0273-2289/1559-0291 (Online)/07/136:327–344/$30.00

Expression of Open Reading Frames in Silkworm Pupal cDNA Library

YAO-ZHOU ZHANG,* JIAN CHEN, ZUO-MING NIE, ZHENG-BING LÜ, DAN WANG, CAI-YING JIANG, PING-AN HE, LI-LI LIU, YU-LAN LOU, LI SONG, AND XIANG-FU WU

College of Life Sciences, Zhejiang Sci-Tech University, Hangzhou 310018, China, E-mail: yaozhou@zist.edu.cn

**Received February 10, 2006; Revised ;
Accepted April 2, 2006**

Abstract

A cDNA library containing 2409 singletons was constructed from whole silkworm pupae (*Bombyx mori*) In addition, the types of genes overexpressed in pupa were analyzed. These genes contained 79 types of proteins with the exception of enzyme, mitochondrial DNA, andribosomal protein. Also analyzed were the expression and nonexpression of open reading frame (ORF) sequences in *Escherichia coli*. cDNA sequences were compared to the silkworm (*B. mori*) genome in the GenBank database and the silkworm cDNA database including the SilkBase and KAIKOBLAST databases and 498 novel expressed sequence tags (ESTs) and 217 unknown ESTs were found. After comparison with all available ORF-complete mRNA sequences from the same organism (fruitfly, mosquito, and apis) in the RefSeq collection, 1659 full-length cDNA were identified. In addition, the structure of silkworm mRNA was analyzed, and it was found that 66.8% of silkworm mRNA tailed with poly(A) contained the highly conserved AAUAAA signal and the signal located 10–17 nucleotides upstream of the putative poly(A). Finally, the composition of nucleotides in promoter region for all ESTs was surveyed. The results imply that the TTTTA box may possess some functions in regulating transcription and expression of some genes.

Index Entries: *Bombyx mori*; construction; full-length cDNA; open reading frame; silkworm; expressed sequence tag.

*Author to whom all correspondence and reprint requests should be addressed.

Introduction

Advances in genomics, proteomics, and transcriptomics has been surprisingly rapid with the completion of the Human Genome Project (HGP) and the Rice Genome Project *(1–4)*. The completed HGP showed that there are only 24,000 coding genes in human genome, less than one-fourth of the 100,000 genes that had been previously forecast *(1)*. This indicates that coding genes are very limited in the genome. The framemap drafts of both *Drosophila* and *Bombyx mori* genomes have been reported recently *(5–7)*. Because ecdysis occurs six times and metamorphosis occurs two times in the 42-d life cycle of *B. mori*, the expression and regulation of genes are very quick and intricate. Silkworm is one of the representative studies of the modes of biochemistry in insects just like *Drosophila*, and one of the ideal materials for insect genetics *(8)*. Recently, great advancements have been achieved in the research of silkworm (*B. bombyx*) cDNA. Mita et al. *(9)* built a *Bombyx* expressed sequence tag (EST) database, SilkBase, on the basis of the construction of 36 cDNA libraries from a variety of tissues of silkworm.

In the present study, a full-length cDNA library of silkworm pupae without brains was constructed. The cDNA sequences were analyzed and characterized, and 629 new genes were identified from the library, which may be owing to the vigorous development in the pupating stage. In this stage, larva develops into pupa, which requires the participation of many new proteins.

Materials and Methods

Reagents

Trizol reagent was purchased from Invitrogen, to extract total RNA from the silkworm pupae. A PolyATtract kit (Promega) was used to isolate silkworm mRNA from total RNA. DNase I (RNase I–free), an M-MLV cDNA Synthesis kit, and all the restriction enzymes were purchased from Takara, for cDNA synthesis and identification of recombinant clones. The plasmid vector, pHelix, was from Boehringer Mannheim.

Silkworm Pupae and Bacterial Strain

Total RNA was purified from a domesticated variety, Qingsong × Haoyue (d 3 after pupating). The bacterial strain TG1, for construction of the cDNA library, is stored in our laboratory.

Construction of cDNA Library of Silkworm Pupae

Total RNA was extracted from silkworm pupae, then digested by DNase I (RNase I–free) to eliminate contaminated genomic DNA. Subsequently, mRNA was isolated from the total RNA with a PolyATtract kit, and cDNA was produced using an M-MLV cDNA Synthesis kit. The obtained double-strand cDNA was blunted by T4 DNA polymerase, then inserted into the vector pHelix treated by *Hin*dII. The ligation products were transformed into TG1-competent cells using the standard method. Finally, the transformed cells were plated on Luria-Bertani (LB) agar containing Amp, X-gal, and isopropyl-β-D-thiogalactoside and incubated overnight at 37°C for screening of blue or white plaques.

Sequencing of cDNA Library

Plasmid clones were randomly selected, and each colony was inoculated into 3 mL of LB medium and incubated at 37°C for 8–10 h with vigorous shaking. Recombinant clones were obtained by miniprep (Rapid Plasmid Miniprep Kit; V-gene, Hangzhou, China), then digested with *Eco*RI/*Hin*dIII. Finally, cDNA insert length was measured by agarose gel electrophoresis against standards and visualized using VDS-CL (Amersham). The identified recombinant clones were stored in LB medium containing 50% glycerol at –86°C. Based on cDNA insert sizes, DNA sequencing for positive clone was performed for single pass or from both ends using a BigDye Teminator v3.1 kit (standard protocol) with ABI Prism 3100-A. The sequencing primers were M13F (5'-TGTAAAACGACGGCCAGT-3') and M13R (5'-CAGGAAACAGCTATGACC-3') for forward and backward direction, respectively. The sequencing analyzers were 3100-Avant Genetic Analyzer and 3130*xl* Genetic Analyzer, purchased from ABI.

