Short Communication

Genomic Organization of the Mouse *src* Gene. Sequencing of *src* Introns Revealed a New Chromosome 2 Microsatellite Marker

(*Mus musculus* / protein tyrosine kinase / c-*src* / introns / microsatellite polymorphism)

V. FUČÍK[1], J. BERAN[1], Z. ČERNÝ[1], J. MÁCHA[2], J. JONÁK[1]

[1]Institute of Molecular Genetics, Academy of Sciences of the Czech Republic, Prague, Czech Republic
[2]Faculty of Science, Charles University, Prague, Czech Republic

Abstract. Ten introns interrupting the coding sequence of the mouse *src* protooncogene were sequenced (in total 11260 bp) and their general characteristics compared with the homologous genes in human and chicken. While the study of genome organization of the *src* gene was performed only in the inbred mouse strain BALB/cHeA (*Mus musculus domesticus*), one special region in the intron 5 was also sequenced in additional mouse strains (*M. musculus musculus* and *M. spretus*), because the discovered CA and GT repeat array differences could serve as a new polymorphic marker in the chromosome No. 2 and help elucidate some evolutionary relations between mouse strains.

c-*src* is the first molecularly defined protooncogene described. It was isolated as the cellular homologue of the v-*src* gene (Stehelin et al., 1976), the transforming portion of the first characterized oncogenic virus (Rous, 1911).

The gene product Src, a protein with tyrosine kinase activity, is a member of the Src-kinase family that contains about ten related molecules with an identical architecture of functional domains and a common regulation mode. The importance of Src and its siblings is documented by the fact that one or more of these genes are present in every higher-animal genome (Brown and Cooper, 1996). Each Src family member consists of about 530 amino acid residues in its non-neuronal form, for example in chicken it is 533 residues, in mouse 535 and in human it is 536 amino acid residues. The Src protein is highly conserved, for instance exons 3–12 share 98% identity between human, mouse and chicken genes. On the other hand, the primary structure of their introns shows only a low resemblance.

Received August 3, 2001. Accepted November 16, 2001.

This work was supported by grant No. 312/96/K205 from the Grant Agency of the Czech Republic (to J. J.).

Corresponding author: Vladimír Fučík, Institute of Molecular Genetics, Academy of Sciences of the Czech Republic, Flemingovo nám. 2, Prague 6, Czech Republic. Fax: +420 (2) 33331274; e-mail: fucik@img.cas.cz.

The human *src* gene sequence available in the database is part of a clone from chromosome 20, accession No. AL133293.

The current Mouse Genome Sequencing Page at NCBI contains ten contigs for chromosome 2 totalling 1635043 bases, but the *src* gene is not included. Known are the data on the regulatory sequences within the 3rd mouse intron, which was thoroughly studied in the laboratory of D. L. Black (Chan and Black, 1995; Modafferi and Black, 1997).

A few sequenced but not specified clones were found when our sequencing data were compared with databases.

In the course of our earlier studies of *src* gene involvement in the embryonic development (Takáč et al., 1992; Habrová et al., 1996; Mácha et al., 1997; Takáč et al., 1998; Jonák, 2000) we decided to determine its genomic structure in mouse. The exon-intron structure of the mouse c-*src* protooncogene was also compared with that of human and chicken c-*src*.

Results and Discussion

We determined the sequences of ten introns totalling 11290 bp separating the translated exons of the *src* gene of the mouse inbred line BALB/cHeA (Fig. 1) and the results were submitted to the EMBL Nucleotide Sequence Database. The accession numbers assigned to individual introns are as follows: AJ312264 (intron 2), AJ313176 (intron 3), AJ312263 (intron 4), AJ313177 (intron 5), AJ31226262 (intron 6), AJ312261 (intron 7), AJ312260 (intron 8), AJ312259 (intron 9), AJ312258 (intron 10) and AJ313104 (intron 11). Here, we present some general characteristics of these regions and compare our data with the human and chicken *src* genomic DNA sequences.

