Construction and Analysis of Full length-enriched and 5'-end-enriched cDNA libraries using the "Oligo-Capping".

Oligo-Capping 法を用いた完全長および5'端特異的 cDNA library の作製と解析

Yutaka Suzuki

鈴木 穣

The University of Tokyo, Graduate School of Arts and Sciences
Department of Life Sciences

東京大学大学院総合文化研究科広域科学専攻生命環境科学系

①

Construction and Analysis of
Full length-enriched and 5’-end-enriched cDNA libraries
using the “Oligo-Capping”.

Oligo-Capping 法を用いた完全長および 5’端特異的 cDNA library の作製と解析

Yutaka Suzuki

鈴木　穣

The University of Tokyo, Graduate School of Arts and Sciences,
Department of Life Sciences.

東京大学大学院総合文化研究科広域科学専攻生命環境科学系

TABLE OF CONTENTS

I. INTRODUCTION 3

II. MATERIALS AND METHODS 10

IIIA. RESULTS AND DISCUSSIONS (I) 13
Construction and Characterization of the "Oligo-capped" cDNA libraries.

IIIB. RESULTS AND DISCUSSIONS (II) 31
Analyses of the "Oligo-capped" cDNA libraries.

IV. CONCLUSION 52

V. REFERENCES 53

VI. ACKNOWLEDGEMENTS 59

I. INTRODUCTION

I) Human Genome Project

In 1958, a theory, called the "central dogma" was proposed[1]. It describes the general flow of the genetic information. In this theory, the hereditary information of all organisms is encoded as a nucleotide sequence of DNA (genome sequence), which is transcribed to RNA. A class of RNA, called mRNA, conveys the information to the cytoplasm. In cytoplasm it serves as a template when the genetic code of DNA is translated into the protein. The protein constitutes the most part of the cell structures and catalyzes the chemical reactions that are essential to keep the homeostasis.

According to the central dogma, the DNA sequence is the original source of every gene. It codes the structure of RNA and protein. It provides the blueprint of gene functions. Since DNA sequence maintains all the primary information, in a sense, it can be said DNA defines the organism. One of the ultimate ways to describe the organism may be the determination of its genome sequence[2), 3)].

In 1970s, technological innovations were made concerning with the DNA manipulation and the DNA analysis. In the early 70s, the DNA cloning technique was introduced[4), 5)]. This technique made it possible to amplify and modify DNA arbitrarily. It became one of the most powerful tools in the molecular biology. In 1975, the DNA sequencing methods were reported[6), 7)]. They became further useful tools for the analysis of the DNA sequence.

With the powerful technology of the molecular biology, the first example of the genome sequence was produced for ϕX174 in 1977[8)]. It consisted of 5,386 base pairs (bp) of DNA. In 1981, the DNA sequence of human mitochondria was determined (16,570bp)[9)]. Through the 1980s, the sequence analyses gradually came up to the larger genome. The genome sequencing of Epstein-Barr virus (172,000bp) and human cytomegalo virus (229,000bp) was finished in 1984 and in 1989, respectively[10), 11)]. In the late 1980s, the possibility of the human genome sequencing (3,000Mb) came to be seriously discussed[12)].

The genome sequence was also expected in the cancer research. The researchers had to detect the genetic rearrangement that occurs during the cancer progression. It required the genome sequence as a standard control. In 1986, Dulbecco mentioned the significance of the genome sequencing in his paper in *Science*[13)]. He said, "If we wish to learn more about cancer, we must now concentrate on the cellular genome."

Around the end of 80s, the stage seemed to be set to sequence the entire human genome. Introduction of the YAC (Yeast Artificial Chromosome) vector and the pulse field gel electrophoresis had accelerated the physical mapping of the microorganism genomes, such as *S. cerevisiae* and *C. elegans*[14)-18)]. PCR had made it possible to quickly amplify the DNA fragment with a faint amount of starting samples at a low cost[19), 20)]. The step of the sequence analysis had been simplified by the development of the DNA auto-sequencing machines[21)].

In 1988 an international organization, called the HUGO (Human Genome Organization) was established. It was organized to carry out the project smoothly with the cooperation of many countries. In 1990 the "Human Genome Project" started in the United States. It set its final goal in the complete determination of the human genome sequence[22)].

II) cDNA analysis in the Human Genome Project

In Japan the human genome project also started in the international cooperation. The Ministry of Education, Science and Culture, the Ministry of Health and Welfare and the Science and Technology Agency took the central role in the early project. Receiving the report of the Science Advisory Council, the Ministry of Education decided to take part in the genome project in 1991[23].

The plan of the Ministry of Education had one distinctive feature compared with the plan of the United States. It set the cDNA (a faithful copy of mRNA) analysis as one of the central parts of the project[24]. In 1992 Matubara, a leader of the genome committee, said in his paper, "Large scale sequencing of cDNA provides a complementary approach to structural analysis of the human genome."[25]

It cannot be directly deduced from the genome sequence how each gene is expressed and carries out various life activities in a cell. Additional efforts should be paid to elucidate the mRNA sequence and the protein functions. The analysis of mRNA has great advantages, since information about the gene functions is concentrated on the mRNA sequence. Through the mRNA analysis, we can obtain the sequence information such as:

a) The mRNA transcription start site, which is indispensable for the exact identification of its promoter region.

b) The 5' untranslated region (5'UTR), which is related to the translation efficiency and the cellular localization of mRNA[26], [27], [28].

c) The protein coding region (CDS).

d) The 3' untranslated region (3'UTR), which is related to the translation efficiency, the cellular localization of mRNA and its stability[26]-[31].

Thus, the mRNA sequence contains the precious information to presume the gene function. The annotation about the gene expression and the protein function could be put to the genome sequence through the cDNA analysis[25], [32].

In the middle of 90s, the cDNA analysis also prevailed outside Japan.

Several projects were formed for large-scale cDNA sequencing. In the United States the Washington University EST Project started in 1994, funded by Merck & Co. and the National Cancer Institute[33)]. At the NCBI (National Center for Biotechnology Information), a database, called the dbEST was constructed. It contains the one-pass sequence data and other information on randomly selected cDNA clones. As a result of intensive efforts, more than 1 million entries have been accumulated in the dbEST[34)].

III) Construction of a full-length-enriched and a 5'-end-enriched cDNA library using the "Oligo-capping".

A large drawback exists in the cDNA libraries that are widely used for the current large-scale cDNA analyses. On many occasions, reverse-transcriptase can not make a full cDNA copy of a mRNA but stops in the middle leaving an incomplete copy. The cDNA libraries made by the conventional methods contain many incomplete cDNA clones (Fig.I-1). They usually lack the 5'-end sequences of the template mRNA. Thus, current cDNA data mainly covers the 3'-ends of mRNA and the information around the 5'-ends still remains poor. Additional work would be required to determine the sequence around the 5'-end for each mRNA species. There are not many genes among the database entries whose transcription start site is clearly defined. To complement this drawback, I considered we should analyze the full-length cDNA, which contains all the sequence of mRNA between the cap structure and the polyA. The cDNA libraries consisting of full-length clones should be essential for that purpose.

Maruyama and Sugano previously reported a new method, called the "Oligo-capping"[35]. This method made it possible to replace the cap structure with the synthetic oligo-nucleotide (5'-oligo). I applied it to the construction of a cDNA library. Using the "Oligo-capped" mRNA as a starting material, I constructed a "full length-enriched cDNA library".

The full length-enriched cDNA library may not include cDNA of long mRNA because the distance between the cap structure and polyA could be beyond the limits of reverse-transcriptase or DNA polymerase for long mRNA molecules. In case that the full-length clone is not obtained at one time, I constructed a "5'-end-enriched cDNA library". This library is expected to cover the 5'-ends of long mRNA.

In this thesis I will describe the construction and characterization of a full length-enriched and a 5'-end–enriched cDNA library in the chapter III-A. With this system, I constructed the "Oligo-capped" cDNA libraries

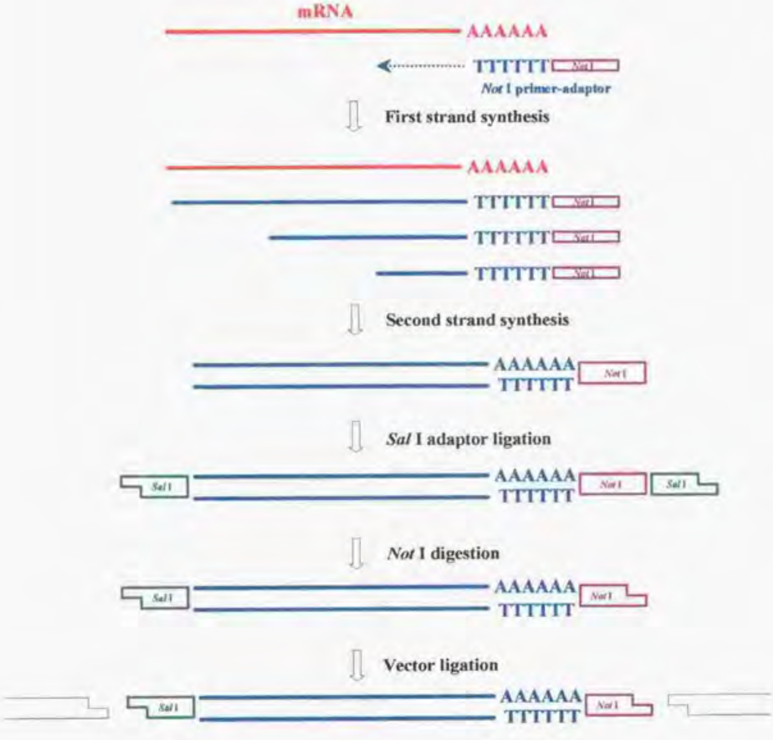


Fig.I-1 Conventional method to construct a cDNA library (Gubler-Hoffman method). Also see the section IIIA-V.

from various kinds of tissues and cultured cells. Through the one-pass sequencing of these libraries, I could reveal several detail features around the 5'-ends of mRNA. In the chapter III-B, I will also report the result of statistical and thermodynamic analyses of our cDNA clones.

II. MATERIALS AND METHODS

I) Cells.

Human neuroblastoma cell line SK-N-MC was obtained from American *Type Culture* Collection and grown as described[36].

II) Isolation of RNA.

Cytoplasmic RNA and Poly A+ RNA were isolated according to the standard method[37]. Oligo-dT cellulose was from Collaborative Biomedical Products and Roche.

III) Oligo-Capping.

Oligo-capping was performed as described[35, 38] with some modifications. In brief, 5 to 10μg of polyA+ RNA was treated with 1.2 units of bacterial alkaline phosphatase (BAP; TaKaRa) in 100μl of 100mM Tris-HCl (pH 8.0), 5mM 2-mercaptoethanol with 100 units of RNasin (Promega) at 37°C for 40 min. After extraction with phenol:chloroform (1:1) twice and ethanol precipitation, the polyA+ RNA was treated with 20 units of tobacco acid pyrophosphatase (TAP)[35] in 100μl of 50mM sodium acetate (pH 5.5), 1mM EDTA, 5mM 2-mercaptoethanol with 100 units of RNasin at 37°C for 45 min. After phenol:chloroform extraction and ethanol precipitation, 2 to 4μg of the BAP-TAP treated polyA+ RNA were ligated with 0.4μg of 5'-oligo (KM-02; 5'-AGC AUC GAG UCG GCC UUG UUG GCC UAC UGG-3') using 250 units of RNA ligase (TaKaRa) in 100μl of 50mM Tris-HCl (pH7.5), 5mM $MgCl_2$, 5mM 2-mercaptoethanol, 0.5mM ATP, 25% PEG8000 with 100 units of RNasin at 20°C for 3 to 16 hours.

IV) cDNA synthesis.

After removing unligated 5'-oligo, cDNA was synthesized with RNaseH free reverse-transcriptase (Superscript II, Gibco BRL). For the full length-enriched library, 10pmol of dT adapter-primer (5'-GCG GCT GAA GAC GGC CTA TGT GGC CTT TTT TTT TTT TTT TTT-3') was used in 50μl with 2 to 4μg of oligo-capped polyA+ RNA. The reaction conditions were as recommended by the supplier and incubated at 42°C for 1 hour. For the 5'-end-enriched cDNA library, 10pmol of random adapter-primer (5'-GCG GCT GAA GAC GGC CTA TGT GGC CNN NNN NC-3') was used and incubated at 12°C for 1 hour and 42°C for another hour.

V) cDNA amplification.

After first strand synthesis, RNA was degraded in 15mM NaOH by incubating at 65°C for 1 hour. The cDNA which is made from 1μg "Oligo-capped" polyA+ RNA was amplified in a volume of 100μl using an XL PCR kit (Perkin-Elmer) with 16pmol of 5' (5'-AGC ATC GAG TCG GCC TTG TTG-3') and 3' (5'-GCG GCT GAA GAC GGC CTA TGT-3') PCR primers. For dT-adapter primer primed cDNA, amplification cycles were 5 to 10 cycles at 94°C for 1 min, 58°C for 1 min, and 72°C for 10 min. For random adapter primer primed cDNA, amplification cycles were 25 to 30 cycles at 94°C for 1 min, 58°C for 1 min, and 72°C for 2 min. PCR products were extracted with phenol:chloroform (1:1) once, ethanol precipitated and digested with *Sfi*I. *Sfi*I-digested PCR products were separated by an agarose gel electrophoresis and products longer than 1,000bp were isolated and cloned into *Dra*III-digested pUC19-FL3 or pME18S-FL3. In this way, we could clone the cDNA into the vector in an orientation-defined manner[39].

VI) Sequencing.

Plasmid DNA was isolated using PI-100 and PI-200 auto-plasmid-isolators (KURABO). Sequences were determined by the dideoxy termination method[7] using an AutoCycle sequencing kit (Pharmacia) and a reaction robot R. O. B. DNA processor (Pharmacia) or BigDye sequencing kit (ABI). The sequence was read by ALF DNA (Pharmacia) and ABI 377XL (ABI) auto-sequencers.

VII) Sequence similarity test.

Sequence similarity of cDNA was tested against GenBank none-redundant nucleotide library (Release 98) using BLASTN[40)] or FASTA[41)] program.

VIII) Database construction.

Database construction was performed using the sequence clustering program, DYNACLUST (DYNACOM).

IX) Secondary structure calculation.

The distribution of the local binding energy of the 3'-end of 18S rRNA to the 5'UTR sequences were calculated by the thermodynamic program, SECDYN2, which calculates the optimal secondary structure composed of more than one RNA molecules, using the algorithm based on the dynamic programming[42)].