Processing and Analysis of cDNA Sequences

Raw sequences were first trimmed to remove vector sequence and poor-quality sequences using ESTprep, a program that reads preprocessing cDNA sequence *(8)*. Then contigs were assembled from redundant reads using SeqManII soft (DNASTAR package). Sequencing of the silkworm genome recently has been completed *(6,7)*, so the genomic sequence of silkworm is available to identify the cloned cDNA sequences at the NCBI genomic BLAST Web site (www.ncbi.nlm.nih.gov/sutils/genom_table.cgi?organism=insects), together with the constructed silkworm cDNA database, which includes SilkBase (http://papilio.ab.a.u-tokyo.ac.jp/silkbase/) and KAIKOBLAST databases (http://kaikoblast.dna.affrc.go.jp/). The processed cDNA sequence was used to query the silkworm genome database with a Blastn algorithm (default parameters). Sequences with an E-value of e-100 or less were classified as significantly matching when compared with the silkworm genomic sequence. Those sequences with an E-value between e-50 and e-99 were given special consideration because this could represent an mRNA matching to different exons of the same gene *(10)*. The remaining sequences were judged as not significantly matching. Similarly, all the sequences were used to compare both SilkBase and KAIKOBLAST to the silkworm cDNA database, but with a different E-value criterion of e-20 to classify *(11)*.

Results

Construction of cDNA Library from Silkworm Pupae

Extraction of Total RNA and Isolation of mRNA

Six fresh pupae stored at –86°C (~6 g) were removed from storage and disinfected with alcohol, and the shells were then shucked off. Subsequently, the pupae were crumbled still in a frozen state, and tissue pieces were transferred to 15-mL centrifuge tubes and treated with Trizol reagent in a proportion of 10:1 (mL/g), i.e., the addition of10 mL of Trizol to 1 g of pupal tissue, for homogenization. Total RNA was extracted according to the manufacturer's (Trizol) instruction after homogenization and pooled together to treat with DNase I (RNase-free). From the treated total RNA, mRNA was purified with a PolyATtract kit. The quality of the total RNA was evaluated by denaturing electrophoresis in a 2% agarose gel. Both 18 S and 28 S RNAs were clearly visible, indicating that the total RNA was fully integrated. The purified mRNA was measured with a spectrophotometer, and the value of A_{260}/A_{280} was 1.95.

cDNA Synthesis and Library Construction

Double-stranded cDNA was synthesized using 5.0 μg of mRNA as template according to the kit's protocol. The ratios of recombination were analyzed by blue/white plaque screening. The results showed that the ratio of recombination was about 91% and that the primary sink size was about 1.8×10^6, thus fulfilling the requirements for library construction.

Identification of Recombinant Clones

White plaque clones were chosen to inoculate and culture. The plasmids were extracted using a rapid plasmid DNA miniprep kit. Then positive clones and the size of extraneous insert segments were identified by digestion with *Eco*RI and *Hin*dIII and 1% agarose gel electrophoresis. Statistical results showed that the average size of the cloning segments was from 100 to 4000 bp (Fig. 1). The length of most segments was between 400 and 1500 bp (Fig. 2).

EST Sequencing and Analysis of cDNA Library of Silkworm Pupae

End Sequencing

Three thousand recombinant clones from the constructed cDNA library were randomly chosen for EST sequencing on the basis of the size of cDNA inserts. Inserts with a length of <1300 bp were adopted for full-insert sequencing. The remaining sequences were adopted for single-direction sequencing. If the obtained cDNA sequence contained a poly(A) tail too long to read, resequencing was carried out from the other end. Of the total cDNA sequences, 2409 effective sequences were obtained after trimming vector sequences, bacteria sequences, and poor-quality sequences.

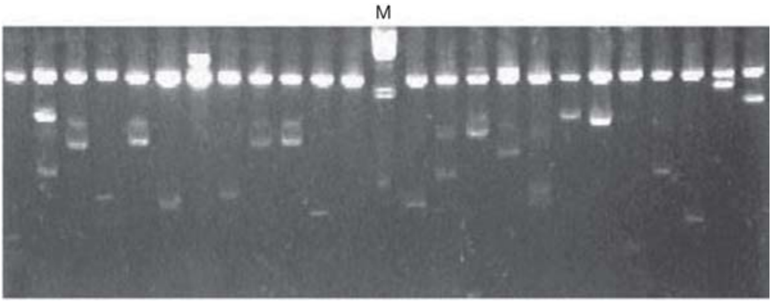


Fig. 1. Identification of recombinant clones by digesting with *Eco*RI plus *Hin*dIII. M is the products of λ DNA digested by *Hin*dIII.

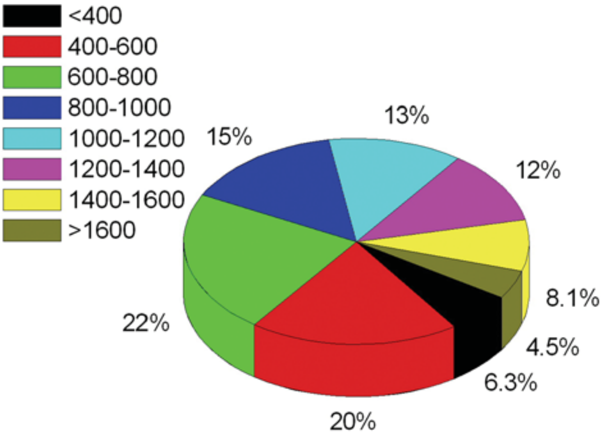


Fig. 2. Size distribution of insert fragments cloned.

cDNA Analysis

Using the Blast Program, we searched the 2409 effective sequences in GenBank, PDB, and Swiss-Prot and found that 629 sequences are highly homologous to some silkworm genes. In addition, 174 sequences are partial sequences of some ribosomal protein, and 110 partial sequences of some mitochondrial DNA. There were also 88 sequences coding various enzyme proteins, such as cytochrome oxidase, NADH dehydrogenase, and adenosine triphosphatase. The other 257 sequences code protein except above, such as cyclophilin, bombyxin, elongation factor, heat-shock protein, BCP inhibitor, promoting protein, troponin, annexin, and antitrypsin precursor (Table 1).