While the primary structure of exons remains highly conserved, all *src* introns of *M. m. domesticus* (BALB/cHeA) turned out to be shorter than those of human (Table 1). The total length of human *src* introns

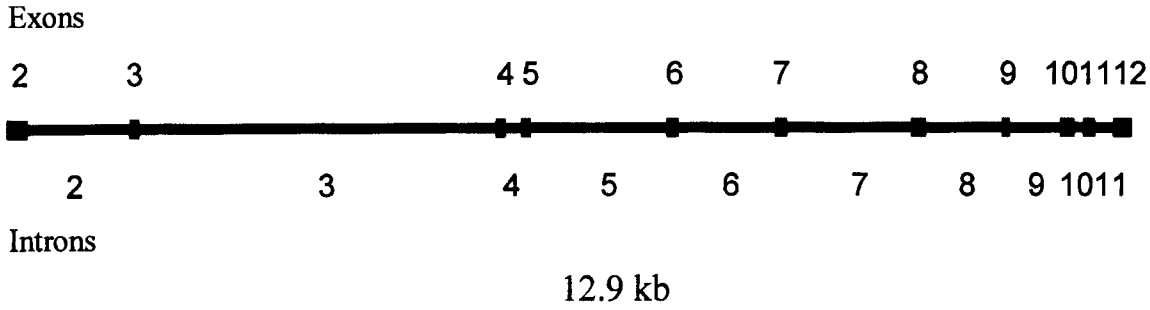


Fig. 1. The exon-intron structure of the mouse c-*src* gene (the distances are drawn to scale).

is more than six thousand bp larger. A still larger difference occurred in comparison with chicken introns, whose length remained more than 7000 bp shorter than that of mouse introns and 13000 bp shorter in comparison to human introns (Table 1). Repeats disclosed in mouse introns with the aid of programmes Repeat Masker[i] and Repeatview[ii] were marked in the sequences submitted to the EMBL database. Repeats from the database of rodents were found in introns 3, 6 and 8. In addition to that, human database-registered repeats were detected in the introns 3 (TIGGER2) and 9 (L1MED_5).

The GC content of individual introns varies between 51.2 to 62.3%, but the mean value of GC calculated from the whole length of 10 introns (total 11290 bp) makes up 56.9%, while the value for the coding region amounts to 60.4% GC (Table 2). It is interesting to note that the GC composition in both exons and introns drifted to higher values from mouse to human.

All junctions between exons and introns conform to the GT-AG rule and the introns interrupt the coding sequences in all three phases, i.e. between or within the coding triplets. Seven out of ten *src* introns are phase 1 introns, i.e. they are inserted between the first and the second residue of a coding triplet (Table 2). Therefore, most of the exons are symmetrical, meaning that they are flanked by introns in the same phase.

Comparison of human/mouse homologies between *src* introns revealed a higher degree of homology between the pairs of longer introns than between the shorter ones. The most conserved is intron 3, especially in the vicinity of the alternative neuronal exons N1 and N2, where the identity between mouse and human is 100%.

Alignment of human and mouse intron sequences confirmed a low level homology also for introns 2, 5, 6 and 9.

Blast[iii] comparison of the introns 5 confirmed the existence of clusters of GT stretches in the human *src*

Table 1. The lengths of chicken, mouse and human src gene introns (bp)

Intron No.	Chicken	Mouse	Human
2	50	1202	1671
3	2040	4076	7720
4	390	170	180
5	1010	1561	1884
6	350	1092	1387
7	85	1408	2260
8	78	872	1307
9	61	566	756
10	118	111	160
11	79	232	290
Total	4261	11290	17615

Our strategy for sequencing the mouse *src* introns made use of the data published by Takeya and Hanafusa, who determined the boundaries between exons and introns of the *src* gene in chicken. They deduced the lengths of introns of the chicken *src* gene by either electron microscopy or DNA sequencing (Takeya and Hanafusa, 1983).

On account of the high homology of the chicken and mouse Src products (Martinez et al., 1987) it was possible to find the boundaries in the mouse *src* gene and to design the pairs of forward and reverse PCR primers located in exons flanking the individual introns. PCR products were purified by agarose electrophoresis and phenolization, and the PCR templates were used for direct sequencing without cloning. Cycle sequencing was performed with a BigDye Terminator (Perkin Elmer Applied Biosystems, Prague, Czech Republic) and electrophoresis of the samples was run in an ABI PRISM 310 genetic analyser. The sequencing proceeded in both directions with the aid of the original PCR primers. As most introns were longer than to be sequenced in a single run, further primers were designed according to the obtained sequences. The sequencing was repeated in case of disagreement between the results from the two directions.