IIIA. RESULTS AND DISCUSSIONS (I)

Construction and Characterization of the "Oligo-capped" cDNA libraries.

1) **Scheme for the construction of the "Oligo-capped" cDNA libraries.**

Eucaryotic mRNA has a specific structure at its 5'-end, called the "cap structure" (Fig. IIIA-1)[43]. As shown in Fig. IIIA-1, tobacco acid pyrophosphatase (TAP) hydrolyzes the cap strucutre and leaves the phosphate group at the 5'-end of mRNA[44]. Bacterial alkaline phosphatase (BAP) can remove the phosphate group that sticks out from the mRNA 5'-end, but cannot destruct the cap structure itself. T4 RNA ligase shows its activity only towards the 5'-end phosphate group.

Making use of these enzyme features in the successive reactions, the "Oligo-capping" enables the replacement of the cap structure with the synthetic 5'-oligo. First, BAP hydrolyses the phosphate group at the truncated mRNA 5'-end from which the cap structure has been taken away. Second, TAP removes the cap structure, leaving the phosphate group at the 5'-end. Finally, RNA ligase binds 5'-oligo to the phosphate group (Fig.IIIA-2A)[35].

With "oligo-capped" mRNA as a starting material, I constructed two new types of cDNA libraries. One is full length-enriched cDNA library, for which the first strand cDNA was synthesized with dT primer. The other is 5'-end-enriched cDNA library. For this library, random primer was used instead of dT primer for the first strand cDNA synthesis. The first strand cDNA was amplified by PCR with the cap-replaced 5'-oligo sequence for the 5' primer. After the size fractionation, the PCR products were cloned into the vector plasmid, pUC19-FL3 or pME18-FL3 in an orientation defined manner (Fig. IIIA-2B, 2C)[38), 39)].

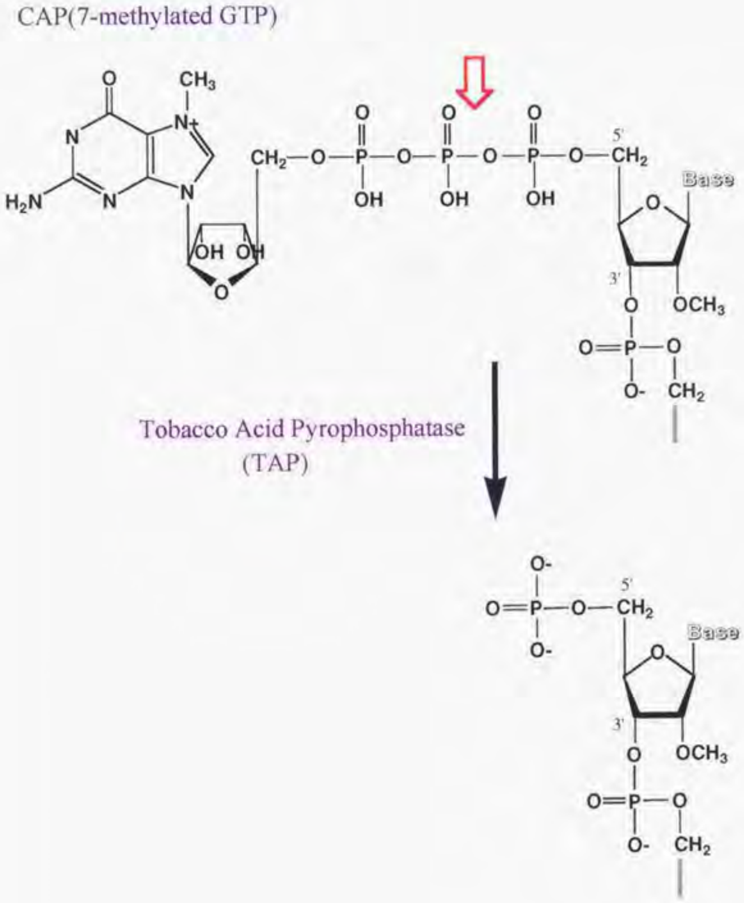


Fig.IIIA-1 Eucaryotic cap structure and the TAP activity.
TAP hydrolyzes the eucaryotic cap structure at the position suggested by a red arrow.

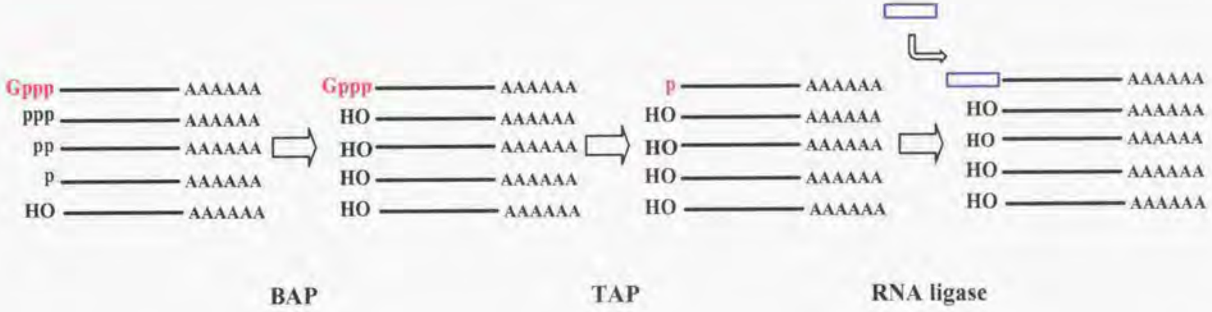


Fig.IIIA-2A “Oligo-capping” procedure.
RNA molecules are represented as solid lines and 5’-oligo as blue boxes. PolyA+RNA consists of RNA molecules with various types of 5’-ends as shown at the left margin. Gppp: cap structure; p: phosphate; OH: hydroxyl.

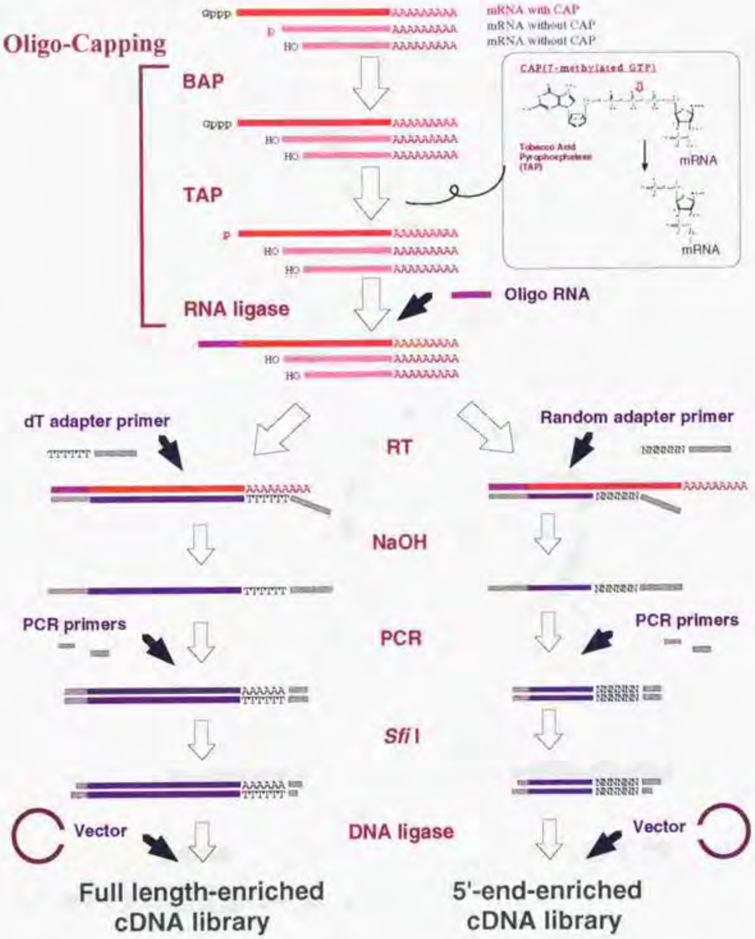


Fig.IIIA-2B Scheme to construct a full length-enriched and a 5'-end-enriched cDNA libraries.

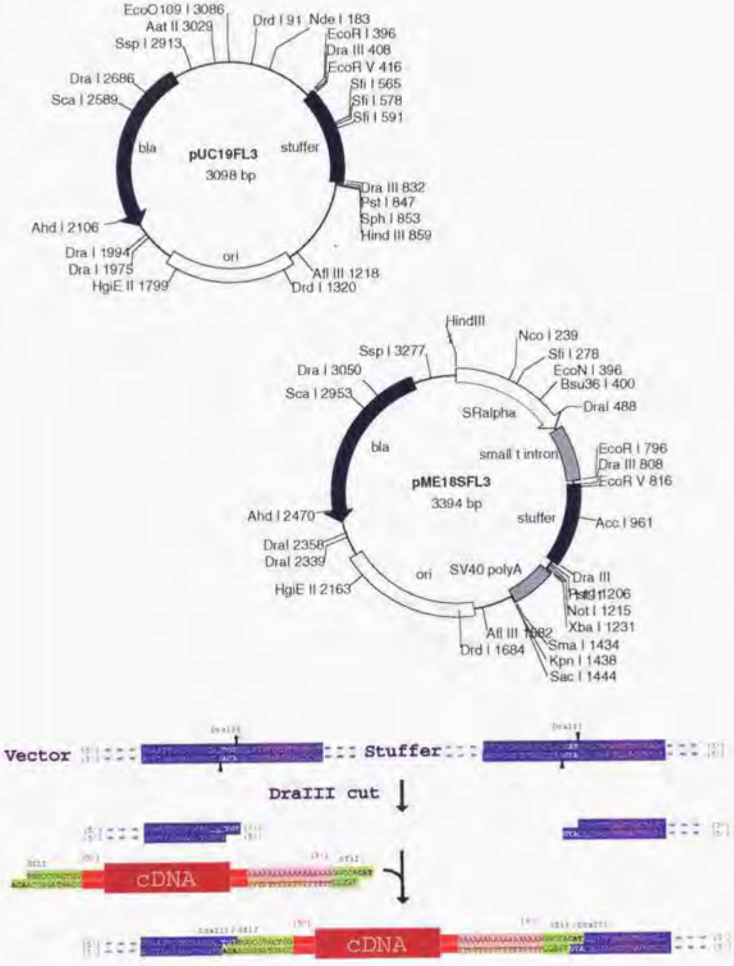


Fig.IIIA-2C **cDNA cloning and the plasmid vectors.**

II) Construction of a full length-enriched and a 5'-end-enriched cDNA library from the human neuroblastoma cell line.

According to the scheme shown in Fig.IIIA-2B, I constructed a full length-enriched and a 5'-end-enriched cDNA library using polyA+ RNA from the human neuroblastoma cell line, SK-N-MC. The size of the cDNA library was about 20,000 clones/μg of polyA+ RNA for the full length-enriched cDNA library and about 200,000 clones/μg for the 5'-end-enriched cDNA library. The average length of cDNA inserts was about 1,500bp for the full length-enriched library and about 1,000bp for the 5'-end-enriched cDNA library.

I then randomly selected cDNA clones from both cDNA libraries and determined the one-pass sequences of the 5'-ends of these clones. Since the 5'PCR primer (5'-AGC ATC GAG TCG GCC TTG TTG-3') has only a part of the 5'-oligo sequence (5'-AGC AUC GAG UCG GCC UUG UUG GCC UAC UGG-3'), the sequence GCCTACTGG at the 5'-end of the clone indicates the ligation of the 5'-oligo at the RNA level. Of 243 clones sequenced, all had the sequence GCCTACTGG. This result suggested that cDNA clones in these libraries were derived only from the "Oligo-capped" mRNA, and the sequences following the 5'-oligo should have come from the mRNA.

The sequence similarity test using these sequences showed that about 40% of the clones matched known genes, about 17% of the clones matched only with expressed sequence tags (ESTs) and the rest did not show any significant similarity with the sequences in the database (Table IIIA-1). The lists of the clones that matched known genes are shown in Table IIIA-2

Table IIIA-1. cDNA clones from the full length-enriched and the 5’-end-enriched cDNA library.

full length-enriched cDNA library		5’-end-enriched cDNA library
	number of clones (%)	number of clones (%)
known total	35 (42%)	62 (39%)
	5’-full 28 (80%)	51 (82%)
	not full 7 (20%)	11 (18%)
EST	15 (18%)	27 (17%)
new	34 (40%)	70 (44%)
total	84 (100%)	159 (100%)

Table IIIA-2. List of clones that matched with known genes.