Analysis of Expression and Nonexpression of ORF Sequences

In all full-length cDNAs, we analyzed possible noticeable cDNA sequences for silkworm growth and pharmaceutical value. These ORF sequences (Table 2) were inserted into the vector pET-28a and the genes

Table 1
Classification of Sequence Homologous to Silkworm Gene Excluding Mitochondrial DNA, Ribosomal Protein, and Enzyme

cDNA	No. of cloned sequences
Elongation factor 1α	16
Vitellogenin	16
Heat-shock protein	12
Bombyxin E-1 precursor	10
Fibroin heavy chain Fib-H	10
Actin 2	9
Annexin B13a	8
Antitrypsin precursor	8
Bmkettin	8
Ubiquitin	8
Bmtitin1	7
FK506-binding protein	7
Tubulin	7
Cyclophilin	6
Retrotransposable elements	5
30K protein	5
BCP inhibitor	5
Profilin	5
Apoptosis protein inhibitor	4
Bombyxin	4
Cuticle protein	4
Hemocytin	4
Hemolin	4
Pleiotrophin-like protein	4
Translationally controlled tumor protein	4
Acyl-CoA-binding protein	3
Endonuclease and reverse transcriptase–like protein	3
Gloverin-like protein	3
Hemolymph 30K protein precursor	3
Ribosome-associated protein P40	3
Wing disclike protein	3
Y-box protein	3
Antibacterial peptide	2
BmP109	2
Bombyrin	2
Disulfide-isomerase like protein	2
Dscam	2
Hemolin-interacting protein	2
Low molecular lipoprotein 30K	2
Lysozyme	2
Silk protein P25	2
Alp-m	1
Apolipophorin III	1
BCP	1

(continued)

Table 1 (Continued)

cDNA	No. of cloned sequences
BmHSC70-4	1
CathD	1
Chilgok 3	1
Chitinase precursor	1
Chitinase-like protein	1
Cholesterol transporter	1
Chorion protein	1
Chymotrypsin inhibitor	1
ComEC/Rec2-related protein	1
Diapause hormone precursor	1
Fibroin light chain	1
Gypsy-Ty3-like retrotransposon Kabuki gene and nested	1
Hsc70/Hsp90-organizing protein HOP	1
Immulectin	1
Internal transcribed spacer	1
Kiser	1
Lark-PA	1
Lebocin 3	1
Lipoprotein	1
Lysozyme precursor	1
Mago-nashi-like protein	1
Masquerade-like serine proteinase homolog	1
Microsatellite	1
Mod(mdg4)-heS00531	1
Non-LTR retrotransposon	1
Promoting protein	1
Odorant receptor	1
Storage-protein SP1	1
Subtelomere region	1
TPA: Mod(mdg4)-heS00531	1
Transposase yabusame	1
tRNA-Gly	1
Troponin I	1
Unknown salivary protein	1
Yabusame-2	1

were expressed in *Escherichia coli*. However, some ORF sequences could not be expressed in *E. coli* (Fig. 3).

First, we divided the ORF sequences into expression and nonexpression groups. We identified that the nonexpression sequences had almost no poly(A) in the 3' end. The G+C content of nonexpression ORFs upstream of ORF was greater than that of expression ORFs, 43 and 41%, respectively. On the other hand, the G+C content of nonexpression ORFs downstream of ORF was lower than that of expression ORFs, 27.4 and 28.7%, respectively (Tables 3 and 4).

Table 2
Expression and Nonexpression of ORF Sequences[a]

Expressed ORFs	Nonexpressed ORFs
Troponin C (1)	Lectin (12)
Tropomyosin (2)	Ferritin subunit precursor (13)
RNA methyltransferase–like (3)	Troponin T (14)
Abnormal wing disclike protein (4)	Zinc finger protein (15)
BmCRABPs (5)	CG13510-PA (16)
Protease inhibitor–like protein (6)	ENSANGP00000009009 (17)
CG7917-PA (7)	
ENSANGP00000008511 (8)	
ENSANGP00000009311 (9)	
ENSANGP00000013177 (10)	

[a]Numbers in parentheses correspond to lanes in Fig. 3.

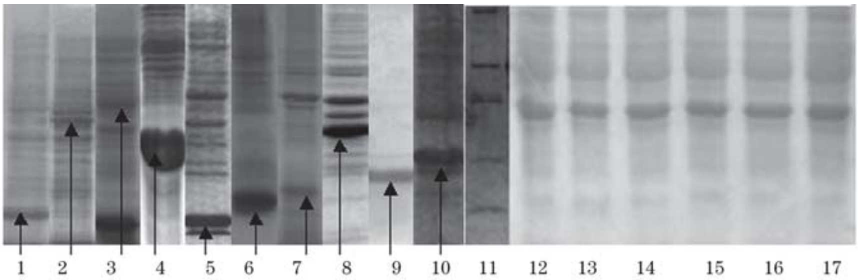


Fig. 3. SDS-PAGE of expression and nonexpression ORFs. The arrows indicate the corresponding proteins. See Table 2 for the ORFs that correspond to the lanes. Lane 11 indicated the molecular weight of protein.

Second, we determined the ORF of each full-length cDNA and analyzed their ORF sequences. The results showed that the repeat sequences occurred in the nonexpression sequences (Table 3). For example, trinucleotides GAA, GAG occurred frequently. Figure 4 presents one nonexpression sequence and its amino acid sequence. It was the only one that contained poly(A) in the 3' end. It can be seen that the occurrence number of GAA, AAA, AAG is very high.

