino acids with a molecular mass of 43 860 Da. Two potential *N*-linked glycosylation sites were found at residues 3 and 191. There was a proline-rich region at the C-terminal portion of the protein (residues 283–353). Moreover, the prolines at residues 325–353 were mostly encoded by a unique repeat of CCCCCA. The amino acid composition of human proacrosin predicted from the cDNA-derived sequence is shown in table 1 together with that of boar proacrosin determined by us [10]. Their compositions were closely similar to each other, and a high content of proline (about 15 mol%) was found in both proacrosins.

The heavy chain of human (pro)acrosin possessed an N-terminal sequence, Ile-Val-Gly-Gly (residues 24–27), which was indicative of an activated serine protease. When the entire sequence of the heavy chain was compared with those of other human serine proteases, significant sequence similarities were found only between residues 24 and 270–290 (fig.3). The percent identity of human (pro)acrosin with human pancreatic trypsin [22] was 33% in the region of residues 24–274, with human pancreatic kallikrein [23] 33% in the region of residues 24–274, with human plasmin B-

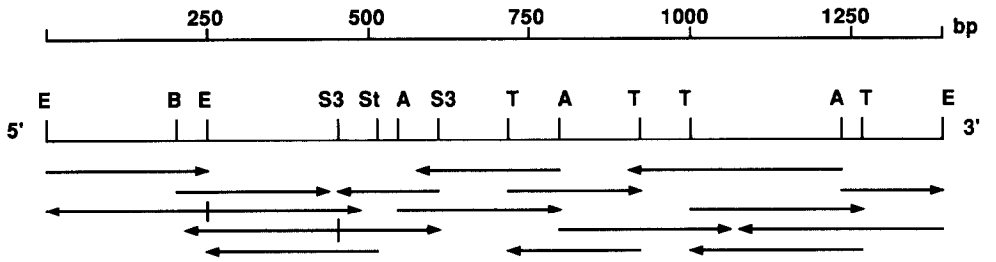

Fig.1. Restriction map and sequencing strategy for the cDNA insert of human proacrosin clone H4. The restriction sites (E, *Eco*RI; B, *Bgl*II; R, *Rsa*I; A, *Alu*I; T, *Tth*HB8I; St, *Stu*I; S3, *Sau*3AI) were used for subcloning into M13mp18 or 19, or pUC19. Arrows indicate the direction and extent of each sequence determination. bp, base pair.

```
                                                CAGGCA GTGCAGGAGT ATG GTT GAG ATG CTA CCA ACT GCC ATT CTG CTG GTC TTG GCA GTG TCC GTG GTT GCT   73
                                                      Met Val Glu Met Leu Pro Thr Ala Ile Leu Leu Val Leu Ala Val Ser Val Val Ala   -1

        AAA GAT AAC GCC ACG TGT GAT GGC CCC TGT GGG TTA CGG TTC AGG CAA AAC CCA CAG GGT GGT GTC CGC ATC GTC GGC GGG AAG GCT GCA  163
Human   Lys Asp Asn Ala Thr Cys Asp Gly Pro Cys Gly Leu Arg Phe Arg Gln Asn Pro Gln Gly Gly Val Arg Ile Val Gly Gly Lys Ala Ala   30
Boar    Arg Asp Asn Ala Thr Cys Asp Gly Pro Cys Gly Leu Arg Phe Arg Gln Lys Leu Glu Ser Gly Met Arg Val Val Gly Gly Met Ser Ala

        CAG CAT GGG GCC TGG CCC TGG ATG GTC AGC CTC CAG ATC TTC ACG TAC --- AAC AGC CAC AGG TAC CAC ACA TGT GGA GGC AGC TTG CTG  250
Human   Gln His Gly Ala Trp Pro Trp Met Val Ser Leu Gln Ile Phe Thr Tyr --- Asn Ser His Arg Tyr His Thr Cys Gly Gly Ser Leu Leu   59
Boar    Glu Pro Gly Ala Trp Pro Trp Met Val Ser Leu Gln Ile Phe Met Tyr His Asn Asn Arg Arg Tyr His Thr Cys Gly Gly Ile Leu Leu

        AAT TCA CGA TGG GTG CTC ACT GCT GCT CAC TGC TTC GTC GGC AAA AAT AAT GTG CAT GAC TGG AGA CTG GTT TTC GGA GCA AAG GAA ATT  340
Human   Asn Ser Arg Trp Val Leu Thr Ala Ala His Cys Phe Val Gly Lys Asn Asn Val His Asp Trp Arg Leu Val Phe Gly Ala Lys Glu Ile   89
Boar    Asn Ser His Trp Val Leu Thr Ala Ala His Cys Phe Lys Asn Lys