Clones from full length-enriched cDNA library

Clone number	Homology[a]	locus name	5' full?[b]	mRNA length[c]	position of 5'[d]
ztv60606	rat MG-160 g-protein	rnu08136	n	5519	-1961
ztv60495	DNA topoisomerase 2	humtopii	n	4792	-2020
ztv60495	DNA topoisomerase 2	humtopii	n	4792	-2020
ztv60326	TEGT	hstegt	n	2600	-771
ztv60212	elongation factor 1-alpha		n	1703	-493
ztv60320	M.mus. E25 homolog	muse25a	n	1635	-22
ztv60431	26S protease S4 regulatory subunit	hum26spsiv	n	1599	-24
ztv60548	Ah receptor	humahre	y	3317	+62
ztv60208	Human protein tyrosine kinase	hsu02680	y	3000	+13
ztv60351	peptide binding protein	humpbp	y	2845	0
ztv60389	Tra1	hstra1	y	2780	+2
ztv60378	TEGT	hstegt	y	2600	+44
ztv60424	hsp90	humhsp90	y	2543	+12
ztv60614	hsc70	hshsc70	y	2403	0
ztv60540	pre B cell enhancing factor	hspbef	y	2376	+75
ztv60542	pre B cell enhancing factor	hspbef	y	2376	+75
ztv60544	pre B cell enhancing factor	hspbef	y	2376	+75
ztv60142	ETS2	humets2pr	y	2269	0
ztv60462	protein phosphatase 1-gamma	hsppp1cc	y	2263	+30
ztv60240	h-sp1 a synaptophysin homolog	hshsp1	y	2130	+8
ztv60647	M-T-D-Cyclohydrolase	hsmmtdc	y	2102	+20
ztv60255	cathepsin B	humctsb	y	2002	+32
ztv60011	ief7442	hsief7442	y	1943	0
ztv60427	ATP synthetase	humatpsas	y	1857	+2
ztv60152	elongation factor 1-alpha		y	1703	0
ztv60301	elongation factor 1-alpha		y	1703	0

ztv60343	elongation factor 1-alpha		y	1703	0
ztv60589	elongation factor 1-alpha		y	1703	0
ztv60629	elongation factor 1-alpha		y	1703	0
ztv60231	ribosomal protein L4	humrsp	y	1418	0
ztv60310	ribosomal protein L4	humrsp	y	1418	0
ztv60278	ferritin	humferrith	y	1198	+1
ztv60220	translationally controled tumor pro.	hstump	y	830	+17
ztv60221	p21 homolog	hsmp21hom	y	819	+26
ztv60155	coatomer	btzcop	y	676	+4

Clones from 5'-end-enriched cDNA library

Clone number	Homology[a]	locus name	5' full?[b]	mRNA length[c]	position of 5'[d]
zrv60084	beta-adaptin	humbadpta	n	5701	-53
zrv60089	ubiquitin activating enzyme E1	humubiqaa	n	3419	-1827
zrv60107	Tra-1	hstra1	n	2780	-59
zrv60074	hsc70	hshsc70	n	2403	-1089
zrv60171	transcription factor SL1	humtfsl1c	n	1703	-44
zrv60002	OXA1HS	hsoxa1hs	n	1551	-82
zrv60012	pre-mRNA splicing factor	humpsf82	n	657	-32
zrv60057	mitochondria		n		
zrv60265	rat dynein heavy chain[e]	ratdyneinc	y	14279	+5[f]
zrv60190	giantin	hsmac	y	10300	-15[f]
zrv60225	PDGF receptor	humpdgfra	y	5427	+96
zrv60278	transferrin receptor	humtrfr	y	5010	0
zrv60281	transferrin receptor	humtrfr	y	5010	0
zrv60141	helix-loop-helix protein	humheb	y	4126	+35
zrv60147	helix-loop-helix protein	humheb	y	4126	+41
zrv60131	2-oxoglutarate dehydrogenase	hum2ogdh	y	4122	0
zrv60227	pIG-A	humpiga	y	3589	0
zrv60064	PM/SCI 100kd nucleolar protein	humaua	y	2834	+43

zrv60092	p1 protein	hsp1h	y	2575	+33
zrv60126	hsp90	humhsp90	y	2543	+12
zrv60292	thyroid hormone binding protein p55	humthbp	y	2514	+29
zrv60161	nucleolin	humnucleo	y	2504	0
zrv60088	nucleolin	humnucleo	y	2504	0
zrv60095	hsc70	hshsc70	y	2403	+3
zrv60269	nuclear protein p68	hsnp68m	y	2323	0
zrv60164	nuclear protein p68	hsnp68m	y	2323	+5
zrv60029	49kd protein	hum49kda	y	2201	0
zrv60065	glucose regulated protein Bip	humgrp78	y	2182	0
zrv60228	AML-2	hsaml2	y	1806	+114
zrv60055	beta-actin	hsactb	y	1802	+12
zrv60136	beta-tubulin	humtbbm40	y	1800	0
zrv60236	beta-tubulin	humtbbm40	y	1800	0
zrv60019	beta-tubulin	humtbbm40	y	1800	0
zrv60035	phosphoglycerate kinase	hspgk1	y	1767	+7
zrv60119	elongation factor 1-alpha		y	1703	0
zrv60137	elongation factor 1-alpha		y	1703	0
zrv60232	elongation factor 1-alpha		y	1703	0
zrv60205	hnRNP A2	humrnpa2a	y	1700	+14
zrv60051	hPGI	humhpgi	y	1685	+26
zrv60037	rabbit progesterone induced protein	rabepip	y	1600	+2
zrv60004	alpha-tublin	humtubak	y	1596	+28
zrv60222	alpha-tubulin	humtubak	y	1596	+28
zrv60259	alpha-tubulin	humtubak	y	1596	+ 34
zrv60230	endonexin2	humenn	y	1592	+5
zrv60255	basigin	humbsg	y	1475	+10
zrv60183	ref-1	s43127	y	1402	+13
zrv60026	elongation factor 1-gamma	hsef1gmr	y	1401	+18
zrv60193	eIF-4A1	hum4ai	y	1383	0
zrv60117	Mus. musculus HMG-1	u00431	y	1308	0

zrv60150	nucleophosmin	humnpm	y	1296	+5
zrv60036	GAPDH	humgapdh	y	1268	+14
zrv60097	lactate dehydrogenase A	hsldhar	y	1260	+1
zrv60066	lactose anhydrogenase B	hsldhbr	y	1260	+45
zrv60184	GTP binding protein G25k	humgpg25k	y	1175	+18
zrv60139	ribosomal protein L15	hsu14966	y	987	+44
zrv60180	tax response element binding protein	humtr107	y?	921	-3
zrv60253	actin capping protein alpha-subunit	hsu03269	y	918	+3
zrv60043	aldehyde dehydrogenase	humaldc01	y	799	+1
zrv60185	ribosomal protein S18	hsrps18	y	549	+3
zrv60142	acidic ribosomal protein P2	humpparp2	y	460	+3
zrv60129	ribosomal protein S15a	hsrps15a	y	450	+31
zrv60207	ribosomal protein S15a	hsrps15a	y	450	+31

[a] The homology search was performed using BLASTN[40] or FASTA [41] against the GenBank entries.

[b] The cDNA clones that were categorized as "full" or "near-full" were scored as "y".

[c] The length is that in the GenBank data. The actual length of cDNA from the full length-enriched cDNA library is similar to these numbers. The actual length of the cDNA from the 5'-end-enriched cDNA library is usually much shorter.

[d] The position of the 5'-end is indicated by the number of nucleotides relative to the known 5' ends Plus sign means our clones were longer than the database sequences.

[e] Also see the text.

III) Evaluation of the content of the cDNA clones that have the mRNA start site in these libraries.

To evaluate the content of cDNA clones that have the mRNA start site, I first analyzed the 5'-end sequences of cDNA clones that came from an abundantly expressed and well-studied human polypeptide chain elongation factor 1 α (EF-1 α) gene. I found 6 cDNA clones of EF-1 α from the full length-enriched cDNA library and 3 clones from the 5'-end-enriched cDNA library. The 5' sequences of these clones are shown in Fig.IIIA-3. The EF-1 α mRNA is 1753bp long, starts with the sequence CTTTTT and its exon 1 consists of 32 bases[45]. Most of the clones had the sequence CTTTTT immediately downstream to the 5'-oligo sequence. They also showed microheterogeneity in the number of T residues after the first C residue and/or the lack of the C residue, which were known from the previous studies[35, 46, 47]. The clone ztv60212 seemed to be non-"full-length", because it totally lacks the CTTTTT sequence. Other clones seemed to have the mRNA start site, because they had the T stretch or retained the first C residue. Thus, 8 clones out of 9 (89%) had the mRNA start site.

In contrast, I found more than 5000 entries that matched the EF-1 α mRNA within the EST database, the dbEST. Of those clones, one clone started with CTTTTT (locus name: SSC26X5) and 251 clones had a part of the exon 1 sequence. Thus, less than 0.02% of the cDNA clones sequenced for EST work had the mRNA start site, and about 5% had exon 1 of EF-1 α. Since the cDNA libraries used for EST work were made mostly by conventional methods, this is a good estimation for the content of the "full-length" cDNA clones of EF-1 α in the conventional cDNA library. Thus, our cDNA libraries seemed to have made a significant improvement in the content of "full-length" cDNA clones at least for the EF-1 α gene.

I then evaluated the content of the cDNA clones with the mRNA start site for other genes. Since the exact mRNA start site is not determined for

Genomic sequence

TATATAAGTGCAGTAGTCGCCGTGAACGTTCTTTTTCGCAACGGGTTTGCCGCCAGAACACA

TATA box **mRNA start site**

Clone name	5'-oligo sequence	cDNA sequence
ztv60212	GGCCTACTGG	GGGTTTGCCGCCAGAACACA
ztv60152	GGCCTACTGG	CTTCGCAACGGGTTTGCCGCCAGAACACA
ztv60301	GGCCTACTGG	CTTTTTCGCAACGGGTTTGCCGCCAGAACACA
ztv60343	GGCCTACTGG	CTTTTTCGCAACGGGTTTGCCGCCAGAACACA
ztv60589	GGCCTACTGG	CTTTCGCAACGGGTTTGCCGCCAGAACACA
ztv60629	GGCCTACTGG	TTTTTTCGCAACGGGTTTGCCGCCAGAACACA
zrv60137	GGCCTACTGG	CTTTTTCGCAACGGGTTTGCCGCCAGAACACA
zrv60119	GGCCTACTGG	CTTTTCGCAACGGGTTTGCCGCCAGAACACA
zrv60232	GGCCTACTGG	TTTTTCGCAACGGGTTTGCCGCCAGAACACA

Fig.IIIA-3 The 5'-end sequences of the clones that matched with the EF-1α mRNA.
The 5' sequences of all the cDNA clones that matched with the EF-1α mRNA are shown. Clones whose name starts with ztv are from the full length-enriched cDNA library and clones with zrv from the 5'-end-enriched cDNA library (see also Table IIIA-2). The sequences derived from the 5'-oligo are aligned with each other. The sequences corresponding to the EF-1α gene were aligned along with the genomic sequence shown above. Gaps shown between the sequence derived from the 5'-oligo and the sequence corresponding to the EF-1α mRNA do not exist in the real sequence. Genomic sequence is corresponding to the EF-1α promoter region. TATA box and the mRNA start site are marked blue and red, respectively.

many genes, a cDNA clone was tentatively scored as "full" if the 5'-end of the clone matched with the putative start site in the promoter data or had the same or longer 5'-end than the "complete cDNA" data in the database. A cDNA clone that had shorter 5'-end but still contained the predicted translation initiator ATG was scored as "near-full". Using these criteria, 28 clones out of 35 (80%) from the full length-enriched cDNA library and 51 out of 62 (82%) from the 5'-end-enriched cDNA library were scored as "full/near-full" (Tables IIIA-1 and IIIA-2). This is in good agreement with the EF-1 α result. Thus, I concluded that the content of cDNA clones with the mRNA start site was around 80% for both libraries.

As described in section I, 100% of the cDNA clones had the 5'-oligo. This indicates that the cloning step itself is 100% specific. Thus, the content of the full-length cDNA clones noted above (about 80%) reflects the specificity of the "Oligo-capping". At present, I do not know why the "Oligo-capping" is not 100% cap specific. Several possibilities include the escape of RNA breakdown products from the BAP treatment and/or RNA breakage during the TAP and the RNA ligation reaction.

IV) cDNA clones of long mRNAs in the 5'-end-enriched cDNA library.

In general, the length of the cDNA in the cDNA libraries was usually in the range of 1,000 to 4,000bp and clones longer than 4,000bp were rare. This could be a serious problem when isolating the mRNA start site of long mRNA using the full length-enriched cDNA library. I constructed a 5'-end-enriched cDNA library in order to by-pass this problem. As shown in Table IIIA-2, the longest clone among the 28 "full/near-full" clones from the full length-enriched cDNA library was about 3,300bp (ztv60542). In contrast, 7 out of 51 clones from the 5'-end-enriched cDNA library had the start site of mRNA whose length is more than 3,500bp. Thus, the 5'-end-enriched library seemed useful for isolating the start site of long mRNA.

I found 2 cDNA clones whose mRNA are more than 10,000bp long from the 5'-end-enriched cDNA library (Table III-2). One is the cDNA of giantin mRNA (zrv60190). The 5'-end of the cDNA clone is about 50 bases short of the putative mRNA start site determined by the primer extension method[48]. Thus, I scored this clone as "near-full". However, the cDNA still had 115bp of the 5'UTR. Since 80% of the cDNA clones in this library had the mRNA start site, it is possible that the clone might represent the 5'-end of mRNA that has been transcribed from a minor start site.

The other clone, zrv60265, showed strong homology with rat dynein heavy chain mRNA. I found two rat dynein heavy chain sequences in the database. As shown in Fig.IIIA-3, the 5'-end of clone zrv60265 matched with the 5'-end of the 14,279bp rat cDNA data (RATDYNEINC)[49]. The actual insert size of the clone was about 1,500bp. The sequence of the 3'-end of the clone also matched with a sequence around 1,300-1,500bp of the 14,279bp clone (Fig.IIIA-4). Though the mRNA length estimated by the Northern analysis was 16,000 to 16,500bp and 15,500bp data of rat dynein heavy chain mRNA, which has 1,200bp longer 5'UTR, has been reported[50], it is noteworthy that both rat and human cDNA (zrv60265) gave a similar 5'-end. Considering the high content of cDNA clones with the mRNA start site, it is possible that this clone may actually have the mRNA start site of human dynein heavy chain mRNA.