[i] Homepage http://repeatmasker.genome.washington.edu/

[ii] Used server Repeatview at the Institute of Advanced Biomedical Technologies (ITBA) http://l25.itba.mi.cnr.it/ genebin/wwwrepeat.pl

[iii] WWW page http://www.ncbi.nlm.nih.gov/gorf/bl2.html

Fig. 2. Sequence comparison of a polymorphic region in the intron 5 of different mouse strains

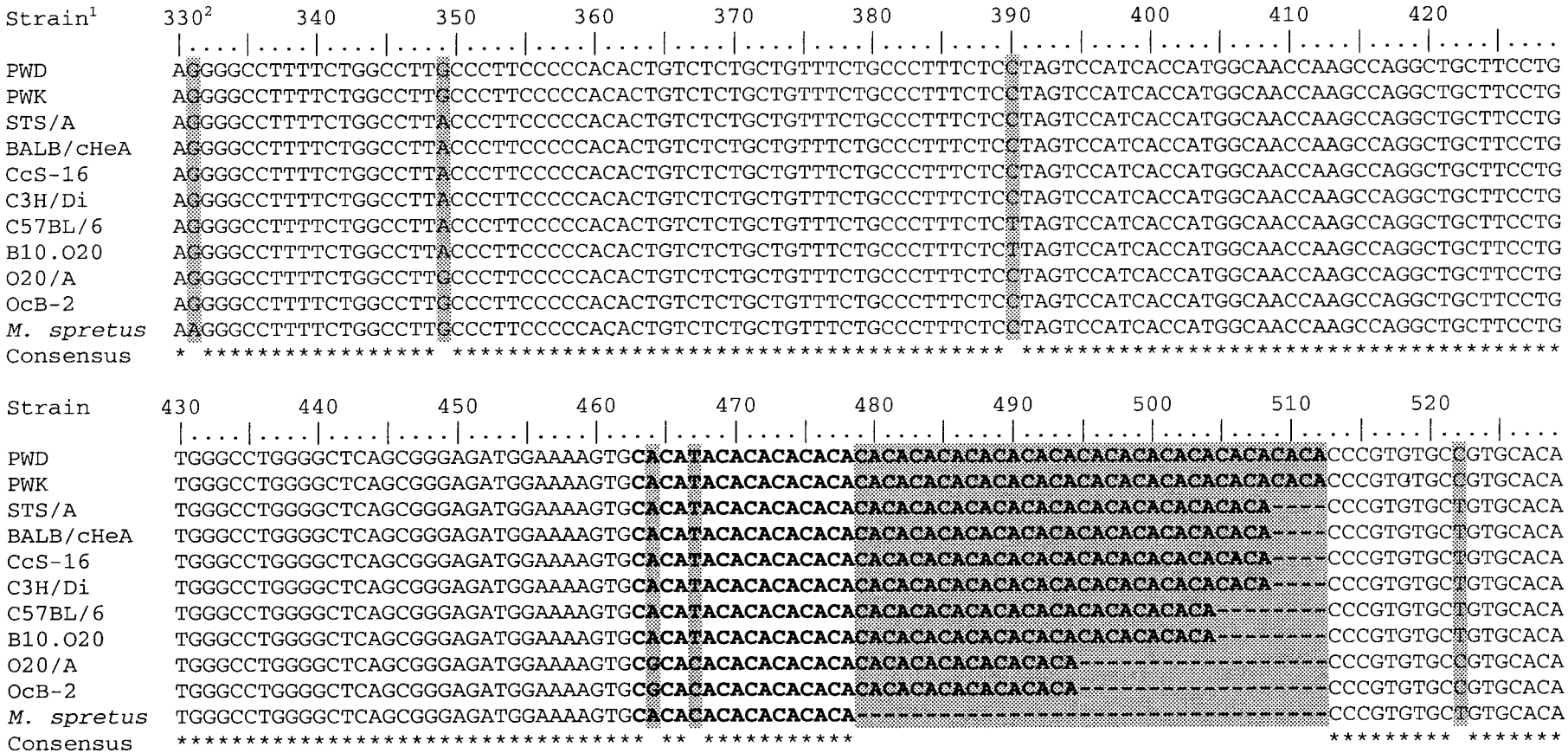
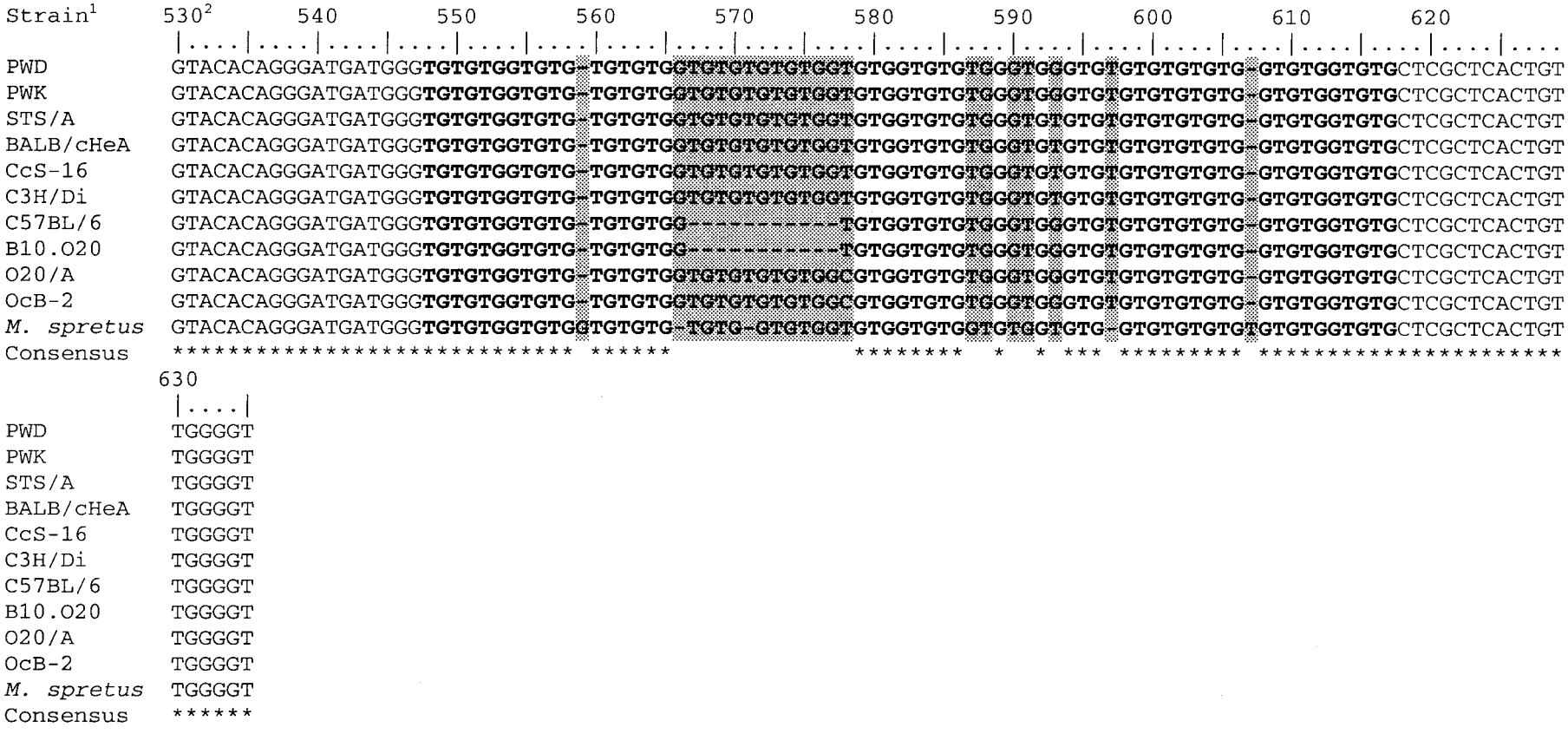


