Artif Life Robotics (2007) 11:135–138
DOI 10.1007/s10015-006-0415-7

ORIGINAL ARTICLE

Makoto Ohkubo · Futoshi Aranishi

A functional motif discovery algorithm for invertebrate EST sequence data

Received and accepted: July 20, 2006

Abstract We have developed a new algorithm for invertebrate expressed sequence tag (EST) analysis, termed as the fmEST algorithm, which consists of a systematic homology search, functional motif scanning, and clustering alignment. This study was undertaken to evaluate the validity of our fmEST algorithm in functional motif discovery for invertebrate EST sequence data. Out of 200 unidentified invertebrate ESTs, including 100 arthropod ESTs and 100 mollusk ESTs, 18 arthropod ESTs and 21 mollusk ESTs were identified as fmESTs that contained functional motifs. The nucleotide lengths of arthropod fmEST and mollusk fmEST sequences were distributed from 388 to 954 bp and from 222 to 742 bp, respectively. This result allowed us to annotate these invertebrate fmESTs as various functional genes, while they showed no significant homology to the gene information recorded in the international DNA databases using the conventional BLAST homology search program. In addition, another 1 arthropod EST and 23 mollusk ESTs were assembled into contigs with any identified fmESTs by clustering alignment. Based on these findings, we have concluded that our fmEST algorithm, involving the functional motif discovery procedure, is a valuable approach, enabling us to break new ground in undeveloped invertebrate EST analysis.

Key words Expressed sequence tag analysis · Invertebrate · Functional motif · fmEST

M. Ohkubo · F. Aranishi (✉)
Department of Biological and Environmental Sciences, Miyazaki University, 1-1Gakuenkibanadai-Nishi, Miyazaki 889-2192, Japan
Tel. +81-985-58-7224; Fax +81-985-58-2884
e-mail: aranishi@cc.miyazaki-u.ac.jp

This work was presented in part at the 11th International Symposium on Artificial Life and Robotics, Oita, Japan, January 23–25, 2006

1 Introduction

Expressed sequence tag (EST) analysis is a mining method for physiological mechanisms by a comprehensive analysis of the expressed genes. EST is the nucleotide sequence information of cDNA that is a complementary single strand of DNA to mRNA expressed in vivo. Molecular biologists used to identify ESTs as the functional gene homologues by computational comparisons of EST sequence data with the gene information recorded in the international DNA databases such as DDBJ (http://www.ddbj.nig.ac.jp/), EMBL (http://www.ebi.ac.uk/embl/), and Genbank (http://www.ncbi.nlm.nih.gov/Genbank/). These findings allowed a survey of the physiological mechanisms occurring in different taxonomical levels of animal, organ, tissue, or cell.[1]

An EST dataset for analyzing the expressed genes usually contains from 500 to 2000 EST sequences. These large numbers in gene information are applicable to broad scientific fields, such as medicine, pharmacology, microbiology, and phylogeny.[2–5] To date, functional genomic studies on invertebrate taxa have increasingly been required from physiological, ecological, and evolutional interests.[6–11] However, gene information on microorganisms and vertebrates occupies the majority of the present databases.[12] In addition, the molecular structures of invertebrate genes differ considerably from those of vertebrate genes. We have developed a new algorithm for invertebrate EST analysis, termed as the fmEST algorithm, which combines functional motif scanning with the traditional systematic homology search and clustering alignment.[12] The functional motifs are the functionally important secondary structures of proteins, such as receptor binding domains, immunological epitopes, and catalytic sites of enzymes. They are encoded by short nucleotide sequences (generally less than 20 nucleotides), and are highly conserved during evolution. Therefore, the functional motifs are a significant clue to annotate the large invertebrate EST dataset.[13]

This study was undertaken to evaluate the validity of our fmEST algorithm in functional motif discovery for 200 unidentified invertebrate ESTs, including 100 arthropod ESTs

and 100 mollusk ESTs, all of which showed no significant homology to the gene information recorded in the international DNA databases using the conventional BLAST homology search program.