a

```
zrv60265        AGTCTGCGGTGGGCTAACGGACGGTCCGGCTTCCGGCGGCCGTTTCTGTCTCTTGCTGGC
RATDYNEINC      -----GCGG-------AC--ACAG--CGTCTTCTGTCGCCGGTTTCTTTCGCTCACTG-C
                     ****       **  ** *  ** **** * ** * ****** ** **  *** *

zrv60265        TGTCTCNCT--GAA---------TCGCGGCCGCCT-TCTCATCGCTCCTGGAAGG---TC
RATDYNEINC      TCCCTCGTTCCGGAGGTAGCTCTTCTCGGCTCTCTCTCTC-TCTCTTCTCTATCTCTCTC
                *  ***  *  * *         ** ****   ** **** ** ** **  *      **

zrv60265        CCGAGCGCGA---CACCATGTCNGANCCCGGGGGCGGCGGCGGCGAAGACGGCTCGGCCG
RATDYNEINC      CCTTCCGCGGATCCGCCATGTCGGAGACC------GGCGGCGGTGAGGACGGCTCGGCCG
                **   ****    * ******* **  **      ******** ** *************

zrv60265        GATTGGAANTGTCGGCCGTGCANAATGTGGCGGACGTGTCGGTGCTGCANAAGCACCTGC
RATDYNEINC      GCCTGGAGGTGTCCGCGGTGCAGAATGTGGCGGACGTGTCGGTGCTGCAGAAGCACCTTC
                *  ****  **** ** ***** ************************** ******** *

zrv60265        GCAAGCTGGTGCCGCTGCTGCTGGAGGACGGCGGCGAAGCGCCGGCCGCGCTGGAAGCGG
RATDYNEINC      GTAAGCTGGTTCCGCTGCTGCTGGAGGACGGCGGCGACGCGCCGGCTGCGCTGGAGGCGG
                * ******** ************************** ******** ******** ****

zrv60265        TGCTGGAGGAGAAGAGCGCCCTGGAGCAGATGCGCAAGTTCCTTTCGGACCCGCAGGTCC
RATDYNEINC      CGCTGGAGGAGAAGAGCGCCCTGGAGCAGATGCGCAAGTTCCTGTCAGACCCGCAGGTCC
                 ****************************************** ** *************

zrv60265        ACACGGTGCTGGTGGAGCGCTCCACGCTCAAAGTGGACNTCGGTGATNAAGGAGAAGAAG
RATDYNEINC      ACACGGTCCTGGTGGAGCGCTCCACCCTCAAAGAGGACGTTGGTGATGAAGGAGAAGAGG
                ******* ***************** ******* **** * ****** ********** *

zrv60265        AAAAAGAATTCATTTCCT
RATDYNEINC      AGAAAGAATTCATTTCCT
                * ****************
```

b

```
zrv60265        ----------------------------------------CTTTGAAGTTTT--AAACTT
RATDYNEINC      ATGTAGCGTATGAAGAGTTTGAAAAGGTCATGGTGGCTTGCTTCGAAGTCTTCCAGACGT
                                                        *** ***** **  * ** *

zrv60265        GGGATGATGAGTATAA--AACTT-AGGTATTGTTGAGAGAAACCGTCAAAAGAAAAAGGG
RATDYNEINC      GGGATGATGAGTATGAGAAACTCCAAGTGCTGCTGAGGGACATCGTCAAGAGGAAGAGGG
                ************** *  ****  * **  ** **** ** * ****** ** ** ****

zrv60265        AAGAAAATCTGAAGATGGTGTGGCGTATCAACCCTGCCCA-AGGAA-CTGCAGGCCCGCC
RATDYNEINC      AGGAGAACCTGAAGATGGTGTGGCGCATCAACCCTGCTCACAGGAAGCTGCAGGCCCGCC
                * ** ** ***************** *********** ** ***** *************

zrv60265        TTGACCAGATGAGAAAATTTAAC-GCCAGCATGAACAGCTAAGAGCTGTTAT--------
RATDYNEINC      TGGACCAGATGAGGAAGTTCCGCCGCCAGCACGAGCAGCTGAGAGCTGTCATTGTCAGAG
                * *********** ** **   * ******* ** ***** ******** **

zrv60265        ---------CTCAGG
RATDYNEINC      TCCTGCGCCCACAGG
                         * ****
```

Fig.IIIA-4 **Sequence alignment of the 5' and 3' sequences of zrv60265 against the rat dynein heavy chain mRNA sequence.**

The homology was searched against the GenBank using BLASTN. The alignment with one of rat dynein heavy chain mRNA sequence gb:RATDYNEINC is shown. a: Alignment of the 5' sequence of the zrv60265: The sequence of 260 bases just downstream of the 5'-oligo sequence was used for the search. b: Alignment of the 3' sequence of zrv60265: The sequence complementary to the sequence of 195 bases just downstream of the random primer sequence was used for the search.

V) Comparison with other methods for cDNA library construction.

On many occasions, reverse-transcriptase can not make a full cDNA copy of mRNA but stops in the middle leaving an incomplete copy. Thus, cDNA libraries made by the conventional methods such as the Gubler-Hoffman method[51] and the Okayama-Berg method[52] contain many incomplete cDNA copies of mRNA. The essence of our method described above is to isolate the "full-length" cDNA from the majority of the incomplete cDNA based on the "Oligo-capping" and PCR.

Kato et al. combined the Okayama-Berg method and an "Oligo-capping" method, which uses a DNA-RNA chimera as 5'-oligo[46]. The Okayama-Berg method requires delicate use of enzymes (especially that of terminal transferase). Furthermore, it may be difficult to make a 5'-end-enriched type library by the Okayama-Berg method, because it uses vector primer for the first strand synthesis. Our PCR based method is relatively simple and the same procedure can be used for the construction of both the full-length library and the 5'-end library.

Edery et al. made a full length-enriched and a 5'-end-enriched cDNA library based on their "Cap Retention Procedure"[53]. Recently, Carninci et al. also made a full length-enriched cDNA library using their "CAP Trapper" method[47]. Both methods use the cap dependent retention (or trapping) to some solid supports for the selection of "full-length" cDNA. Only mRNA-cDNA hybrids whose cDNA extended to the cap were retained (or trapped) to solid supports and then can be selectively cloned. This selection principle can be modified for our "Oligo-capping" based method using biotinylated 5'-oligo.

Our PCR based method also has an advantage in the selective amplification of cDNA that has both 5'-oligo and adapter primers. Initially PCR had a high mutation rate and difficulty in amplifying long DNA. However, introduction of the long PCR method greatly improved both the fidelity of the reaction and the length of PCR products[54), 55]. The majority of the cDNA in our full length-enriched library ranges from 1000 to 4000bp, similar to the inserts of most cDNA libraries made by non-PCR methods. PCR has other drawbacks, such as bias in the profile due to the difference in PCR efficiency among cDNA clones, and the generation of a high number of sister clones. At present, I do not have enough data to assess the extent of these problems. Judging from the clones listed in TableIIIA-2, the libraries seemed divergent enough to use for the generation of 5' ESTs.

IIIB. RESULTS AND DISCUSSIONS (II)

~Analyses of the "Oligo-capped" cDNA libraries.

I) Construction and the large-scale sequencing of the "Oligo-capped" cDNA libraries.

The large-scale analyses of the 5'-ends of mRNA were attempted using full length-enriched and 5'-end-enriched cDNA libraries ("Oligo-capped" cDNA libraries). The "Oligo-capped" cDNA libraries were constructed from about 30 kinds of human tissues and cultured cells (Table IIIB-1). the cDNA clones were randomly selected from the "Oligo-capped" cDNA libraries and the 5'-end sequences of about 10,000 clones were determined in total. The most intensively sequenced were the full length-enriched cDNA libraries of ileum and colon, and the 5'-end-enriched cDNA library of SK-N-MC. (3150, 3374, 1660 clones respectively).

In order to evaluate how many fractions of the cDNA clones from each library contains the full-length clones, I employed the same criteria shown in the chapter IIIA. I selected the clones whose sequence matched with function-known genes. Sequence similarity tests showed about 50% matched with known genes. Among them, in general, about 50-60% had the same or longer 5'-ends than the "complete cDNA" data in the database ("full" clones). About 5-10% had shorter 5'-ends but still contained the predicted translation initiator ATG codon ("near-full" clones). The others lacked the initiator ATG ("not-full" clones) (Table IIIB-1).

A database, named DYNACLUST, was constructed with the 5'-end sequence data. In the database, about 1,200 species of full/near-full cDNA 5'-ends for the function-known genes have been accumulated so far. The average length of the corresponding mRNA was 2.0 kb. Though the majority of the mRNA were less than 3 kb long, I found the 5'-ends of mRNA whose length is more than 5 kb among the entries from the 5'-end-enriched cDNA library (Fig.IIIB-1).

Table IIIB-1 List of the "Oligo-Capped" cDNA libraries constructed from human tissues and cultured cells.

Origin	library-type[a]	Library Name	Ave.Ins.Size (kb)[b]	known#	full/near-full#	not full#	full%	EST#	NEW/Others#
Adipocyte (A)	dT	FATb	2.0	13	9	4	69	3	6
Adipocyte (B)	dT	HsfA	1.7	15	8	7	53	7	5
Adipocyte (C)	dT	fata	1.4	16	14	2	88	2	8
Embryonal Brain	dT	hemb	1.4	18	10	8	56	7	10
Whole Embryo	dT	HemB	1.5	21	14	7	67	7	17
Colon	dT	ColF	1.5	1923	886	1037	46	533	918
Colon Mucosa	dT	HgtA	1.9	6	6	0	100	8	6
Ileum	dT	kaia	2.0	937	568	369	61	546	1667
Duodenum	dT	JYUf	1.6	21	14	7	67	8	18
Liver	dT	HlvA	1.6	15	12	3	80	3	3
Mammary gland	dT	NYUb	1.9	8	5	3	63	6	12
Uterus	dT	WmbA	1.5	21	11	10	52	4	13
Lymphonode	dT	htlb	2.1	11	7	4	64	6	2
Neuro Blastoma (A)	dT	NblC	2.3	9	4	5	44	10	12
Neuro Blastoma (B)	dT	NblG	1.3	14	9	5	64	7	17
Neuro Blastoma (C)	dT	NblL	1.8	37	26	11	70	12	14
Cao2	dT	caoa	2.2	5	5	0	100	2	9
HepG2	dT	HEPa	2.2	8	5	3	63	0	1
JCRD	dT	jcrd	1.7	18	11	7	61	10	2
KATO-III	dT	kt3a	1.7	23	19	4	83	5	3
MKN28	dT	MKNa	2.0	10	7	3	70	2	1
NT-2	dT	w	1.7	293	184	109	63	111	59
NT-2 (differentiated)	dT	ntra	2.2	18	11	7	61	4	6
Y79	dT	y79a	2.2	33	25	8	76	13	6
SK-N-MC	dT	Ziv6	2.0	35	28	7	80	15	34
SK-N-MC	dR	Zrv6	2.4	1087	819	268	75	398	123

[a] dT represents the full length-enriched and dR the 5'-end-enriched cDNA library.

[b] The length used to calculate the average length is those in the GenBank data. The actual length of cDNAsfrom the full length-enriched cDNA library is similar to these numbers. The actual length of the cDNA from the 5'-end-enriched cDNA library is usually much shorter

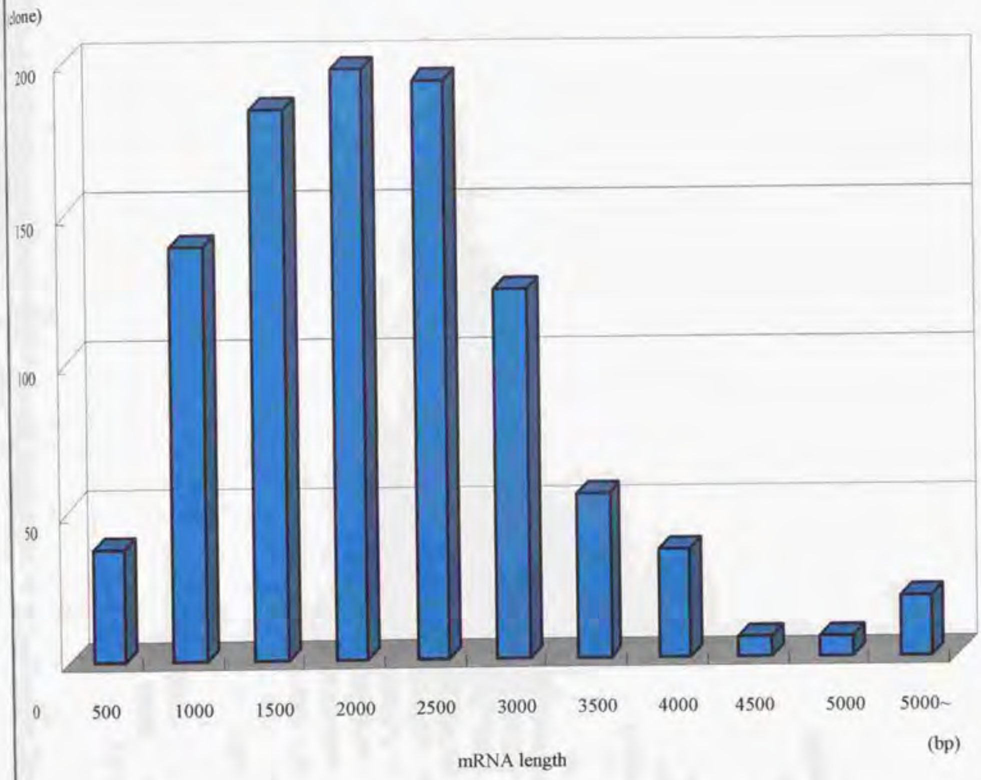


Fig.IIIB-1 **Length distribution of the full/near-full clones in the DYNACLUST.**

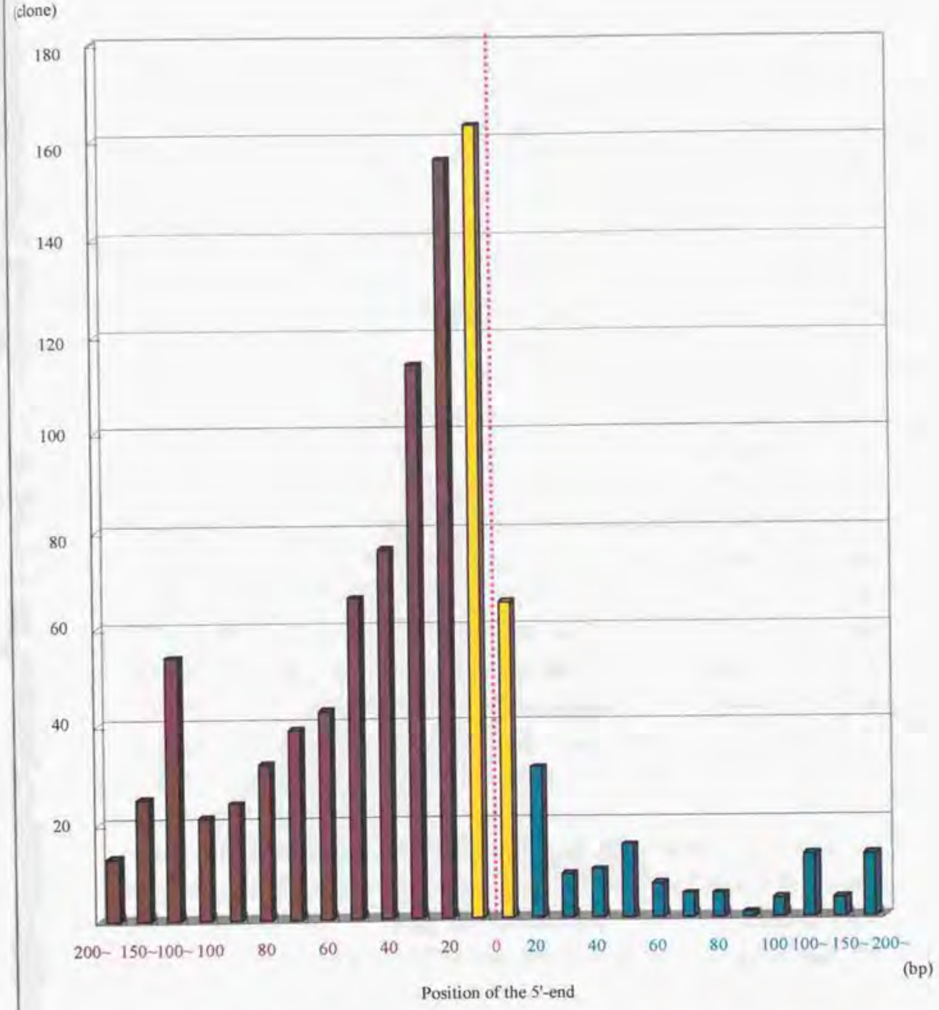


Fig.IIIB-2 **Comparison of the 5'-end of the "Oligo-Capped" cDNA clones with the GenBank entries.**

The position of the 5'-ends were compared between the "Oligo-Capped" full/near-full clones and the GenBank entries. The X-axes shows the position of the 5'-end of the "Oligo-Capped" clone relative to the known 5'-endof the GenBank entries. The red bars represent our clones had longer 5' -end. The blue bars represent our clones were shorter. Yellow bars represent the differences were within 10 bp.