Fig. 2. (continued)



[1]PWD and PWK are inbred lines of *Mus musculus musculus* (Pavljuková and Forejt, 1981; Gregorová and Forejt, 2000). STS/A, BALB/cHeA, C3H/Di, C57BL/6, B10.O20 (full name: B10.O20/R164/Dem), O20/A are inbred lines of the *Mus musculus domesticus* subspecies. CcS-16 and OcB-2 are recombinant congenic strains (RCS) from the CcS/Dem and OcB/Dem series, respectively, a genetic tool designed to map the genes involved in the multigenic control of complex genetic traits (e.g. Lipoldová et al., 2000). A series of RCS were produced using the background and donor strain pair: BALB/cHeA and STS/A (the CcS/Dem series) and O20 and B10.O20 (the OcB/Dem series). The RC strains are homozygous and carry 12% of genes from the donor strain and 88% of genes from the background strain (Stassen et al., 1996). The sample of a different species included in this comparison was from *Musculus spretus*. DNA samples of *M. m. domesticus* were a generous gift from Dr. Lipoldová, the others from Dr. Gregorová (Inst. Mol. Genet., Prague). The alignment was processed with the help of the biological sequence editor Bioedit (Hall, 1999).

[2]The numbers reflect the distance in bp from the 5' end of intron 5.

Table 2. GC composition of individual introns and exons in mouse and human src genes

No.	Mouse exons	Human exons	Mouse introns	Human introns	Phase of introns
	%GC	%GC	%GC	%GC	
2	73.28	74.80	60.48	62.78	1
3	53.00	53.00	57.58	58.43	2
4	61.54	63.64	62.35	65.56	2
5	56.73	53.85	55.35	60.56	1
6	54.67	58.67	59.43	62.22	1
7	62.18	67.95	55.97	57.74	1
8	53.89	56.11	53.10	48..28	1
9	55.84	59.74	51.24	58.07	0
10	56.49	59.74	54.95	58.13	1
11	56.06	59.85	57.33	59.66	1
12	62.68	66.99			
Mean	60.02	62.76	56.94	58.60	

Note: The mean %GC was computed for the whole lengths of all exons and introns, respectively. The phase of introns denotes after which nucleotide in the coding triplet the intron is inserted.

gene at a similar position, i.e. in the first third of the intron, but not as compact as in the mouse.

In the mouse intron 5, the GT region is preceded by a CA stretch, a new microsatellite marker for chromosome 2 that we designated D2Img1 (DNA segment, chromosome 2, Institute of Molecular Genetics 1). We have also sequenced this region in several other mouse inbred lines and the results are demonstrated in Fig. 2. The whole intron 5 is 1561 bp long in the BALB/cHeA strain. After the G462 (counted from the 5' end of the intron), which is common to all eleven strains tested, we can see the CA stretches of different lengths and, 35 bp apart, a sequence composed of G and T (which means also CA in the antiparallel strand). Yet, while the CA sequences in the depicted plus strand vary in the number of dinucleotide repeats, the length of the GT region remains constant (with the exception of C57BL/6, and B10.O20) but contains examples of several single-nucleotide insertions/deletions or base changes (see the GGT and GGGT interspersed between $(GT)_n$ repeats). This characteristic could be explained as another example of different propensity of the leading and the lagging strand to accumulate different types of mutations (Freudenreich et al., 1997).

The GT regions in both black inbred lines (C57BL/6 and B10.O20) shown in Fig. 2 are 11 bp shorter than in other *M. musculus* strains. This finding is in agreement with the available sequence of one RIKEN-library clone of the C57BL/6 strain (accession No. AK0176290). The clone is annotated as a reverse transcript of mRNA, but in fact it encompasses the 3' part of intron 5, the whole exon 6 and the 5' part of intron 6.

Sequences of human and mouse *src* genes were also scanned for the occurrence of CpG islands[iv]. In the human sequence we discovered four short islands, 305, 501, 241 and 212 base pairs long. Three of them are present within the exon 2 and the adjacent intron 2. The fourth one encompasses the 3' part of intron 9, the whole exon 10 and the 5' end of intron 10. No CpG island was found upstream of the translated region.

In the mouse *src* gene only one CpG island, 253 bp long, was found within the intron 2. In chicken we could not include the introns in the search for the CpG islands, because their full sequence is not available (Takeya and Hanafusa, 1983), but in the cDNA, two islands were discovered at positions corresponding to exons 2–4 and 5–7, respectively.

The generally observed CpG depletion of the vertebrate DNAs is commonly explained by methylation of cytosine, followed by deamination to thymine. This process seems to be more rapid in rodents. Antequera and Bird (1993) have brought examples of coding sequences of mouse genes in which, in contrast to the human homologues, the CpG dinucleotides have been replaced at the equivalent positions by TpG in 40–72%.

In the *src* gene we could not confirm such pronounced differences. Mutability of CpG dinucleotides in coding sequences is strictly influenced by their position in the coding triplet. For instance, CpG dinucleotides in human c-*src* were found changed to TpG in mouse or chicken in about 14%, but only in case the C was situated at the third (wobble) place of the triplet.

The mean densities of CpG dinucleotides per 100 bp of cDNA found in chicken, mouse and human are 5.4, 5.3 and 6.1, respectively. Corresponding values for mouse and human introns are 1.8 and 2.5, respectively.