        ACA TAT GGG AAC AAT AAA CCA GTA AAG GCG CCT GTG CAA GAG AGA TAT GTG GAG AAA ATC ATC ATT CAT GAA AAA TAC AAC TCT GCG ACA  430
        Thr Tyr Gly Asn Asn Lys Pro Val Lys Ala Pro Val Gln Glu Arg Tyr Val Glu Lys Ile Ile Ile His Glu Lys Tyr Asn Ser Ala Thr  119

        GAG GGA AAT GAC ATT GCC CTC GTG GAG ATC ACC CCT CCC ATT TCG TGT GGG CGC TTC ATT GGG CCG GGC TGC CTG CCC CAC TTG AAG GCA  520
        Glu Gly Asn Asp Ile Ala Leu Val Glu Ile Thr Pro Pro Ile Ser Cys Gly Arg Phe Ile Gly Pro Gly Cys Leu Pro His Leu Lys Ala  149

        GGC CTC CCC AGA GGC TCC CAG AGC TGC TGG GTG GCC GGC TGG GGA TAT ATA GAA GAG AAA GCC CCC AGG CCA TCA TCT ATA CTG ATG GAG  610
        Gly Leu Pro Arg Gly Ser Gln Ser Cys Trp Val Ala Gly Trp Gly Tyr Ile Glu Glu Lys Ala Pro Arg Pro Ser Ser Ile Leu Met Glu  179

        GCA CGT GTG GAT CTC ATC GAC CTG GAC TTG TGT AAC TCG ACC CAG TGG TAC AAT GGG CGC GTT CAG CCA ACC AAT GTG TGC GCG GGG TAT  700
        Ala Arg Val Asp Leu Ile Asp Leu Asp Leu Cys Asn Ser Thr Gln Trp Tyr Asn Gly Arg Val Gln Pro Thr Asn Val Cys Ala Gly Tyr  209

        CCT GTA GGC AAG ATC GAC ACC TGC CAG GGA GAC AGC GGC GGG CCT CTC ATG TGC AAA GAC AGC AAG GAA AGC GCC TAT GTG GTC GTG GGA  790
        Pro Val Gly Lys Ile Asp Thr Cys Gln Gly Asp Ser Gly Gly Pro Leu Met Cys Lys Asp Ser Lys Glu Ser Ala Tyr Val Val Val Gly  239

        ATC ACA AGC TGG GGG GTA GGC TGT GCC CGT GCC AAG CGC CCC GGA ATC TAC ACG GCC ACC TGG CCT TAT CTG AAC TGG ATC GCC TCC AAG  880
        Ile Thr Ser Trp Gly Val Gly Cys Ala Arg Ala Lys Arg Pro Gly Ile Tyr Thr Ala Thr Trp Pro Tyr Leu Asn Trp Ile Ala Ser Lys  269

        ATT GGT TCT AAC GCT TTG CGT ATG ATT CAA TCG GCC ACC CCT CCA CCG CCC ACC ACT CGA CCG CCC CCG ATT CGA CCC CCC TTC TCC CAC  970
        Ile Gly Ser Asn Ala Leu Arg Met Ile Gln Ser Ala Thr Pro Pro Pro Pro Thr Thr Arg Pro Pro Pro Ile Arg Pro Pro Phe Ser His  299

        CCT ATC TCT GCT CAC CTT CCT TGG TAT TTC CAA CCG CCC CCT CGA CCA CTT CCA CCC CGA CCA CCG GCA GCC CAG CCC CGA CCC CCA CCT 1060
        Pro Ile Ser Ala His Leu Pro Trp Tyr Phe Gln Pro Pro Pro Arg Pro Leu Pro Pro Arg Pro Pro Ala Ala Gln Pro Arg Pro Pro Pro  329

        TCA CCC CCG CCC CCA CCC CCA CCT CCA GCC TCA CCT TTA CCC CCA CCC CCA CCC CCA CCC CCA CCT ACA CCC TCA TCT ACC ACA AAA CTT 1150
        Ser Pro Pro Pro Pro Pro Pro Pro Pro Ala Ser Pro Leu Pro Pro Pro Pro Pro Pro Pro Pro Pro Thr Pro Ser Ser Thr Thr Lys Leu  359

        CCC CAA GGA CTT TCT TTT GCC AAG CGC CTA CAG CAG CTC ATA GAG GTC TTG AAG GGG AAG ACC TAT TCC GAC GGA AAG AAC CAT TAT GAC 1240
        Pro Gln Gly Leu Ser Phe Ala Lys Arg Leu Gln Gln Leu Ile Glu Val Leu Lys Gly Lys Thr Tyr Ser Asp Gly Lys Asn His Tyr Asp  389

        ATG GAG ACC ACA GAG CTC CCA GAA CTG ACC TCG ACC TCC TGA TCTGACCTGG TTCTCAACAG ACCCAGTGAG CCCTTCACTC CTGAGAAAAA      1332
        Met Glu Thr Thr Glu Leu Pro Glu Leu Thr Ser Thr Ser                                                              402

        GGAAAGATGA AATAAATAAA TAAACATATA TATATAGATA TAAAAAAAAA AAAAAA                                                    1388
```