The mRNA start site has not been determined for many of these genes. As shown in Fig.IIIB-2, our clones had, in average, 45bp longer 5'-ends than the Genbank entries. The 5'-ends of our clones can represent the mRNA start sites that have not been reported. Considering the first exon is a short exon in many cases, many of the Genbank entries may miss the first exons. It could be a serious drawback when an identification of the promoter region is attempted with the current database entries.

II) Comparison of the database between the DYNACLUST and the dbEST.

Using our cDNA clones that matched known genes, I compared the DYNACLUST with the dbEST about the content of full/near-full clones. I selected the genes that I found among the DYNACLUST entries more than 10 times (Table IIIB-2). I compared the content of full/near-full clones for these genes. Figure IIIB-3 shows the result of the comparison for the polypeptide chain elongation factor 2 (EF-2). According to our criteria, 9 out of 13 clones (69%) from the DYNACLUST entries were full/near-full clones. I found 140 entries for EF-2 in the dbEST but none of them hit within the 50bp of the 5'-ends of the full/near-full clones. I made the same comparison for the other genes listed in Table IIIB-2. In most of the cases, the difference was significant. The DYNACLUST could complement the dbEST with the sequence around the mRNA start site.

It is also intriguing that the content of full/near-full clones varied between mRNA species. Among the genes listed in the Table IIIB-2, some genes are known to have long mRNA half-life[26), 27), 30)]. Considering these genes had high content of full/near-full clones, it is possible that the difference reflects the stability of the mRNA molecule. The error of the cap-replacement with the "Oligo-capping" might increase if only a small fraction of mRNA remained intact.

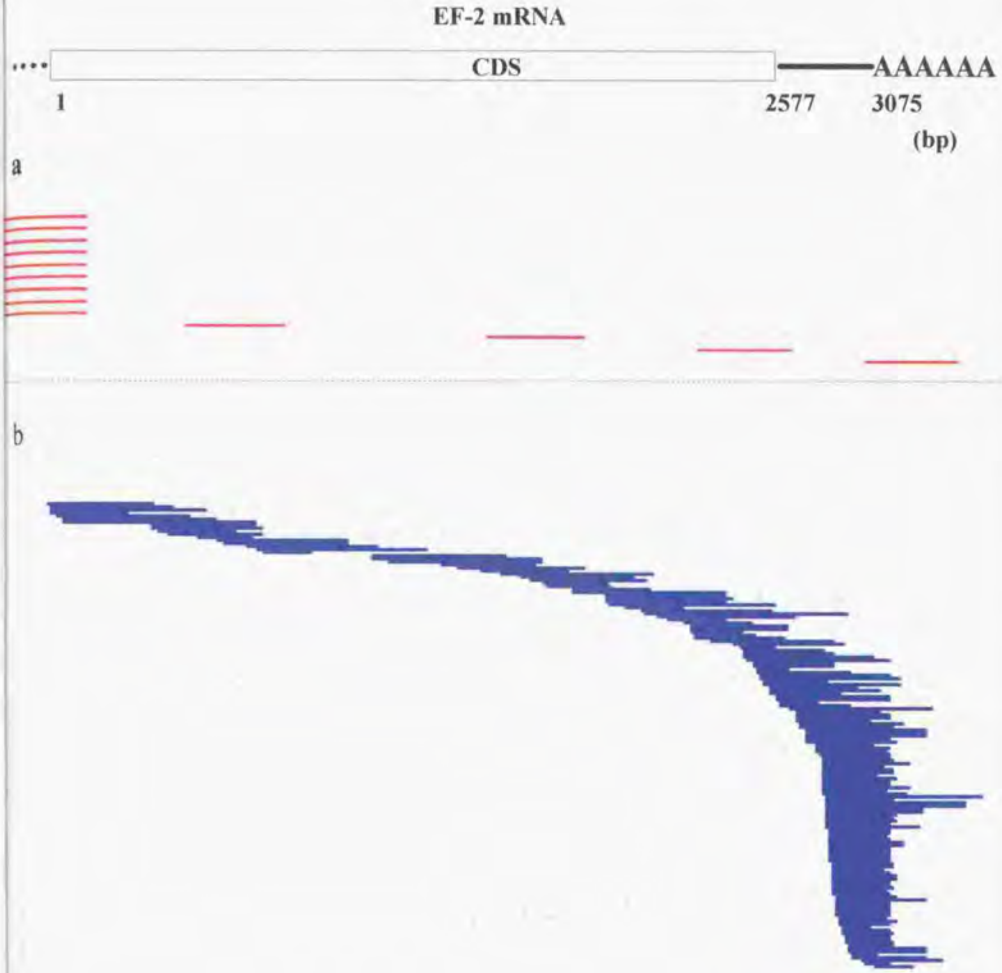


Fig.IIIB-3 **Comparison of the 5'-ends of EF-2 cDNA between the "Oligo-Capped" clones and the dbEST entries.**

The distribution was compared between the "Oligo-Capped" cDNA clones and the dbEST entries along the EF-2 mRNA. The sequences derived from the 5'-ends of "Oligo-Capped" cDNA clones (a) and the dbEST entries were aligned along with the EF-2 complete cDNA sequence shown above. The length of each bar corresponds to the reported sequence length.

Table IIIB-2

List of cDNA clones that were most frequently isolated from the "Oligo-Capped" cDNA libraries.

cDNA clone[a]	mRNA length[b]	Clone#[c]	full#[d]	%[e]	EST#[f]	fullEST#[g]	%[h]	TATA[i]	Start Site[j]
EF-1α	1752	165	151	92	5076	251	4.9	y	y
β-actin	1802	74	71	96	3171	0	0.0	y	y
polyubiquitin	2192	62	42	68	1805	39	2.2	y	y
HLA-Cw5	1089	45	42	93	1176	10	0.9	n	y
hsp70	2177[k]	35	33	94	702	30	4.3	y	y
alpha enolase	1755	35	33	94	1173	6	0.5	n	y
glucose transporter type 3	3915	32	32	100	ND[l]	11	ND[l]	n	n
CGM2	2292	31	7	23	252	1	0.4	n	n
α-tubulin	1596	30	29	97	2367	213	9.0	n	y
apolipoprotein B-100	14121	30	1	3.3	112	0	0.0	n	n
hnRNP G	1894	29	29	100	72	3	4.2	n	n
β-tubulin (clone m40)	2600/1800	24	10	42	1288	90	7.0	n	n
nuclear matrix protein 55	1791	24	16	67	60	8	13.3	n	y
nucleolin	2531	23	17	74	136	1	0.7	n	y
ferritin L chain	822	23	23	100	1927	250	13.0	y	y
tra1	2780	22	13	59	257	8	3.1	n	n
lamin C	2404	21	19	90	283	1	<0.1	y	y
heat shock protein 70-1	2417	21	20	95	344	5	1.5	n	y
amyloid precursor-like protein 2	2366	21	21	100	203	0	0.0	n	n
HRPL4	1418	20	18	90	802	99	12.3	n	n
vimentin	1766	20	20	100	607	1	0.2	y	y
TH binding protein p55	2514	20	16	80	249	0	0.0	y	y
PPHα	2920	20	17	85	9	0	0.0	n	n
epithelin	2178	20	20	100	176	1	0.6	n	y

Table IIIB-2 (*continued*)

cDNA clone[a]	mRNA length[b]	Clone#[c]	full#[d]	%[e]	EST#[f]	fullEST#[g]	%[h]	TATA[i]	Start Site[j]
TEGT	2600	18	8	44	553	10	1.8	n	n
hsp86	2912	18	18	100	948	22	2.3	y	y
MXR7	2300	18	16	89	142	0	0.0	n	y
hnRNPcore protein A1	1747	17	17	100	631	79	12.5	n	y
ftp-3	2220	17	14	82	294	1	0.3	n	n
mitochondrial matrix protein P1	2227	16	16	100	225	17	7.6	n	y
progesterone-induced protein	2158[m]	16	11	69	35	5	14.3	n	n
HLA-DR	1304	16	16	100	1077	103	9.6	y	y
HHCPA78 homolog	2704	16	16	100	491	17	3.5	n	n
lupus p70 (Ku) autoantigen	2123	15	15	100	287	23	8.0	n	n
lactate dehydrogenase-A	1661	15	15	100	262	18	6.9	n	y
ATP synthase alpha subunit	1883	15	14	93	247	4	1.6	n	y
PAP-H=pancreatitis-associated	797	15	14	93	22	1	4.5	n	y
serum albumin (HSA)	2055	15	15	100	2060	0	0.0	y	y
OS-9 precurosor mRNA	2736	14	13	93	147	4	2.7	n	y
Ig rearranged H-chain	ND[m]	ND[n]	ND[n]	ND[n]	ND[n]	ND[n]	ND[n]	n	n
inter-alpha-trypsin inhibitor	3089	14	8	57	85	0	0.0	n	n
EF-2	3075	13	9	69	140	0	0.0	n	n
glutaminyl-tRNA synthetase	2437	12	10	83	182	4	2.2	n	n
ADP/ATP carrier protein	1228	12	11	92	407	18	4.4	y	y
collagen binding protein 2	2047	12	12	100	222	9	4.1	n	n
4F2 antigen heavy chain	2304	12	11	92	261	5	1.9	n	y
thymosin β-10	453	12	12	100	364	87	23.9	y	y
SKB1Hs	1996	12	11	92	75	3	4.0	n	n
selenoprotein P	2038	12	8	67	199	19	9.5	y	y

Table IIIB-2 (*continued*)

cDNA clone[a]	mRNA length[b]	Clone#[c]	full#[d]	%[e]	EST#[f]	fullEST#[g]	%[h]	TATA[i]	Start Site[j]
ferritin heavy chain	1198	12	12	100	1661	81	4.9	y	y
cytokeratin20	1267	12	12	100	336	0	0.0	n	y
amiloride-binding protein	2473	12	9	75	62	0	0.0	n	y
acidic ribosomal phosphoprotein P0	1097	11	10	91	1819	250	13.7	n	n
M2-type pyruvate kinase	2287	11	10	91	796	45	5.7	n	y
monocarboxylate transporter 1	2578	11	11	100	45	5	11.1	n	n
initiation factor 4B	3878	11	8	73	369	9	2.4	n	n
β-2 microglobulin	433	11	11	100	187	76	40.6	n	y
lysosomal acid lipase	2626	11	7	64	74	0	0.0	n	y
Wilm's tumor-related protein	744	11	10	91	1246	226	18.1	n	y
selenophosphate synthetase 2	2253	11	11	100	98	5	5.1	n	n
Cctg	1901	10	10	100	253	4	1.6	n	n
KIAA0174	2348	10	8	80	139	1	0.7	n	n
HRPS20	505	10	10	100	1041	250	24.0	n	n
nucleobindin	1650	10	6	60	32	0	0.0	n	n
GTP-binding protein G25K	1175	10	9	90	147	8	5.4	n	n
serine/threonine protein kinase	2370	10	2	20	159	4	2.5	n	y
colon mucosa-associated (DRA)	2881	10	9	90	3	0	0.0	n	y
MTP	3224	10	6	60	17	0	0.0	y	y
neuroleukin	1987	10	10	100	103	0	0.0	n	y
IFN-inducible γ2 protein	2608	10	7	70	153	7	4.6	y	y
β-tubulin (clone B3T)	1648	9	9	100	1014	1	0.1	y	y
KIAA0064	2043	8	8	100	121	6	5.0	n	n

[a] The cDNA clones that were isolated from the "oligo-capped" cDNA libraries more than 10 times were listed in order of their redundancy.

[b] The length is those in the GenBank data. The actual length of cDNAs from the full length-enriched cDNA library is similar to these numbers. The actual length of the cDNAs from the 5'-end-enriched cDNA library is usually much less.

Table IIIB-2 (*continued*)

[c] The number of clones that were isolated from the "Oligo-Capped" cDNA libraries.

[d] The number of "full/near-full " clones that were isolated from the "Oligo-Capped" cDNA libraries.

[e] The frequency at which the corresponding cDNA was isolates as a "full/near-full" clone from the "Oligo-Capped" cDNA libraries.

[f] The number human EST entries that showed the sequence simirality (P<0.0001) against the complete cDNA sequence.

[g] The number human EST entries that showed the sequence simirality (P<0.0001) against the 50bp of the 5'-end sequence of the "full/near-full" clone.

[h] The frequency at which the corresponding cDNA was isolates as a "full/near-full" clone from the human EST database.

[i] The gene reported to have TATA box at its promoter region were scored as "y".

[j] The gene whose mRNA start site that were reported were scored as "y".

[k] The data of C.aethiops hsp70 mRNA.

[l] The sequence similarity test could not be performed due to the internal *Alu* sequence.

[m] The data of Rabbit endometrial progesterone-induced protein (EPIP) mRNA.

[n] The cDNA sequences were highly divergent.

III) The sequence analysis around the transcription start site.

The most frequently isolated gene from the "Oligo-capped" cDNA libraries was EF-1α (Table IIIB-2). I found 165 full/near-full EF-1α cDNA among the DYNACLUST entries. Again, I checked the microheterogeneity of the 5'-end of EF-1α mRNA as shown in the chapter IIIA. Figure IIIB-4 shows the change in the number of T residues following the start site C. It varied from 2 to 12. The majority of the EF-1α clones contained five T residues, which is identical to its genome sequence, but a certain fraction of our clones consisted of those whose number of T residues was not identical to its genomic sequence.