EST Analysis

After aligning and clustering the total obtained ESTs, 2409 singletons were assembled. In the contigs, homologies with other gene entries were discovered by comparison with the silkworm genome and cDNA databases. The results suggested that all of the sequences could be divided into five groups (Table 5), which were named mitochondrial DNA, ribosomal protein, known ESTs (significantly homologous to both the silkworm genome and cDNA databases), unknown ESTs (not homologous to the silkworm genome database but obviously homologous to the silkworm

Table 3
Analysis of Expressed ORF Sequences

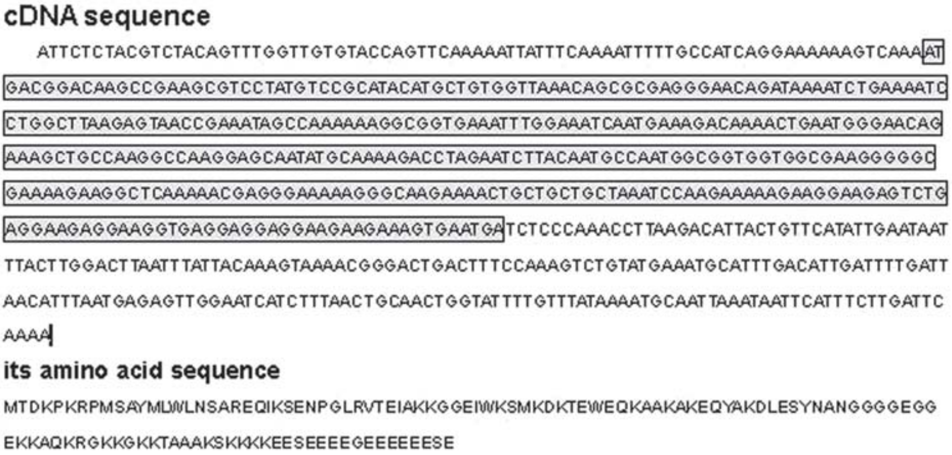
ORF	G+C content in 3' end (%)	G+C content in 5' end (%)	Poly(A) in 3' end	Repeats in ORF
Troponin C	31.7	50	No	GAC
Tropomyosin	29.8	61.5	Yes	GCC, GAC, GAG, GAA, AAG, CAG
RNA methyl-transferase-like	12.8	41.2	Yes	No
Abnormal wing disc-like protein	23.8	42.3	No	No
Protease inhibitor–like protein	40.2	45.7	Yes	No
BmCRABPs	35.2	28	Yes	No
CG7917-PA	38	39.3	Yes	GAA
ENSANGP00000008511	11.7	31	Yes	No
ENSANGP00000009311	40	37	Yes	GAA, AAA
ENSANGP00000013177	24.1	33.9	Yes	No
Average	28.7	41		

Table 4
Analysis of Nonexpressed ORF Sequences

ORF	G+C content in 3' end (%)	G+C content in 5' end (%)	Poly(A) in 3' end	Repeats in ORF
Lectin	23.7	56.7	No	AAC
Ferritin subunit precursor	26.3	38.8	No	CTC, GCC
Troponin T	38.4	55.9	No	GAG, GAA
Zinc finger protein	23.5	35.6	Yes	GAG, GAC, TGT, TGC
CG13510-PA	25	40.2	Yes	No
ENSANGP00000009009	27.2	30.8	Yes	GAA, AAA, AAG
Average	27.4	43		

cDNA database), and novel ESTs (highly homologous to the silkworm genome database and no matching to sequences in the silkworm cDNA database), respectively. The results showed that 1410 sequences are known ESTs, 217 are unknown ESTs, and 498 are novel ESTs. Of the total EST sequences, the percentage of the known ESTs, unknown ESTs, and novel ESTs was 58.5, 9, and 20.7%, respectively.

Again, we analyzed the cloned cDNA to identify full-length cDNA clones using the method of Strausberg et al. *(12)*. We found that 4212 putative ORFs were coded by 1659 full-length cDNAs, of which 1167 ESTs contained full ORF (Table 6). Briefly, we identified all possible ORFs in cDNA sequences by locating the standard and alternative stop and start codons



cDNA sequence

```
    ATTCTCTACGTCTACAGTTTGGTTGTGTACCAGTTCAAAAATTATTTCAAAATTTTTGCCATCAGGAAAAAAGTCAAAAT
GACGGACAAGCCGAAGCGTCCTATGTCCGCATACATGCTGTGGTTAAACAGCGCGAGGGAACAGATAAAATCTGAAAATC
CTGGCTTAAGAGTAACCGAAATAGCCAAAAAAGGCGGTGAAATTTGGAAATCAATGAAAGACAAAACTGAATGGGAACAG
AAAGCTGCCAAGGCCAAGGAGCAATATGCAAAAGACCTAGAATCTTACAATGCCAATGGCGGTGGTGGCGAAGGGGGC
GAAAAGAAGGCTCAAAAACGAGGGAAAAAGGGCAAGAAAACTGCTGCTGCTAAATCCAAGAAAAAGAAGGAAGAGTCTG
AGGAAGAGGAAGGTGAGGAGGAGGAAGAAGAAAGTGAATGATCTCCCAAACCTTAAGACATTACTGTTCATATTGAATAAT
TTACTTGGACTTAATTTATTACAAAGTAAAACGGGACTGACTTTCCAAAGTCTGTATGAAATGCATTTGACATTGATTTTGATT
AACATTTAATGAGAGTTGGAATCATCTTTAACTGCAACTGGTATTTTGTTTATAAAATGCAATTAAATAATTCATTTCTTGATTC
AAAA
```

its amino acid sequence

```
MTDKPKRPMSAYMLWLNSAREQIKSENPGLRVTEIAKKGGEIWKSMKDKTEWEQKAAKAKEQYAKDLESYNANGGGGEGG
EKKAQKRGKKGKKTAAAKSKKKKEESEEEEGEEEEEESE
```

Fig. 4. One nonexpression sequence and its amino acid sequence.

Table 5
Classification of All ESTs

cDNA category	No. of clones	Percentage
I. Mitochondrial DNA	110	4.6
II. Ribosomal protein	174	7.2
III. Known ESTs	1410	58.5
IV. Unknown ESTs	217	9
V. Novel ESTs	498	20.7
Total	2409[a]	100

[a] In the 2409 sequences, there are 2233 sequences that have been submitted to GenBank. DN985369-DN985373, and DY230449-DY231514.

with ORF finder (see GenBank software). Then the deduced amino acid sequences were used to Blast against the same organism (such as fruitfly or mosquito) in the RefSeq collection.