2 fmEST algorithm

The traditional systematic approach to EST analysis consists of a BLAST homology search and subsequent CLUSTAL W[14] multiple alignment of EST sequences to the gene information recorded in the international DNA databases. Several researchers into invertebrate EST analysis have reported that these traditional procedures enabled us to identify a limited part of the EST dataset as the functional gene homologues,[6–11] because there is less gene information on invertebrates than on vertebrates recorded in the DNA databases, and there are divergent molecular structures between invertebrate and vertebrate genes. The BLAST homology search program allows us to obtain a little information about the functional genes for certain invertebrate EST sequence data based only on short homologous nucleotide sequences.

The fmEST algorithm consists of three computational analytical steps: a systematic homology search, functional motif scanning, and clustering alignment (Fig. 1). Following a BLAST homology search, the nucleotide sequences of unidentified ESTs were deduced to amino acid sequences and applied to one of the functional motif search programs, such as InterProScan (http://www.ebi.ac.uk/InterProScan/), PROSITE (http:// www.expasy.org/prosite/), or Pfam (http://www.sanger.ac.uk/Software/Pfam/). During subsequent CLUSTAL W multiple alignment, nucleotide sequences determined to contain the same functional motif were assembled into a single contig, which was the cDNA transcribed from a single mRNA.

For example, an unidentified EST (GenBank accession CI999386) was obtained from the kuruma prawn *Marsupenaeus japonicus*. Although this EST showed no significant homology to the carboxypeptidase gene of the sea hare *Aplysia californica* (GenBank accession U37755) by the BLAST homology search (Score = 46.1, Expect value = 0.079), the prawn EST was found to contain the functional motif of zinc carboxypeptidase (InterProScan accession IPR000834) by functional motif scanning. There were two unique amino acid sequences, i.e., PEFKYVANMHGNEV and LPSMNPDGWQ, both of which were zinc-binding sites involved in the catalysis of carboxypeptidases. When the deduced amino acid sequences were then compared between the prawn EST and the sea hare carboxypeptidase by CLUSTAL W multiple alignment, these local short segments were highly conserved (Fig. 2). This result enabled us to annotate the unidentified prawn EST as a homologous gene encoding the zinc carboxypeptidase of sea hare.

3 Invertebrate EST analysis

A total of 200 unidentified invertebrate ESTs (100 arthropod ESTs, i.e., 50 crustacean ESTs plus 50 insect ESTs, and 100 mollusk ESTs, i.e., 50 bivalve ESTs plus 50 gastropod ESTs) were obtained from the Genbank dbEST database under the accession numbers from CI999365 to CI999414 in kuruma prawn *M. japonicus*, from DT777319 to DT777368 in flour beetle *Tribolium castaneum*, from DR409845 to DR

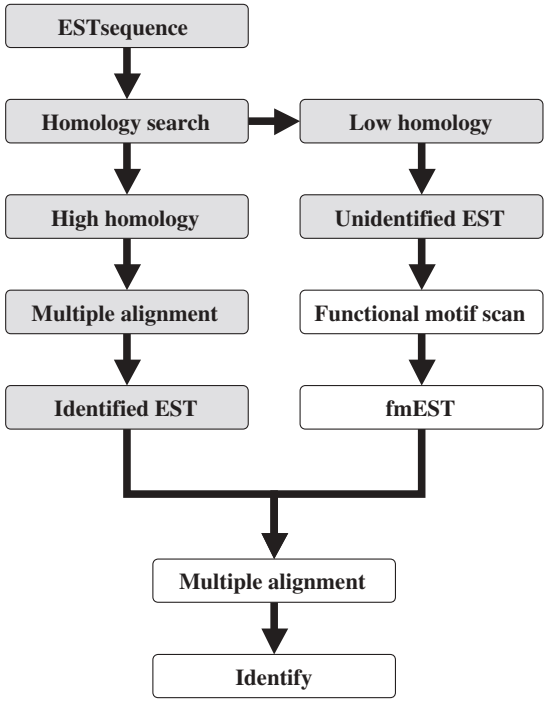

Fig. 1. Schematic procedure of the fmEST algorithm. *Gray boxes* indicate the traditional procedure of EST analysis