Using the redundant clones for other genes, the 5'-ends of the clones were compared with each other. The exact 5'-ends slightly differed between the clones for many genes. I selected the genes whose transcription machinery was well-studied among the genes listed in Table IIIB-2. For 35 genes in the list, the promoter sequence was available. Among them, 17 genes were reported to contain the TATA-like element in their promoter region. The 5'-ends of our clones were mapped onto the promoter sequence for each gene. Figure IIIB-5A shows the results of the mapping for ferritin heavy chain and lysosomal acid lipase (LAL). Ferritin heavy chain is a TATA-containing gene and LAL is a TATA-less gene. Figure IIIB-5B shows the distribution of mRNA start sites for other genes. The color intensity of the red boxes reflects the rate at which the corresponding base was used as a transcription start site. The mRNA start sites of the TATA-containing genes seemed restricted in a small area. Compared with that, the mRNA start sites of the TATA-less genes seemed scattering over a relatively wide area. This feature may reflect the difference in the machinery that regulates the transcription initiation between TATA-containing and TATA-less genes.

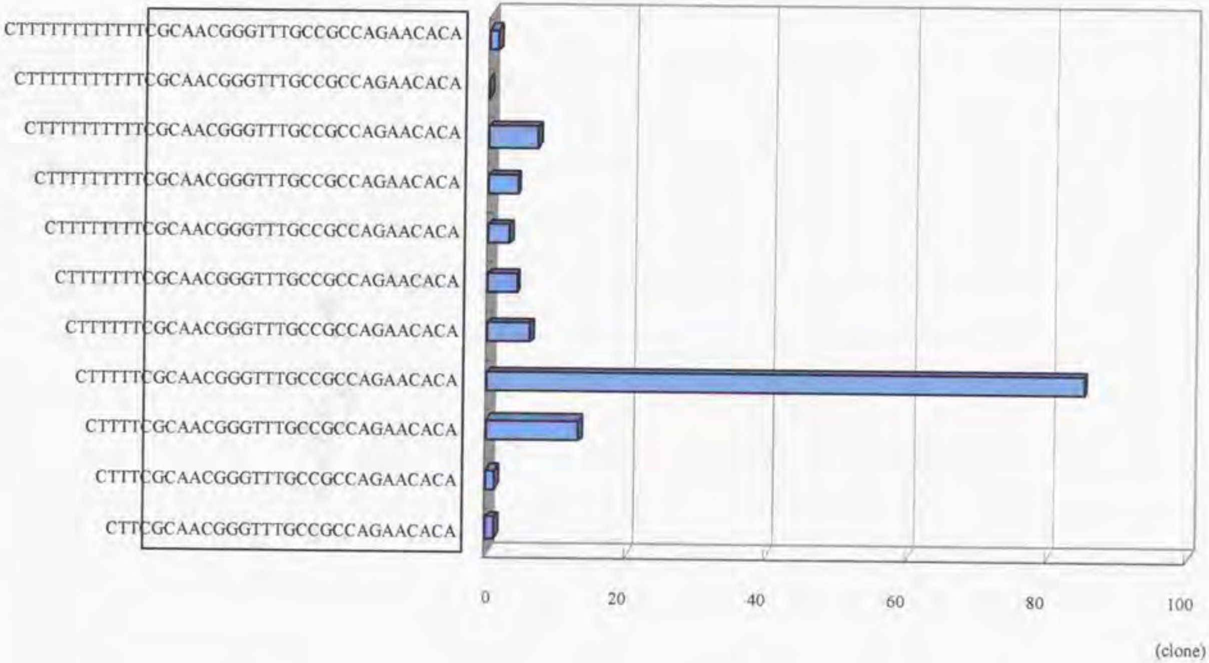


Fig.IIIB-4 **Statistical analysis of the microheterogeneity of EF-1α 5'-ends.**

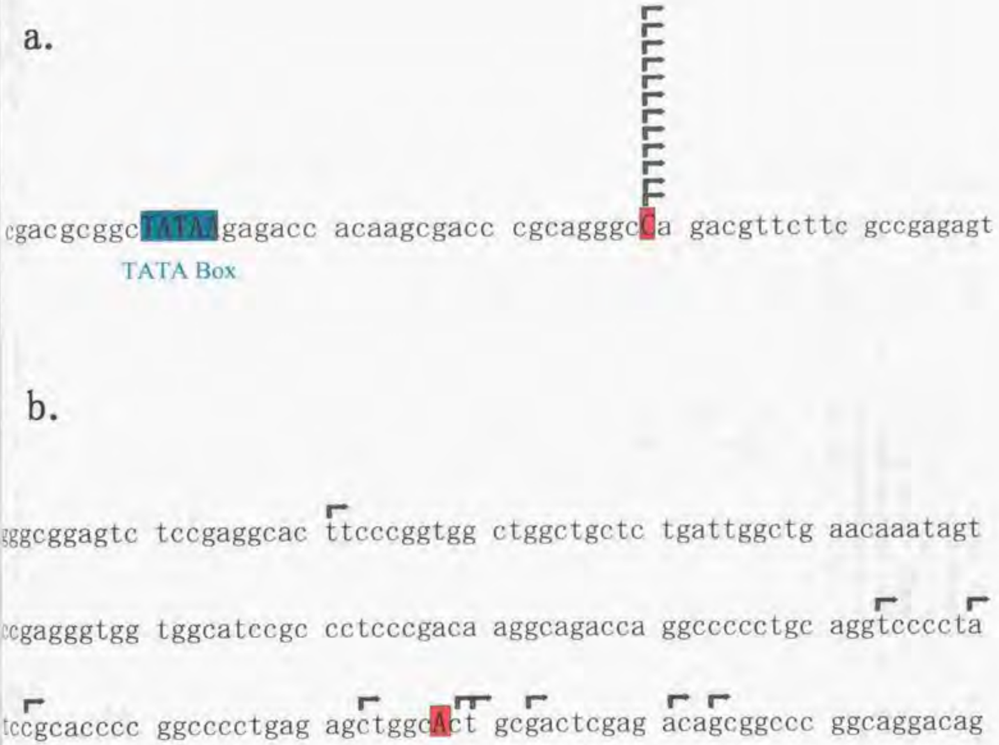

a.
cgacgcggcTATAAgagacc acaagcgacc cgcagggcCa gacgttcttc gccgagagt
TATA Box
b.
gggcggagtc tccgaggcac ttcccggtgg ctggctgctc tgattggctg aacaaatagt
ccgagggtgg tggcatccgc cctcccgaca aggcagacca ggccccctgc aggtccccta
tccgcacccc ggcccctgag agctggcAct gcgactcgag acagcggccc ggcaggacag

Fig.IIIB-4 Distribution of mRNA startsites.
The 5' sequences derived from the "oligo-capped" cDNA clones were aligned against the genomic sequence of ferritin heavy chain (a) and lysosomal acid lipase (b). Each arrow represents the 5'-end of each clone. The predicted transcription start site and TATA box are marked with red and blue boxes, respectively.

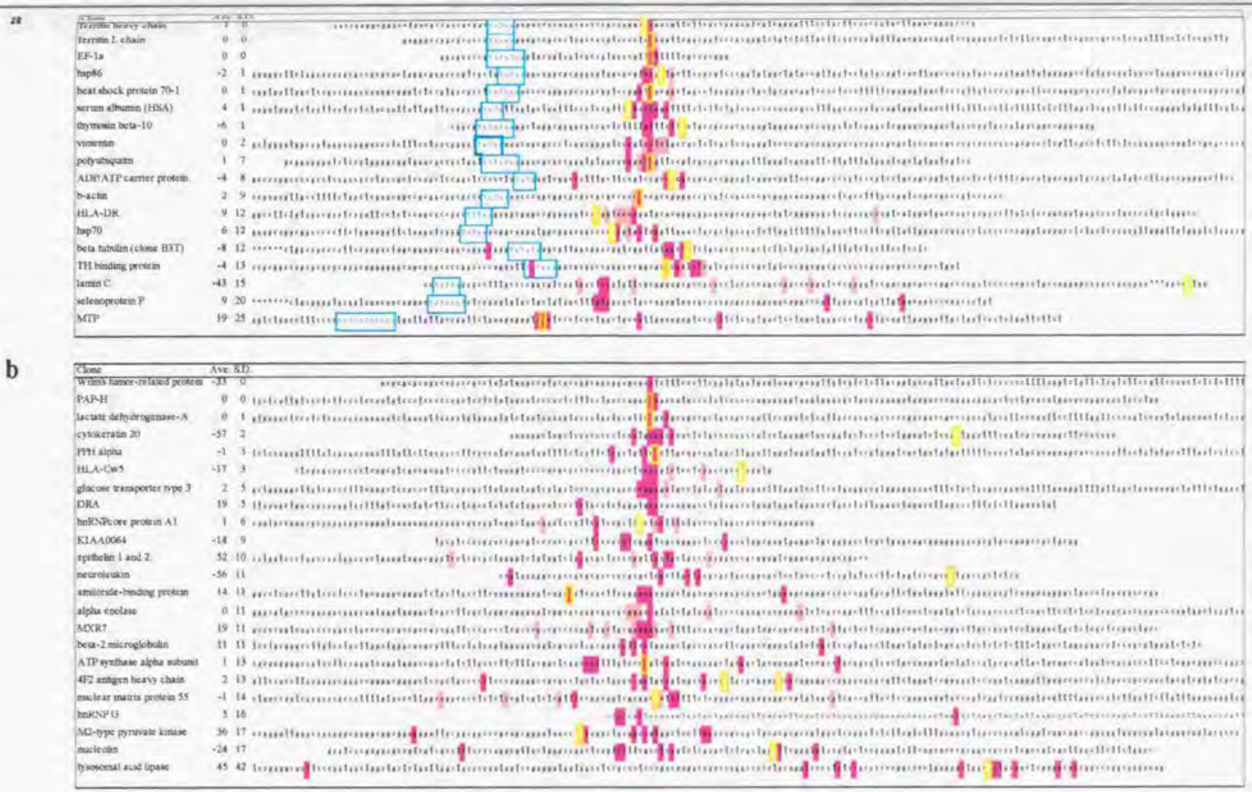


Fig.IIIB-5B Distribution of the mRNA start sites for TATA-containing/TATA-less genes.

The 5' sequences derived from the full/near-full clones of TATA-containing (a) and TATA-less (b) genes listed in Table IIIB-2 were aligned against the genomic sequence. The color intensity of red boxes reflects the rate at which the corresponding base was assigned as an mRNA start site. Previously reported mRNA start sites and TATA boxes are marked with yellow and blue boxes, respectively. For each clone, the average and the standard deviation were calculated for the relative distance between the reported start site and the observed 5'-ends. The clones were sorted in order of the standard deviation.

IV) Statistical Analysis of the 5' UTR.

Using the 5'-end sequence of our full/near-full clones, the statistical analyses of the 5'UTR were attempted. The 5'-boundary of the 5'UTR, which is the mRNA start site, had been determined for our full/near-full clones by the "Oligo-capping". The 3'-boundary of the 5'UTR is the start site of the protein coding sequence (CDS). As for 1010 mRNA species among our full/near-full clones, the CDS start site was described in the database. Combination of these data made it possible to extract the exact 5'UTR sequences from our database. The length distribution of 5'UTR and its relation with the mRNA length were shown in Fig.IIIB-6A and B. The average length of the 5'UTR was 122bp. There was little correlation between the length of 5'UTR and that of mRNA (the correlation coefficient was 0.27). Regardless of its mRNA length, the length of 5'UTR seemed concentrated under 200bp.

The UTR has been recently reported to play a crucial role in the translation control and the cellular localization of mRNA[26)-31)]. A database, called the UTRdb has been developed[56), 57)]. It contains a collection of 5'UTR sequences for the genes whose mRNA start sites are described in the GenBank. 421 entries have been registered for the human genes in the UTRdb. It means the description about the 5'-boundaries of the 5'UTR is missing from the rest of human function-known genes (7496 species). Our cDNA sequences seemed useful to enrich the information about the 5'UTR.

V) Structural Analysis of the 5' UTR.

According to the typical ribosome-scanning model, the 5'UTR is the path for a small (40S) ribosomal subunit. In the eucaryote, the 40S ribosomal subunit is first recruited to the cap structure. It linearly scans the 5'UTR for the initiator ATG. It pauses around the initiator ATG until a large (60S) subunit joins. When the 60S subunit is combined to the 40S subunit, the ribosome becomes ready to initiate the translation[27)].

It is not fully understood how the 40S subunit pauses around the initiator ATG in the eucaryote. In the procaryote, there is a consensus sequence, called the Shine-Dalgarno sequence (SD sequence) just upstream to the initiator ATG. It is complementary to the 3'-end of the 16S ribosomal RNA (rRNA). The base-pair interaction between the 16S rRNA and the SD sequence energetically stabilizes the binding of the small ribosomal subunit to the mRNA[58]. In the eucaryote, the machinery that can serve to form a stable complex near the initiator ATG has not been reported.

Using the SECDYN2, I calculated the free energy of the optimum secondary structure for the 3'-end of the 18S rRNA. The obtained optimum structure resembled the reported structure of the 16S rRNA 3'-end[59] (Fig.IIIB-7A). In 16S rRNA, the sequence marked red is predicted to interact with the SD sequence. It sticks out from the stable stem-loop structure. The 9 bases of the 3'-end of the 18S rRNA also stuck out from the stem-loop structure and seemed free to interact with mRNA.

I calculated the the binding energy between the 9 bases of 18S rRNA 3'-end and the 5'UTR of mRNA. The upper panel of Fig.IIIB-7B shows the free energy distribution of a local secondary structure along the 5'UTR of receptor tyrosine kinase (RTK) mRNA. The lower panel shows the binding energy distribution between the 18S rRNA and the 5'UTR of RTK. I performed the same analysis for all the 5'UTR sequences that I extracted from our full/near-full clones (Fig.IIIB-7C). For many genes, the 18S rRNA and the 5'UTR sequence could form an energetically stable complex around the initiator ATG. This feature can serve to understand the ribosomal pausing in respect of the energetic stability.