Fig.2. Nucleotide sequence and deduced amino acid sequence of human proacrosin. The predicted amino acid sequence is shown below the nucleotide sequence. The nucleotides are numbered in the 5'- to 3'-direction. The amino acids are numbered from the N-terminus of the predicted light chain sequence, and the residues from the N-terminal side of residue 1 are indicated by negative numbers. The N-terminal sequences of the light (wavy underline) and heavy chains (broken underline) of boar (pro)acrosin determined by the protein analysis [4,5,10] are also indicated. Potential *N*-linked glycosylation sites are marked with asterisks. A series of polyadenylation signal is underlined.

Table 1

Amino acid composition of boar and human proacrosins

Amino acid	Residues/molecule (mol%)	
	Boar proacrosin	Human proacrosin
Asx	24 (5.7)	28 (7.0)
Thr	23 (5.5)	24 (6.0)
Ser	24 (5.7)	28 (7.0)
Glx	43 (10.3)	29 (7.2)
Pro	64 (15.3)	61 (15.2)
Gly	40 (9.6)	34 (8.5)
Ala	28 (6.7)	27 (6.7)
Val	26 (6.2)	20 (5.0)
Cys	12 (2.9)	12 (3.0)
Met	6 (1.4)	5 (1.2)
Ile	19 (4.5)	22 (5.5)
Leu	28 (6.7)	28 (7.0)
Tyr	12 (2.9)	14 (3.5)
Phe	12 (2.9)	8 (2.0)
Lys	19 (4.5)	20 (5.0)
His	5 (1.2)	10 (2.5)
Arg	28 (6.7)	21 (5.2)
Trp	5[a] (1.2)	11 (2.7)
Total	418	402

[a] Data from [5]

The amino acid composition of boar proacrosin was determined by acid hydrolysis [10], and the nearest integers are indicated

chain [24] 35% in the region of residues 24–273, and with human factor X [25] 33% in the region of residues 24–292. A comparison of the amino acid sequence of the C-terminal portion of human proacrosin including the proline-rich segment with the sequences listed in the National Biochemical Research Foundation protein library did not establish any significant sequence similarity with other known proteins.

4. DISCUSSION

Here, we have isolated a cDNA clone encoding a mammalian (pro)acrosin for the first time. The cloned human proacrosin cDNA is 1388 bp long (figs 1,2). Two potential ATG start codons, which are only 6 nucleotides apart, are found near the beginning of the open reading frame (fig.2). Since the translation of eukaryotic mRNAs is mostly initiated from the first AUG [26], we assume that the first ATG is the start codon of human proacrosin.

According to Siegel et al. [18], who have recently purified and partially characterized human sperm proacrosin, the human zymogen shows an apparent molecular mass of 52–55 kDa on SDS-

Fig.3. The deduced amino acid sequence of the human (pro)acrosin heavy chain and its sequence similarity with other human serine proteases. Dashes represent gaps introduced to maximize the sequence identity. The identical amino acids are indicated by black-background letters. The identified or predicted active site residues of serine proteases are shown by asterisks. The last amino acids in the human proacrosin sequence are numbered at the right.

polyacrylamide gel electrophoresis, and is autoactivated to a 49 kDa form followed by further conversion into several lower molecular mass forms. The activation process is very similar to that of boar proacrosin [6,7]. The molecular mass predicted for human proacrosin (43860 Da) is about 10 kDa lower than that determined by SDS-polyacrylamide gel electrophoresis [18]. This discrepancy may be due to the anomalous electrophoretic behaviour of the probably glycosylated proacrosin.

The deduced amino acid sequence of human proacrosin confirms that its mature form, acrosin, belongs to the serine protease family (fig.3). A sequence comparison with other serine proteases indicates that His-69, Asp-123, and Ser-221, all located in the heavy chain, constitute the active site of human acrosin. Two cysteines in the light chain at residues 6 and 10 seem to form the interchain disulfide bonds with two cysteines in the heavy chain (probably residues 135 and 143). It is also likely that 4 intrachain disulfide bonds are formed within the heavy chain between appropriate cysteine residues (probably between residues 54 and 70, 158 and 227, 190 and 206, and 217 and 247).