We found that about 3% mRNA sequence is noncoding mRNA with a poly(U) sequence at the 5' terminal that is complementary to the poly(A) sequence at the 3' terminal, forming a particular cucurbit-shaped secondary structure that still has a 2- to 80-nucleotide (nt) poly(A) terminal. These noncoding mRNA were identified as small nucleolar RNA (snoRNA) and has the characteristic of snoRNA. snoRNA in insects has not yet been reported; therefore, the discovery of *B. mori* snoRNA and analysis of the structural characteristics, especially the discovery of the sequences with a poly(U) head and poly(A) tail, provided the foundation for research of snoRNA in insects. Our relative article has been submitted for publication.

Table 6
Analysis of Full-Length cDNA for Putative Full ORFs

cDNA	No. of clones	Percentage
Full-length cDNA[a]	1659	100
Containing ORF	1167	70.3
No ORF	492	29.7

[a]Includes some noncoding mRNAs.

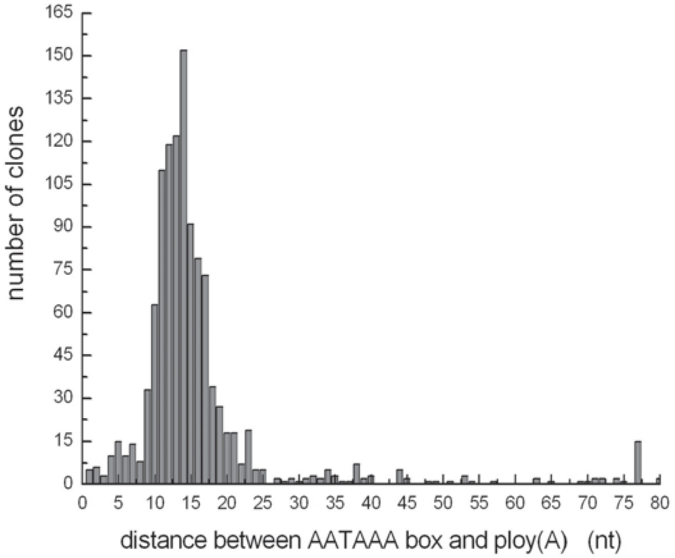


Fig. 5. Analysis of AATAAA box location away from poly(A). Position 0 is the putative poly(A) site.

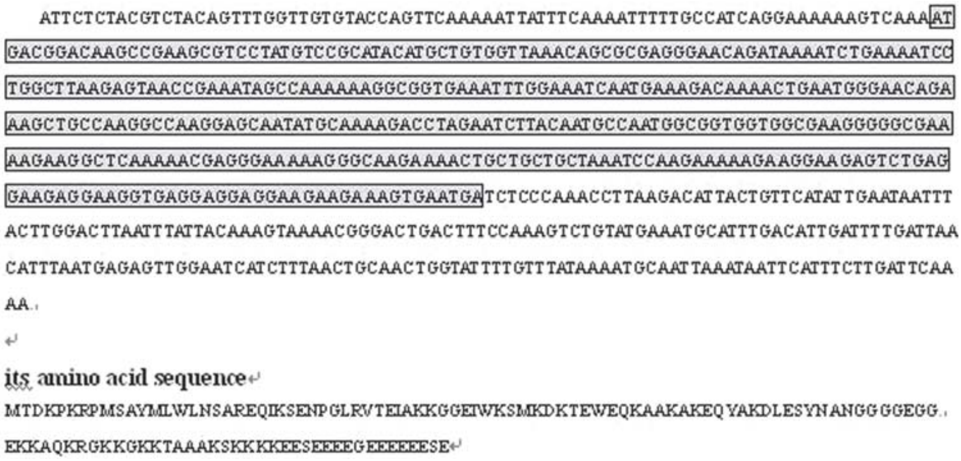
Analysis of AATAAA Box in mRNA poly(A)

In our study, we determined that there are 1609 mRNAs with poly(A). Of these, 1187 mRNAs, about 73.8%, have AATAAA box in poly(A). Searching the distance between each AATAAA box and poly(A) of the sequence, we found that the most frequent position of AATAAA box is from 10 to 17 bp, which contains 809 mRNAs. There are 152 mRNAs for which the distance between each AATAAA box and poly(A) is 14 bp, 122 mRNAs for 13 bp, 119 mRNAs for 12 bp, and 110 mRNAs for 11 bp (Fig. 5).

Analysis of Upstream Region in mRNAs

Using statistical methods, we analyzed the regulation of base composition in the upstream region of ORF to distinguish functional location in promoter of mRNAs. The working process was as follows:

First, we searched ORFs in all singletons using the DNAStar program package. The results revealed that 2233 ORFs are direct coding sequences,



```
ATTCTCTACGTCTACAGTTTGGTTGTGTACCAGTTCAAAAATTATTTCAAAATTTTTGCCATCAGGAAAAAAGTCAAAAT
GACGGACAAGCCGAAGCGTCCTATGTCCGCATACATGCTGTGGTTAAACAGCGCGAGGGAACAGATAAAATCTGAAAATCC
TGGCTTAAGAGTAACCGAAATAGCCAAAAAAGGCGGTGAAATTTGGAAATCAATGAAAGACAAAACTGAATGGGAACAGA
AAGCTGCCAAGGCCAAGGAGCAATATGCAAAAGACCTAGAATCTTACAATGCCAATGGCGGTGGTGGCGAAGGGGGCGAA
AAGAAGGCTCAAAAACGAGGGAAAAAGGGCAAGAAAACTGCTGCTGCTAAATCCAAGAAAAAGAAGGAAGAGTCTGAG
GAAGAGGAAGGTGAGGAGGAGGAAGAAGAAAGTGAATGATCTCCCAAACCTTAAGACATTACTGTTCATATTGAATAATTT
ACTTGGACTTAATTTATTACAAAGTAAAACGGGACTGACTTTCCAAAGTCTGTATGAAATGCATTTGACATTGATTTTGATTAA
CATTTAATGAGAGTTGGAATCATCTTTAACTGCAACTGGTATTTTGTTTATAAAATGCAATTAAATAATTCATTTCTTGATTCAA
AA.
```

its amino acid sequence

```
MTDKPKRPMSAYMLWLNSAREQIKSENPGLRVTEIAKKGGEIWKSMKDKTEWEQKAAKAKEQYAKDLESYNANGGGGEGG.
EKKAQKRGKKGKKTAAAKSKKKKEESEEEEGEEEEEESE
```