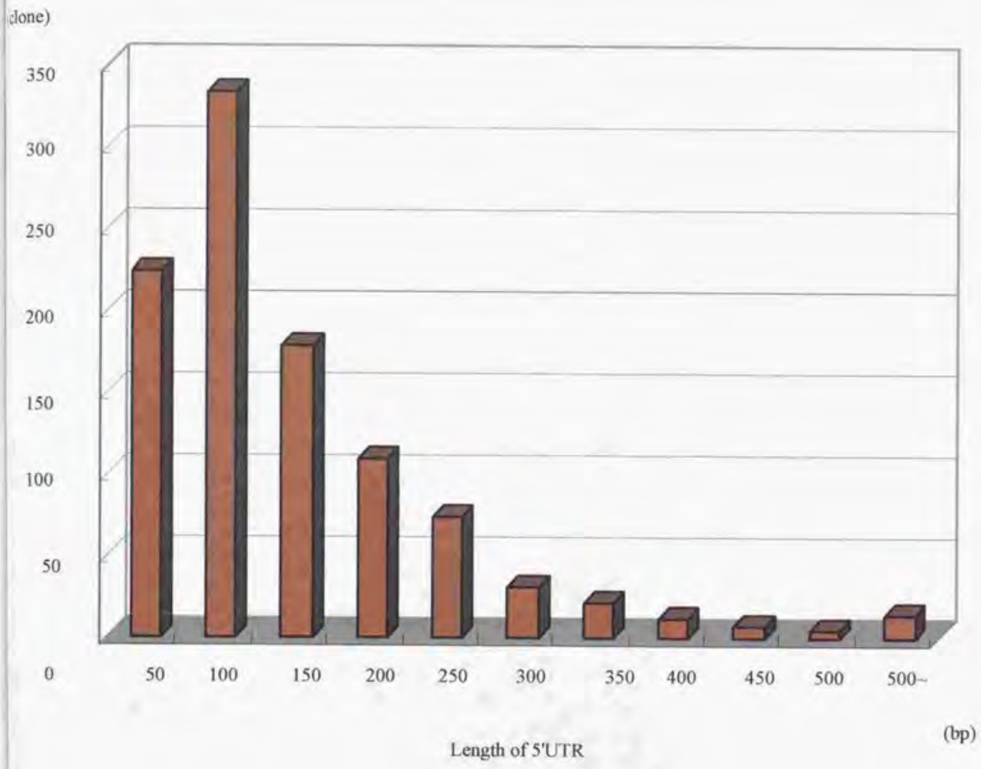


Fig.IIIB-6A **Distribution of the 5'UTR length.**

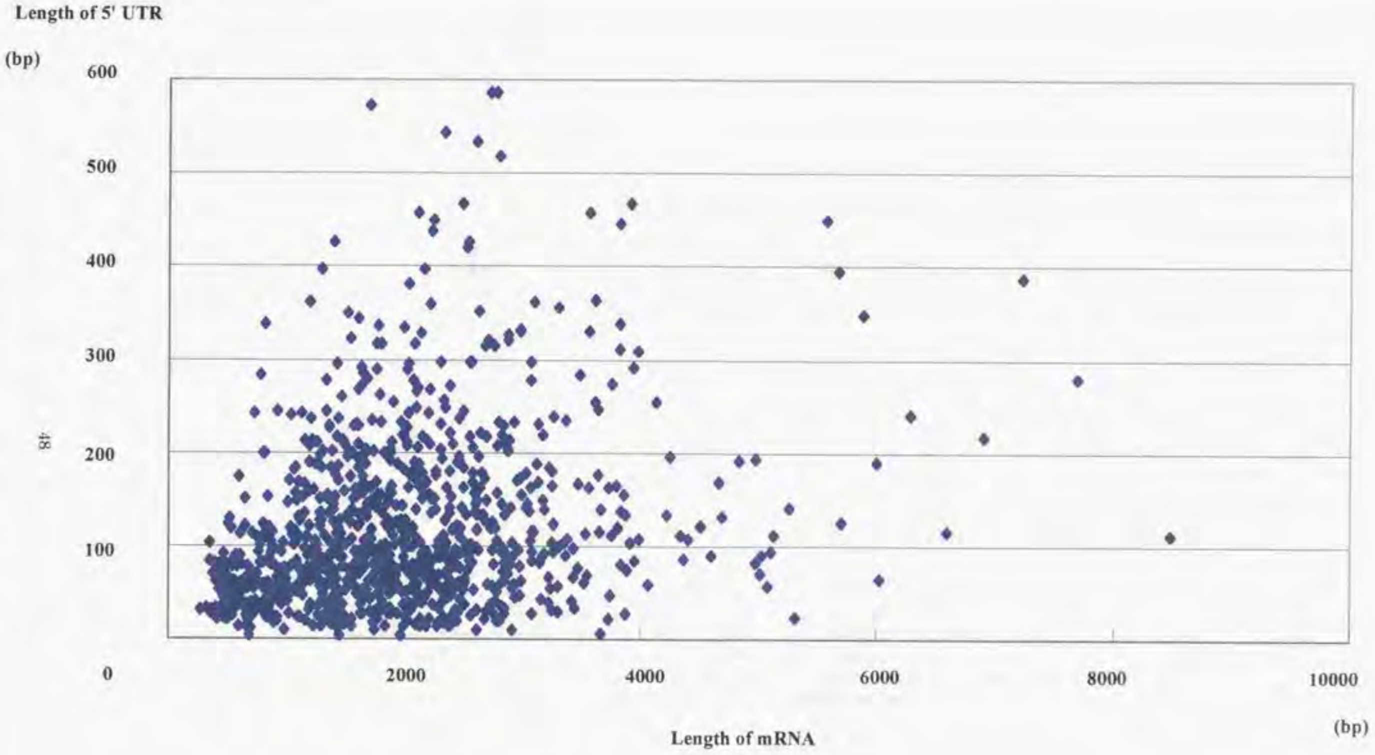


Fig.IIIB-6B Relation between the length of 5'UTR and the length of mRNA.

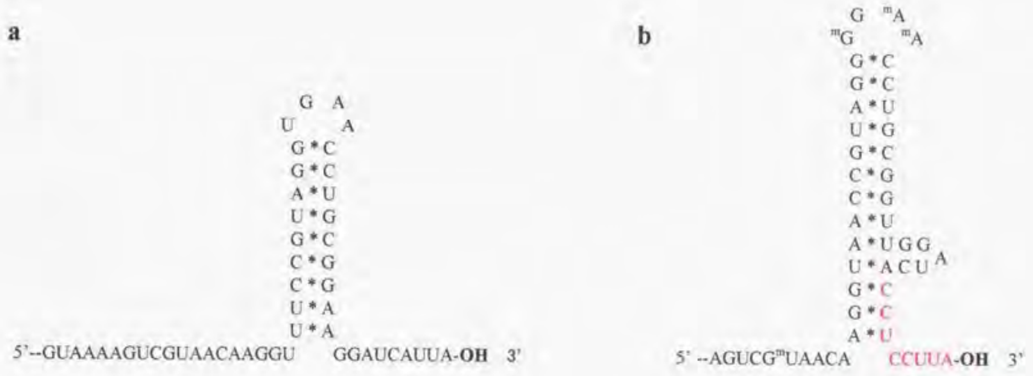


Fig.IIIB-7A Secondary structure of the 18S and the 16S rRNA 3'-end.
a: The optimum secondary structure was calculated for the 50 bases of the 18S rRNA 3'-end using SECDYN2. b: The optimum secondary structure of the E.coli 16S rRNA 3'-end was obtained by the calculation using the thermodynamic parameter table by Salser. The bases maked with red letters are predicted to interact with the SD sequence.

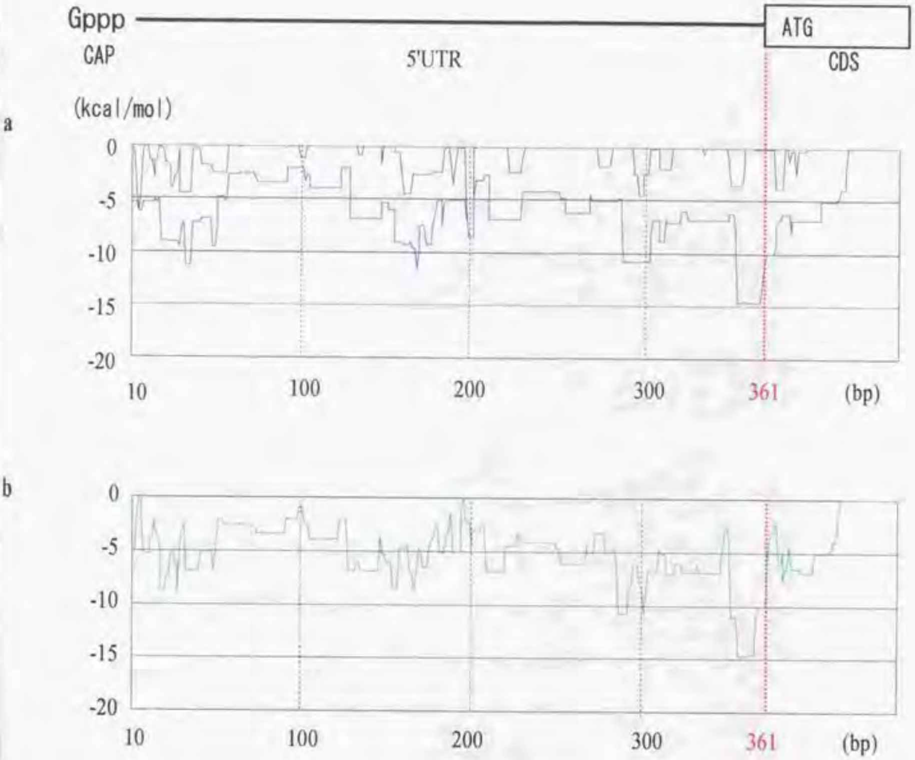


Fig.IIIB-7B **Free energy distribution of the secondary structure calculated along the 5'UTR of RTK mRNA.**

a: The free energy was calculated using the 20 bases of the 5'UTR sequence of RTK mRNA with (blue line) or without (black line) the 3'-end of the 18S rRNA. The calculated sequence was slid along the 5'UTR to the 50 bases downstream to the CDS start site. The X-axes shows the center position of the sequence used for the calculation. b: The binding energy between the 5'UTR and the 18S rRNA was plotted along the 5'UTR.

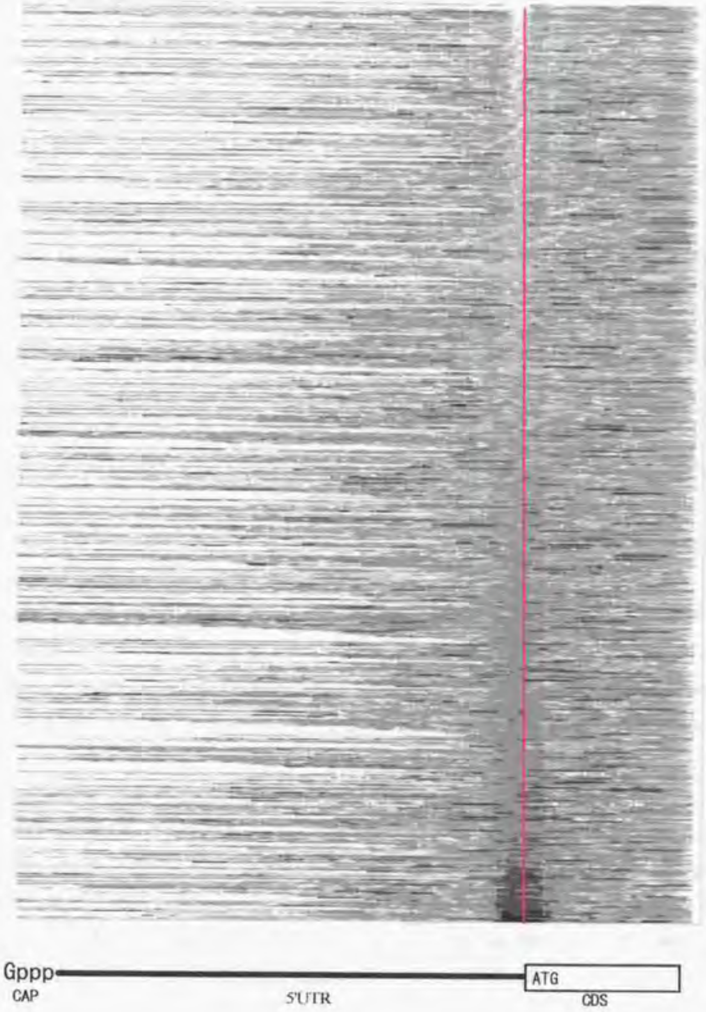


Fig.IIIB-7C Distribution of binding energy between the 5'UTR and the 18S rRNA. The distribution of the binding energy between the 5'UTR and the 18S rRNA were calculated for the full/near-full clones in the DYNACLUST. The color intensity reflects the binding energy at each region of the 5'UTR. The red line shows the CDS start site. The clones were sorted in order of their average binding energy that were calculated with the sequences within the 5 bases from the CDS start sites.

IV. CONCLUSION

In this thesis, I described the construction and the large-scale analyses of new types of cDNA libraries.

In the chapter IIIA, I described a new method to construct a full length-enriched cDNA library and a 5'-end-enriched cDNA library based on the "Oligo-capping". The content of cDNA clones that have the mRNA start site in both libraries was estimated at around 80%. This is a significant improvement of the content compared to the cDNA libraries made by conventional methods. Furthermore, the 5'-end-enriched cDNA library seemed to contain the cDNA clone with the mRNA start site of the long mRNA.

In the chapter IIIB, I reported the results of the one-pass sequence analysis of the "Oligo-capped" cDNA libraries.

As a result of the sequencing effort of 10,000 clones from the "Oligo-capped" cDNA libraries, the 5'-end sequences for more than 1,000 function-known genes have been accumulated so far. With the sequence data, I performed the statistical and thermodynamic analyses about the mRNA start sites and the 5'UTR. The current database may not be suitable for this purpose, since the descriptions about the mRNA start site, which is the 5'-boudary of the 5' UTR, are missing from their entries in many cases. Continuous sequencing of our libraries would bring further information about the mRNA start site and the 5'UTR. Our approach may be useful to generate the 5'ESTs that contain the information, which is missing from the current database.

V. REFERENCES

1. Crick, F. Central dogma of molecular biology. *Nature* **227,** 561-3 (1970).
2. Hess, E. L. Origins of molecular biology. *Science* **168,** 664-9 (1970).
3. Weaver, W. Molecular biology: origin of the term. *Science* **170,** 581-2 (1970).
4. Simmon, V. F. and Lederberg, S. Degradation of bacteriophage lambda deoxyribonucleic acid after restriction by Escherichia coli K-12. *J Bacteriol* **112,** 161-9 (1972).
5. Jackson, D., Symons, R. and Berg, P. Biochemical Method for Inserting New Genetic Information into DNA of Simian Virus 40: Circular SV40 DNA Molecules Containing Lambda Phage Genes and the Galactose Operon of E. coli. *Proc. Natl. Acad. Sci. USA* **69**, 2904-2909 (1972).
6. Maxam, A. M. and Gilbert, W. A new method for sequencing DNA. Proc. Natl. Acad. Sci.. USA **74,** 560-4 (1977).
7. Sanger, F., Nicklen, S. and Coulson, A. R. DNA Sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. USA* **74**, 5463-5467 (1977).
8. Sanger, F., Air, G. M., Barrell, B.G., Brown, N.L., Coulson, A.R., Fiddes, C.A., Hutchison, C.A., Slocombe, P.M. and Smith, M. Nucliotide sequence of bacteriophage phi X174 DNA. *Nature* **265,** 687-95 (1977).
9. Anderson, S., Bankier, A.T., Barrell, B.G., de Bruijn, M. H. L., Coulson, A. R., Drouin, J., Eperon, I. C., Nierlich, D. P., Roe, B. A., Sanger, F., Schreier, P. H., Smith, A. J. H., Staden, R. and Young,I.G. Sequence and organization of the human mitochondrial genome. *Nature* **290**, 457-465 (1981).
10. Baer, R., Bankier, A. T., Biggin, M. D., Deininger, P. L., Farrell, P.