The nucleotide sequence data reported in this paper have been submitted to GenBank and have been assigned the accession numbers: AJ312264, AJ313176, AJ312263, AJ313177, AJ312262, AJ312261, AJ312260, AJ312259, AJ312258 and AJ313104.

References

Antequera, F., Bird, A. (1993) Number of CpG islands and genes in human and mouse. *Proc. Natl. Acad. Sci. USA* **90,** 11995-11999.

Brown, M. T., Cooper, J. A. (1996) Regulation, substrates and functions of Src. *Biochim. Biophys. Acta* **1287,** 121-149.

Chan, R. C., Black, D. L. (1995) Conserved intron elements repress splicing of a neuron-specific c-*src* exon in vitro. *Mol. Cell Biol.* **15,** 6377-6385.

Freudenreich, C. H., Stavenhagen, J. B., Zakian, V. A. (1997) Stability of a CTG/CAG trinucleotide repeat in yeast is

[iv] We used the CpG Island finder and plotting tool (EMBOSS) at European Bioinformatics Institute server http://www.ebi.ac.uk/emboss/cpgplot/

dependent on its orientation in the genome. *Mol. Cell Biol.* **17,** 2090-2098.

Gregorová, S., Forejt, J. (2000) PWD/Ph and PWK/Ph inbred mouse strains of *Mus m. musculus* subspecies – a valuable resource of phenotypic variations and genomic polymorphisms. *Folia Biol. (Praha)* **46,** 31-41.

Habrová, V., Takáč, M., Navrátil, J., Mácha, J., Češková, N., Jonák, J. (1996) Association of Rous sarcoma virus DNA with *Xenopus laevis* spermatozoa and its transfer to ova through fertilization. *Mol. Reprod. Dev.* **44,** 332-342.

Hall, T. A. (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucl. Acids Symp.* **41,** 95-98.

Jonák, J. (2000) Sperm-mediated preparation of transgenic *Xenopus laevis* and transmission of transgenic DNA to the next generation. *Mol. Reprod. Dev.* **56,** 298-300.

Lipoldová, M., Svobodová, M., Krulová, M., Havelková, H., Badalová, J., Nohynková, E., Holáň, V., Hart, A. A., Volf, P., Demant, P. (2000) Susceptibility to *Leishmania major* infection in mice: multiple loci and heterogeneity of immunopathological phenotypes. *Genes Immun.* **1,** 200-206.

Mácha, J., Štursová, D., Takáč, M., Habrová, V., Jonák J. (1997) Uptake of plasmid RSV DNA by frog and mouse spermatozoa. *Folia Biol. (Praha)* **43,** 123-127.

Martinez, R., Mathey-Prevot, B., Bernards, A., Baltimore, D. (1987) Neuronal pp60c-*src* contains a six-amino acid insertion relative to its non-neuronal counterpart. *Science* **237,** 411-415.

Modafferi, E. F., Black, D. L. (1997) A complex intronic splicing enhancer from the c-*src* pre-mRNA activates inclusion of a heterologous exon. *Mol. Cell Biol.* **17,** 6537-6545.

Pavljuková, H., Forejt, J. (1981) Three inbred strains derived from wild mice carrying Hybrid sterility mutations. *Mouse News Lett.* 65.

Rous, P. (1911) A sarcoma of the fowl transmissible by an agent separable from the tumor cells. *J. Exp. Med.* **13,** 121-149.

Stassen, A. P., Groot, P. C., Eppig, J. T., Demant, P. (1996) Genetic composition of the recombinant congenic strains. *Mamm. Genome* **7,** 55-58.

Stehelin, D., Varmus, H. E., Bishop, J. M., Vogt, P. K. (1976) DNA related to the transforming gene(s) of avian sarcoma viruses is present in normal avian DNA. *Nature* **260,** 170-173.

Takáč M., Černá J., Habrová V., Štokrová J., Rychlík I., Jonák J. (1992) Microinjection of cloned proviral Rous sarcoma virus DNAs into *Xenopus laevis* one cell embryos. *Folia Biol. (Praha)* **38,** 65-77.

Takáč, M., Habrová, V., Mácha, J., Češková, N., Jonák, J. (1998) Development of transgenic *Xenopus laevis* with a high c-*src* gene expression. *Mol. Reprod. Dev.* **50,** 410-419.

Takeya, T., Hanafusa, H. (1983) Structure and sequence of the cellular gene homologous to the RSV *src* gene and the mechanism for generating the transforming virus. *Cell* **32,** 881-890.