Fig. 2. Alignment of deduced amino acid sequences of the kuruma prawn EST (Genbank accession CI999386) and sea hare carboxypeptidase (Genbank accession U37755)

```
Kuruma prawn EST          -----MDTSKMLS--QCVLVFAAAVAIVSYQSIEASQDTEENGSKKTESTIFFHHTYEE
Sea hare carboxypeptidase MCVHSSADTMKYCWGHVGVLLLLLVASVCVQAASVAKPGTSGNGTPSKSEFVFKHHNNEE
                                ** *       **:: . .::    :  *. .*. **: .... :* **. **

Kuruma prawn EST          MVSLMYEVNKACPEVTRIYNLSEPSVEKRNLTVLEITENPGVHVPGKPEFKYVANMHGNE
Sea hare carboxypeptidase LEQVLRETAEKCKDVTRLYALSEPSVRNVPLWVIEFSDNPGQHDLLEPEFKYVANMHGNE
                          : .:: *. : * :***:* ******.:  * *:*::*** *   :**************

Kuruma prawn EST          VVGKEMVLYFLVALCEEYKRGDKLANFIVSQTRVHVLPSMNPDGWQKAYKELQEKGEAGW
Sea hare carboxypeptidase VLGRELLLALAHYLCQGWKDGDEEIKKLIKSTRIHLLPSMNPDGWQLATDTGGKDYLR—
                          *:*:*::* :   **: :* **:  : ::.. **:*:********** * .   :.
```

Table 1. Summary of EST datasets for functional motif scanning

Phylum	Class	Species	No. of clones	Nt length (average)	Accession No.[a]
Arthropoda	Crustacea	Kuruma prawn (*Marsupenaeus japonicus*)	50	152–678 (480)	CI999365–CI999414
	Insecta	Flour beetle (*Tribolium castaneum*)	50	644–1025 (890)	DT777319–DT777368
Mollusca	Bivalvia	Akoya pearl oyster (*Pinctada fucata*)	50	173–592 (350)	DR409845–DR409894
	Gastropoda	Snail (*Biomphalaria glabrata*)	50	147–743 (518)	DT725365–DT725414

[a] Accession number on the Genbank dbEST database

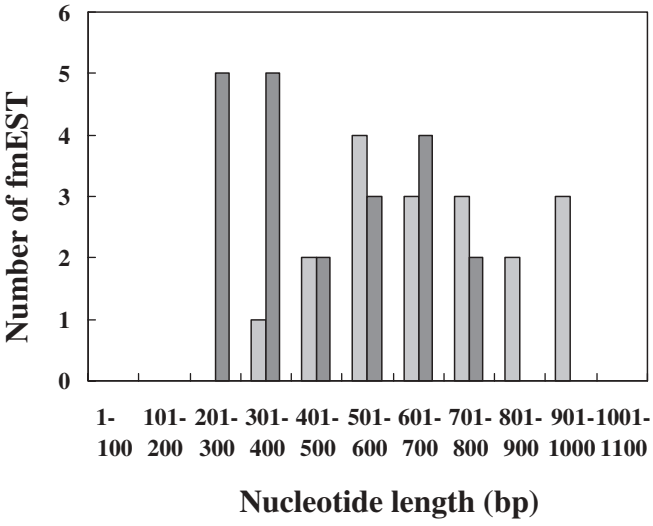

Fig. 3. Distribution of the nucleotide length of arthropod ESTs (*light bars*) and mollusk ESTs (*dark bars*) containing some functional motifs