J., Gibson, T. J., Hatfull, G., Hudson, G. S., Satchwell, S. C., Seguin, C., et al. DNA sequence and expression of the B95-8 Epstein-Barr virus genome. *Nature* **310,** 207-11 (1984).

11. Chee, M. S., Bankier, A. T., Beck, S., Bohni, R., Brown, C. M., Cerny ,R., Horsnell, T., HutchisonIII, C. A., Kouzarides, T., Martignetti, J. A., Preddie, E., Satchwell, S. C., Tomlinson, P., Weston, K. M. and Barrell, B. G. Analysis of the protein-coding content of the sequence of human cytomegalovirus strain AD169. *Curr. Top. Microbiol. Immunol.* **154**, 125-169 (1990).
12. Noll, H. Sequencing the human genome. *Science* **233,** 143 (1986).
13. Dulbecco, R. A Turning Point in Cancer Research: Sequencing the Human Genome. *Science* **231**, 1055-1056 (1986).
14. Schwartz, D. C. and Cantor, C. R. Separaiton of Yeast Chromosome-sized DNAs by Pulse Field Gel Electrophoresis. *Cell* **37**, 67-75 (1984).
15. Pavan, W. J., Hieter, P., Sears, D., Burkhoff, A., Reeves, R. H. High-efficiency yeast artificial chromosome fragmentation vectors. *Gene* **106,** 125-7 (1991).
16. Burke, D. T. The role of yeast artificial chromosomes in generating genome maps. *Curr. Opin. Genet. Dev.* **1**, 69-74 (1991).
17. Anand, R. Yeast artificial chromosomes (YACs) and the analysis of complex genomes. *Trends Biotechnol* **10,** 35-40 (1992).
18. Coulson, A., Kozono, Y., Lutterbach, B., Shownkeen, R., Sulston, J. and Waterston, R. YACs and the C. elegans genome. *Bioessays* **13,** 413-7 (1991).
19. Mullis, K., Faloona, F., Scharf, S., Saiki, R., Horn, G. and Erlich, H. Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. *Cold Spring Harb Symp Quant Biol* **51,** 263-73 (1986).
20. Rose, E. A. Applications of the polymerase chain reaction to genome analysis. *FASEB J.* **5,** 46-54 (1991).

21. Hunkapiller, T., Kaiser, B. F., Koop, B. F. and Hood, L. Large-Scale and Automated Sequence Determination. *Science* **254**, 59-67 (1991).

22. Olson, M. V. The human genome project. *Proc. Natl. Acad. Sci. USA* **90,** 4338-44 (1993).

23. Ikawa, Y. Human genome efforts in Japan. *FASEB J.* **5,** 66-9 (1991).

24. Matsubara, K. Progress on the plan for human genome project in Japan. *Tanpakushitsu Kakusan Koso* **36,** 1542-50 (1991).

25. Okubo, K., Hori, N., Matoba, R., Niiyama, T., Fukushima, A., Kojima, Y. and Matsubara, K. Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression. *Nat. Genet.* **2,** 173-9 (1992).

26. Ross, J. Control of messenger RNA stability in higher eukaryotes. *Trends Genet* **12,** 171-5 (1996).

27. Sonenberg, N. mRNA translation: influence of the 5' and 3' untranslated regions. *Curr. Opin. Genet. Dev.* **4,** 310-5 (1994).

28. Wilhelm, J. E. and Vale, R. D. RNA on the move: the mRNA localization pathway. *J. Cell. Biol.* **123,** 269-74 (1993).

29. Curtis, D., Lehmann, R., Zamore, P. D. Translational regulation in development. *Cell* **81,** 171-8 (1995).

30. Decker, C. J. and Parker, R. Mechanisms of mRNA degradation in eukaryotes. *Trends. Biochem. Sci.* **19,** 336-40 (1994)

31. Singer, R. H. The cytoskeleton and mRNA localization. *Curr. Opin. Cell Biol.* **4,** 15-9 (1992).

32. Adams, M. D., Kelley, J. M., Gocayne, J.D., Dubnick, M., Polymeropoulos, M. H., Xiao, H., Merril, C. R., Wu, A., Olde, B., Moreno, R. F., et al. Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* **252,** 1651-6 (1991).

33. Boguski, M. S. and Lowe, T. M., Tolstoshev, C. M. dbEST--

database for "expressed sequence tags". *Nat. Genet.* **4,** 332-3 (1993).

34. Boguski, M. S. The Turning Point in Genome Research. *Trends in Biochemical Sciences* **20,** 295-6 (1995).

35. Maruyama, K. and Sugano, S. Oligo-capping: a simple method to replace the cap structure of eucaryotic mRNAs with oligoribonucleotides *Gene* **138**, 171-174 (1994).

36. Biedler, J. L., Helson, L. and Spengler, B. A. Morphology and growth, tumorigenicity, and cytogenetics of human neuroblastoma cells in continuous culture. : phenotypic reversion to normal growth behavior of Chinese hamster cells. *Cancer Res.* **33,** 2643-2652 (1973).

37. Sambrook, J., Fritsch, E. F. and Maniatis, T. Molecular Cloning : A Laboratory Manual, 2nd ed. Cold Spring Harbor Laboratory. (1989)

38. Suzuki, Y., Yoshitomo, K., Maruyama, K., Suyama A. and Sugano, S. Construction and characterization of a full length-enriched and a 5'-end-enriched cDNA library. *Gene* **200,** 149-156 (1997).

39. Seed, B. and Aruffo, A. Molecular cloning of the CD2 antigen, the T-cell erythrocyte receptor, by a rapid immunoselection procedure. *Proc. Natl. Acad. Sci. USA* **84,** 3365-3369 (1987).

40. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215,** 403-410 (1990).

41. Pearson, W. R. and Lipman, D. J. Improved tools for biologic sequence comparison. *Proc. Natl. Acad. Sci. USA* **85**, 2444-2448 (1988).

42. Wada, A. and Suyama, A. Local stability of DNA and RNA secondary structure and its relation to biological functions. *Prog Biophys Mol Biol* **47,** 113-57 (1986).

43. Furuichi, Y. and Miura, K. A blocked structure at the 5' terminus of mRNA from cytoplasmic polyhedrosis virus. *Nature* **253**, 374-375 (1975).

44. Shinshi, H., Miwa, M., Kato, K., Noguchi, M. Matushima, T. and Sugimura, T. A novel phosphodiesterase from cultured tobacco cells. *Biochemistry* **15,** 2185-2190 (1976).

45. Uetsuki, T., Naito, A., Nagata, S. and Kaziro, Y. Isolation and characterization of the human chromosomal gene for polypeptide chain elongation factor-1 alpha. *J. Biol. Chem.* **264,** 5791-5798 (1989).

46. Kato, S., Sekine, S., Oh, S. W., Kim, N. S. Umezawa, Y., Abe, N., Yokoyama-Kobayashi, M. and Aoki, T. Construction of a human full-length cDNA bank. *Gene* **150,** 243-250 (1994).

47. Carninci, P., Kvam, C., Kitamura, A., Ohsumi, T., Okazaki, Y., Itoh, M., Kamiya, K., Sasaki, N., Izawa, M., Muramatsu, M., Hayashizaki, Y. and Scheider, C. High-efficiency full-length cDNA cloning by biotinylated CAP trapper. *Genomics* **37,** 327-336 (1996).

48. Seelig, H. P., Schranz, P., Schroter, H., Wiemann, C., Griffiths, G. and Rentz, M. Molecular genetic analysis of a 376-kilodalton Golgi complex membrane protein (giantin). *Mol. Cell. Biol.* **14,** 2564-2576 (1994).

49. Mikami, A., Paschal, B. M., Mazumdar, M. and Vallee, R. Molecular cloning of the retrograde transport motor cytoplasmic dynein (MAP 1C). *Neuron* **10,** 787-796 (1993).

50. Zhang, Z. Tanaka, Y., Nonaka S., Aizawa, H., Kawasaki, H., Nakata, T. and Hirokawa, N. The primary structure of rat brain (cytoplasmic) dynein heavy chain, a cytoplasmic motor enzyme. *Proc. Natl. Acad. Sci.. USA* **90,** 7928-7932 (1993).

51. Gubler, U. and Hoffman, B. J. A simple and very efficient method for generating cDNA libraries. *Gene* **25,** 263-269 (1983).

52. Okayama, H. and Berg, P. High-efficiency cloning of full-length cDNA. *Mol. Cell. Biol.* **2,** 161-170 (1982).

53. Edery, I., Chu, L. L., Sonenberg, N. and Pelletier, J. An efficient strategy to isolate full-length cDNAs based on an mRNA cap

retention procedure (CAPture). *Mol. Cell. Biol.* **15,** 3363-3371 (1995).

54. Barnes, W. M. PCR amplification of up to 35-kb DNA with high fidelity and high yield from lambda bacteriophage templates. *Proc. Natl. Acad. Sci. USA* **91,** 2216-2220 (1994).

55. Cheng, S., Fockler, C. Barmes, W. M. and Higuchi, R. Effective amplification of long targets from cloned inserts and human genomic DNA. *Proc. Natl. Acad. Sci. USA* **91,** 5695-5699 (1994).

56. Pesole, G., Liuni, S., Grillo, G. and Saccone, C. Structural and compositional features of untranslated regions of eukaryotic mRNAs. *Gene* **205,** 95-102 (1997).

57. Pesole, G., Liuni, S., Grillo, G. and Saccone, C. UTRdb: a specialized database of 5'- and 3'-untranslated regions of eukaryotic mRNAs. *Nucleic Acids Res.* **26,** 192-5 (1998).

58. Jacques, N. and Dreyfus, M. Translation initiation in Escherichia coli: old and new questions. *Mol. Microbiol.* **4,** 1063-7 (1990).

59. Studnicka, G. M., Rahn, G. M., Cummings, I. W. and Salser, W. A. Computer method for predicting the secondary structure of single-stranded RNA. *Nucleic Acids Res.* **5,** 3365-87 (1978).

VI. ACKNOWLEDGEMENT

I would like to express my heartfelt thank to Dr. Suyama and Dr. Sugano for their generous teaching and the helpful advice. Without their moral support, I could not have overcome many hardships to accomplish this work.

I'm especially grateful to Ms. Hata, Ms. Shirai, Ms. Takahashi, Mr. Sasaki, Mr. Watanabe, Dr. Komatsu and Mr. Tsunoda. They provided all the sequence data used to this work.

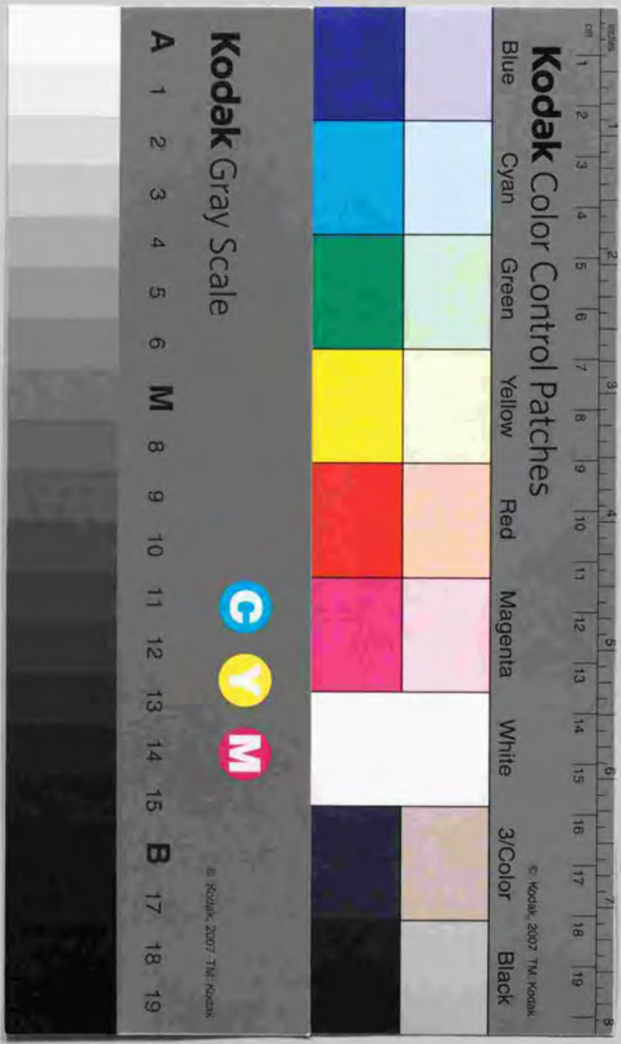
I express thanks to all my colleagues in the laboratory in the IMSUT and the HRI, Dr. Isogai, Dr. Ohta, Dr. Kataoka, Ms. Shioda, Ms. Nakagawa, Mr. Yoshida, Mr. Imai, Ms. Shibui, Mr. Ohmori, Mr. Hida and Ms Tsuchida. They always made an exciting and stimulating environment in the laboratory. I have also enjoyed the many discussions with Mr. Takano and Mr. Uchikoga and all the members of the Suyama laboratory.

I also thank to Ms. Okada for her assistant in experiments and Ms. Yamashita for her help in office works.

I am particularly thankful to Mr. Nakagawa, Ms. Koyanagi, Ms. Mizuno and Ms. Kuwabara. They encouraged and supported me kindly. I appreciated their kindness and friendship.

I was thankfully supported by the scholarship from the Japan Society for the Promotion of Science.

Finally, I present this thesis to my parents and grandparents with a lot of thanks for their great love and care.



Kodak Color Control Patches
Blue
Cyan
Green
Yellow
Red
Magenta
White
3/Color
Black
© Kodak, 2007 TM: Kodak
Kodak Gray Scale
A 1 2 3 4 5 6 M 8 9 10 11 12 13 14 15 B 17 18 19
C Y M
© Kodak, 2007 TM: Kodak