Fig. 6. Results of alignment by Clustal1.83.

and 1979 reverse coding sequences. Second, we took a substring of 100 bp in length in front of each ORF and divided 30 sequences into a group. Third, Using the ClustalX program package, we aligned sequences in each group and observed more frequent oligonucleotide in all sequences (Fig. 6). Fourth, we searched all possible oligonucleotides that contain only C and G and searched TATA box and CAAT box based on the structure of the promoter. Finally, we randomly selected three groups and searched all possible oligonucleotides that contain only A and T. The results indicated that oligonucleotides that contain only C and G are infrequent in sequences, <20%, and that the frequency of CAAT box is <10%. In all possible 5-mers with A and T, however, the most frequent oligonucleotide is TTTTA. In addition, TTTATT is the most frequent oligonucleotide in all possible 6-mers with A and T (Figs. 7–12).

We then searched the position of TTTTA box in each sequence. Figure 13 shows the number of the TTTATT box for each sequence located 0–100 nt upstream of ATG. The most frequent positions are 50–60 bp in front of ATG. The research of TTTTA box in promoter are not seen so far. The results showed that the TTTTA box may play an important role in the expression and regulation of genes. Further experiments are needed in order to determine the real function of TTTTA box.

Discussion

We successfully constructed a silkworm pupa cDNA library, in which the size of cDNA varies from 100 to 4000 bp. In addition, we discovered some noncoding mRNAs for regulation. Three percent of those belong to snoRNAs (noncoding mRNA) that have a poly(U) in the 5' end. Blast analysis with the silkworm pupa EST database showed that 20.7% new cDNA

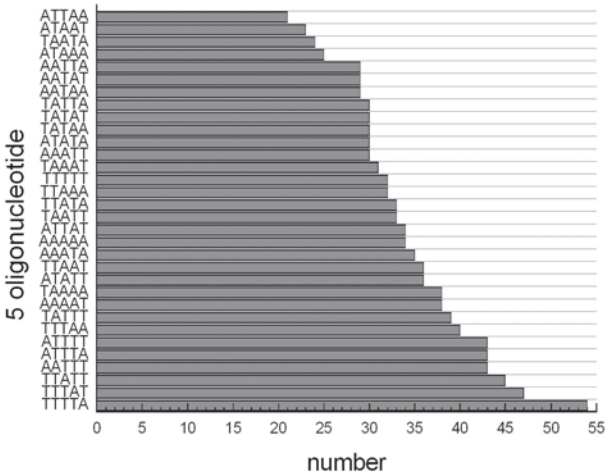


Fig. 7. Numbers of occurrence for all 5-mers with A and T in first group.

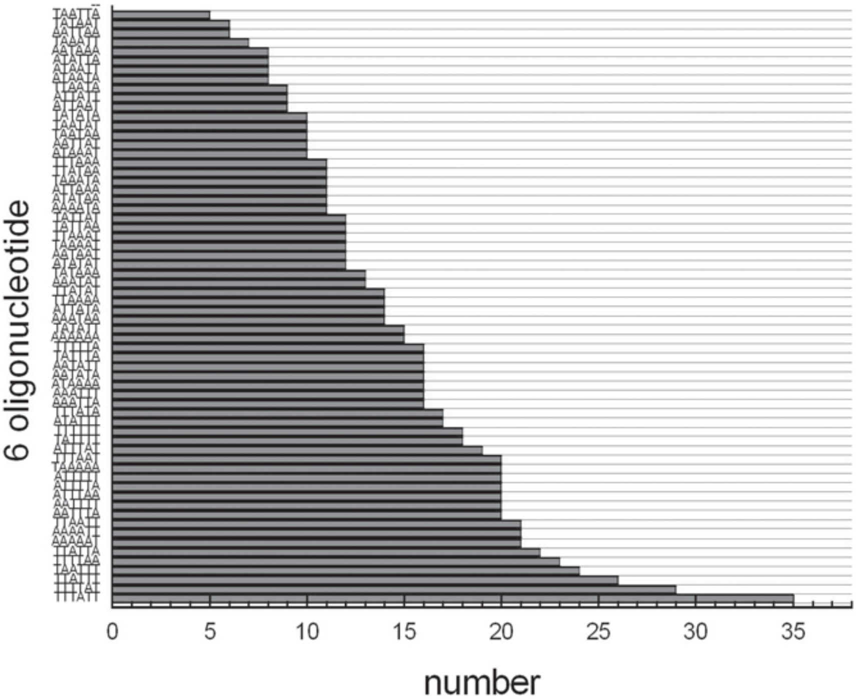


Fig. 8. Numbers of occurrence for all 6-mers with A and T in first group.

sequences and 9% unknown cDNA had been identified, which indicated that cDNA in the library was abundant. This method overcomes the disadvantage that some small cDNAs often are lost in constructing cDNA library *(13)*.

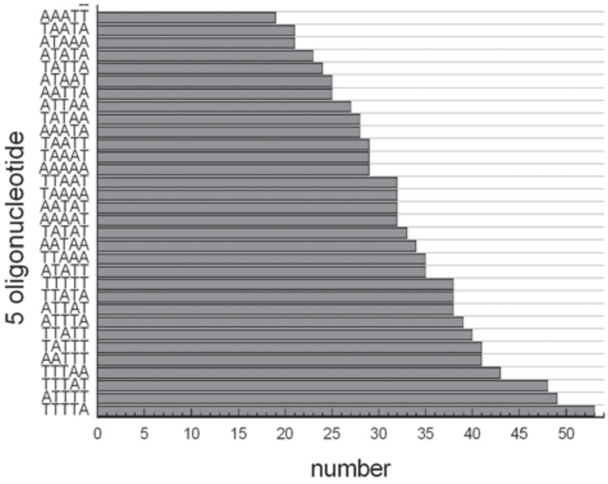


Fig. 9. Numbers of occurrence for all 5-mers with A and T in second group.