We also predict that in human acrosin the C-terminus of the heavy chain is either Arg-276, -289, -294, -314, -319, or -326. The rationale for this prediction is two-fold. Firstly, boar acrosin has the proline-rich segment as a C-terminal extension, which is split off during the conversion of the zymogen to the mature form [10]. The proline-rich segment of human proacrosin is located at residues 283–353, in particular at residues 325–353. If this region is cleaved off during the maturation, the cleavage site must be an arginine residue located at the N-terminal side of the proline-rich segment. Secondly, the stretch spanning residues 24–270 of the human proacrosin heavy chain exhibits significant sequence similarities with other serine proteases, as mentioned above, and thus seems to have to be retained in the active, mature enzyme. At any rate, it may be concluded that, like the boar enzyme, human acrosin is synthesized as a single-chain polypeptide (proacrosin), and the zymogen is then converted to the mature form by the proteolytic removal of the C-terminal proline-rich segment and by the cleavage of the peptide bond between the segments corresponding to the light and heavy chains.

Acknowledgement: This work was supported in part by a research grant from the Scientific Research Funds of the Ministry of Education, Science, and Culture of Japan.

REFERENCES

[1] Bhattacharyya, A.K. and Zaneveld, L.J.D. (1982) in: Biochemistry of Mammalian Reproduction (Zaneveld, L.J.D. and Chatterton, R.T. eds) pp.119–151, Wiley, New York.
[2] Schleuning, W. and Fritz, H. (1974) Hoppe-Seyler's Z. Physiol. Chem. 355, 125–130.
[3] Polakoski, K.L. and McRorie, R.A. (1973) J. Biol. Chem. 248, 8183–8188.
[4] Fock-Nüzel, R., Lottspeich, F., Henschen, A., Müller-Esterl, W. and Fritz, H. (1980) Hoppe-Seyler's Z. Physiol. Chem. 361, 1823–1828.
[5] Fock-Nüzel, R., Lottspeich, F., Henschen, A. and Müller-Esterl, W. (1984) Eur. J. Biochem. 141, 441–446.
[6] Polakoski, K.L. and Parrish, R.F. (1977) J. Biol. Chem. 252, 1888–1894.
[7] Parrish, R.F. and Polakoski, K.L. (1978) J. Biol. Chem. 253, 8428–8432.
[8] Töpfer-Petersen, E. and Henschen, A. (1987) FEBS Lett. 226, 38–42.
[9] Jones, R., Brown, C.R. and Lancaster, R.T. (1988) Development 102, 781–792.
[10] Baba, T., Michikawa, Y., Kawakura, K. and Arai, Y. (1989) FEBS Lett. 244, 159–162.
[11] Zaneveld, L.J.D., Dragoje, B.M. and Schumacher, G.F.B. (1972) Science 177, 702–703.
[12] Gilboa, E., Elkana, Y. and Rigbi, M. (1973) Eur. J. Biochem. 39, 85–92.
[13] Schleuning, W.D., Hell, R. and Fritz, H. (1976) Hoppe-Seyler's Z. Physiol. Chem. 357, 855–865.
[14] Tobias, P.S. and Schumacher, G.F.B. (1977) Biochem. Biophys. Res. Commun. 74, 434–439.
[15] Anderson, R.A. jr, Beyler, S.A., Mack, S.R. and Zaneveld, L.J.D. (1981) Biochem. J. 199, 307–316.
[16] Elce, J.S. and McIntyre, E.J. (1982) J. Biochem. 60, 8–14.
[17] Siegel, M.S. and Polakoski, K.L. (1985) Biol. Reprod. 32, 713–720.
[18] Siegel, M.S., Bechtold, D.S., Kopta, C.I. and Polakoski, K.L. (1986) Biochim. Biophys. Acta 883, 567–573.
[19] Huynh, T., Young, R.A. and Davis, R.W. (1985) in: DNA Cloning (Glover, D.M. ed.) vol.1, pp.49–78, IRL, Oxford.
[20] Benton, W.D. and Davis, R.W. (1977) Science 196, 180–182.
[21] Sanger, F., Nicklen, S. and Coulson, A.R. (1977) Proc. Natl. Acad. Sci. USA 74, 5463–5467.
[22] Emi, M., Nakamura, Y., Ogawa, M., Yamamoto, T., Nishide, T., Mori, T. and Matsubara, K. (1986) Gene 41, 305–310.
[23] Fukushima, D., Kitamura, N. and Nakanishi, S. (1985) Biochemistry 24, 8037–8043.
[24] Wiman, B. (1977) Eur. J. Biochem. 76, 129–137.
[25] Leytus, S.P., Chung, D.W., Kisiel, W., Kurachi, K. and Davie, E.W. (1984) Proc. Natl. Acad. Sci. USA 81, 3699–3702.
[26] Kozak, M. (1984) Nucleic Acids Res. 12, 857–872.