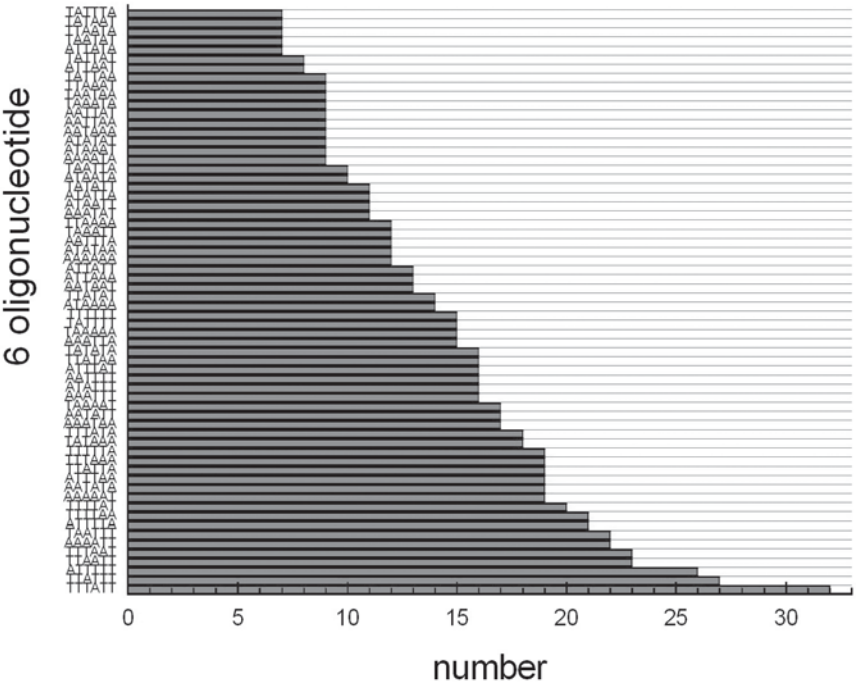


Fig. 10. Numbers of occurrence for all 6-mers with A and T in first group.

About 1659 full-length cDNAs were discovered in the library, including 492 noncoding full-length cDNAs. The forecasts of ORF in these cDNAs showed that each cDNA had two ORFs. The data contributed to the further study of protein for *B. mori*. Statistical analysis indicated that the first nucle-

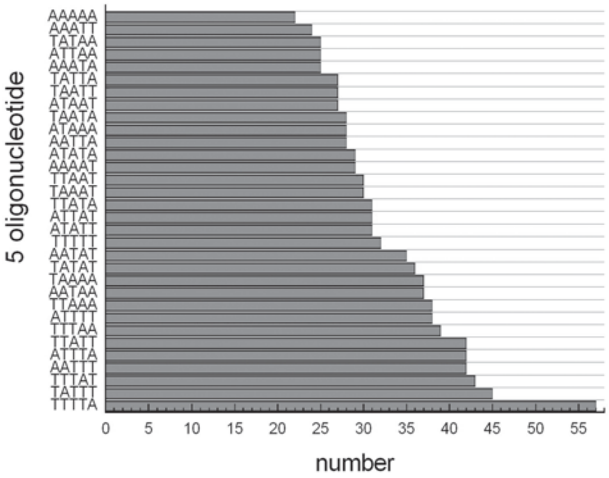


Fig. 11. Numbers of occurrence for all 5-mers with A and T in third group.

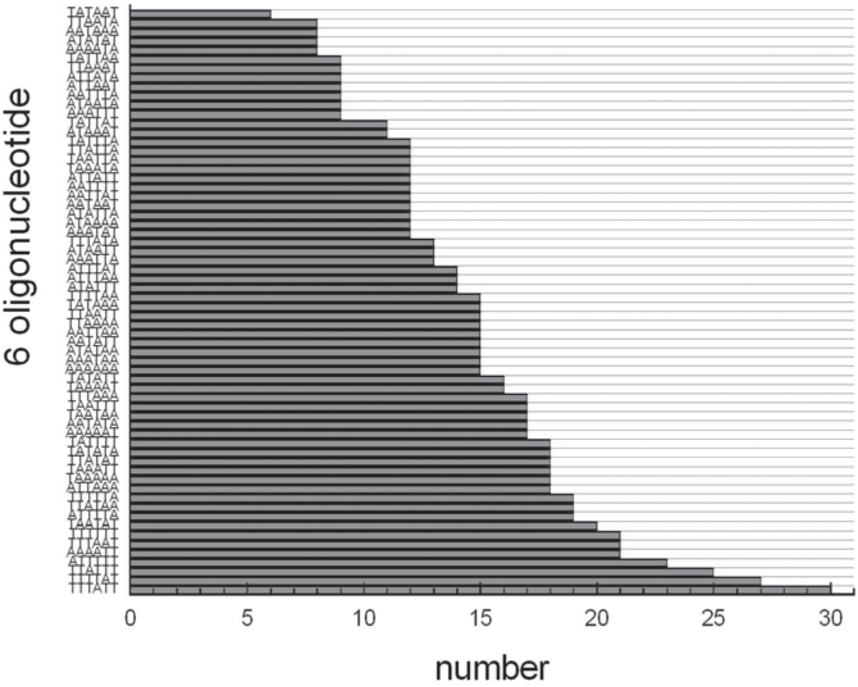


Fig. 12. Numbers of occurrence for all 6-mers with A and T in third group.

otide in the 5' end includes C (37%), A (23%), G (24%), and U (16%), and this phenomenon occurred in many cDNAs, which indicates that the structures of the mRNA 5' end might be variable, not unchangeable, and might relate to the vehement spallation.

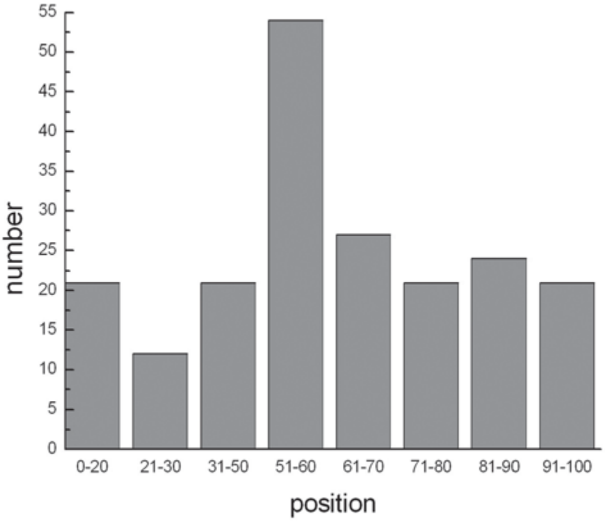


Fig. 13. Position of TTTTA.

We found that the mode of mRNA splicing and the pattern of intron are various, and that the characteristic of the transcription region is nearly clear; these findings will be published in another article.

We also found the mode of promoter based on a statistical method. We believe that the TTTTA box is possibly a functional signal in eukaryote RNA expression and regulation.

Through the analysis of some expression and nonexpression ORF sequences, we found that the G+C contents in the upstream and downstream region of nonexpression ORF sequences and expression ORFs are different and contain repeat sequences in nonexpression ORFs.

Some scholars deem that there is no terminator for the transcription of eukaryote mRNA because the start site of transcription is clear and the terminator is not clear *(14–17)*. There have been many reports about protein factors for eukaryotic transcription terminator. However, the signal sequence for the transcription stop site in genome has not been found to date *(14–17)*. According to the rule that poly(A) is appended to the 3' end of mRNA *(18)*, we found that there was an AATAAA box before poly(A), and accurate addition of most poly(A) was owing to the presence of AATAAA box. There is a distance of 10–30 nt between AATAAA box and the tail *(19–21)*. Some cDNA had two or even more AATAAA boxes, as previously reported *(22)*. Why do eukaryote mRNAs terminate transcription and add tail behind the AATAAA box? This phenomenon indicates that AATAAA box was the signal not only for tail appending, but also for mRNA termination; otherwise the transcription could not effectively terminate. Why do some mRNAs have several AATAAA boxes in the 3' end of mRNA? We find that several signals contribute to speed-down and termination of mRNA transcription. Behind the AATAAA box there are 10–30 nt, which is a suitable region for RNA polymerase binding in efficiently transcripted genes. For the mRNA with one AATAAA box, the 12 nt make the AATAAA box locate at the center of RNA polymerase, thereby terminating transcription. We believe that AATAAA box is one of the termination signals in eukaryote RNA transcription, and that the copy number of mRNA is dependent on the number of AATAAA box in series.

Acknowledgments

This study was supported by grants from the National Natural Science Foundation of China (no. 30470034), the National Basic Research Program of China (no. 2005CB121006), and the National High Technology Research and Development Program of China (863 Major Program) (2005BA711A07).

References

1. Venter, J. C., Adams, M. D., Myers, E. W. et al. (2001), *Science* **291,** 10,304–10,351.
2. Goff, S. A., Ricke, D., Lan, T. H., et al. (2002), *Science* **296,** 92–100.
3. Eisenberg, D., Marcotte, E. M., Xenarios, I., Yeates, T. O. (2002), *Nature* **405,** 823–826.
4. Sasaki, I., Matsumoto, T., Yamamoto, K., et al. (2002), *Nature* **420,** 312–316.
5. Adams, M. D., Celniker, S. E., Holt, R. A., et al. (2000), *Science* **287,** 2185–2195.
6. Mita, K., Kasahara, M., Sasaki, S. (2004), *DNA Res.* **11,** 27–35.
7. Xia, Q., Zhou, Z., Lu, C., et al. (2004), *Science* **306,** 1937–1940.
8. Goldsmith, M. R. (1995), in *Molecular Model Systems in the Lepidoptera*. Goldsmith, M. R. and Wilkins, A. S. ed., Cambridge University Press, pp. 21–76.
9. Mita, K., Mormyo, M., Okano, K., et al. (2003), *Proc. Natl. Acad. Sci. USA* **100,** 14,121–14,126.
10. Scheetz, T. E., Trivedi, N., Roberts, C. A., et al. (2003), *Bioinformatics* **19,** 1318–1324.
11. Blackshear, P. J., Lai, W. S., Thorn, J. M., et al. (2001), *Gene* **267,** 71–87.
12. Strausberg, R. L., Feingod, E. A., Grouse, L. H., et al. (2003), *Proc. Natl. Acad. Sci. USA* **99,** 16,899–16,903.
13. Haas, S., Vingron, M., Wiemann, S. (1998), *Nucleic Acids Res.* **26,** 3006–3012.
14. Karamysheva, Z. N., Karamyshev, A. L., Ito, K., et al. (2003), *Nucleic Acids Res.* **31,** 5949–5956.
15. Poole, E. S., Askarian-Amiri, M. E., Major, L. L., et al. (2003), *Prog. Nucleic Acid Res. Mol. Biol.* **74,** 83–121.
16. Chavatte, L., Kervestin, S., Favre, A., et al. (2003), *EMBO* **22,** 1644–1653.
17. Salas-Marco, J. and Bedwell, D. M. (2004), *Mol. Cell Biol.* **24,** 7769–7778.
18. Colgan, D. F. and Manley, J. L. (1997), *Genes Dev.* **11,** 2755–2766.
19. Chen, F., MacDonald, C. C., Wilusz, J. (1995), *Nucleic Acids Res.* **23,** 2614–2620.
20. Beaudoing, E., Freier, S., Wyatt, J. R., et al. (2000), *Genome Res.* **10,** 1001–1010.
21. Edwalds-Gilbert, G., Veraldi, K. L., Milcarek, C. (1997), *Nucleic Acids Res.* **25,** 2547–2561.
22. Graber, H. J., Cantor, C. R., Mohr, S. C., et al. (1999) *Proc. Natl. Acad. Sci. USA* **23,** 14,055–14,060.