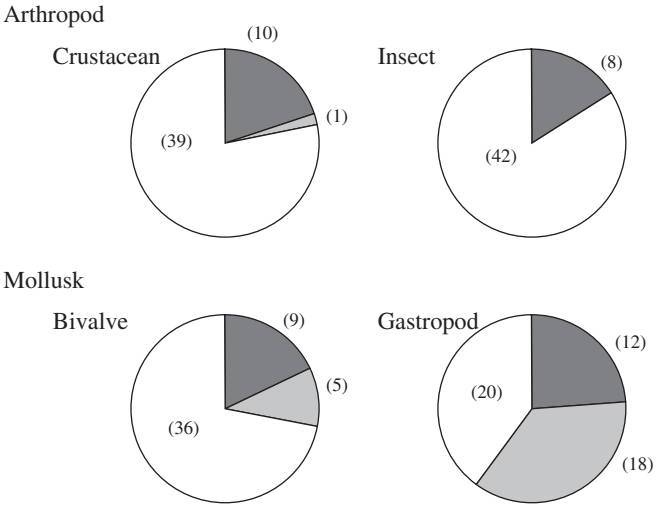

Fig. 4. Proportion of ESTs containing some functional motifs (fmEST, *dark shading*). ESTs assembled into fmEST contigs (*light shading*) and unidentified ESTs (*white*) in the arthropod and mollusk EST datasets analyzed. The number of ESTs is shown in parentheses

409894 in akoya pearl oyster *Pinctada fucata*, and from DT725365 to DT725414 in snail *Biomphalaria glabrata*, respectively (Table 1).

Functional motif scanning was carried out interactively on the web using the InterProScan functional motif search program operated by the European Bioinformatics Institute. The option of a "translation table" was set to "Standard Code" for all ESTs. The option of the "minimum open reading frame size" was set to 20, 50, 100, and 150 for individual ESTs. The other parameters were set to default values. Subsequent clustering alignment was carried out using the TIGR gene indices clustering tools (TGICL) with a default setting.[15]

In general, the nucleotide lengths of EST sequences obtained from various organisms are distributed from 100 to 1000 bp. Those of the 100 arthropod EST and 100 mollusk EST sequences were distributed from 152 to 1025 bp and from 147 to 743 bp, respectively. The average nucleotide length of the 50 flour beetle EST sequences was estimated to be 890 bp, which was higher than that of other organisms analyzed (Table 1). Insect taxa have so far been the most well understood invertebrate class in recent molecular biology,[16] and their ESTs with long nucleotide sequences can be prepared more easily than those of other invertebrate taxa. In contrast, the average nucleotide length of the 50 akoya pearl oyster EST sequences was lower than that of other organisms analyzed (Table 1). This is probably because much less information is available on molecular biological techniques for bivalves.

The nucleotide lengths of 18 arthropod EST and 21 mollusk EST sequences containing same functional motifs were distributed from 388 to 954 bp and from 222 to 742 bp, respectively (Fig. 3). This result indicates that the functional motifs are difficult to find in short EST sequences less than approximately 200 bp. Several studies on invertebrate EST analysis have reported that the EST dataset generally contains from 500 to 2000 ESTs, and approximately 20% to 30% of these can be identified as functional genes using the conventional BLAST homology search program.[6–11] In this study, our fmEST algorithm enabled us to identify an additional 39 ESTs containing some functional motifs from 200 unidentified invertebrate ESTs (Fig. 4). In addition, another 1 arthropod EST and 23 mollusk ESTs were assembled into contigs, with some fmESTs identified by clustering alignment, while no functional motifs were observed in these 24 EST sequences. The results allowed us to annotate an additional 63 unidentified ESTs as various functional genes, which corresponded to approximately 30% of invertebrate ESTs analyzed.

4 Conclusion

This study demonstrated the validity of our fmEST algorithm in functional motif discovery for invertebrate EST sequence data. We concluded that our fmEST algorithm, combining functional motif scanning with the traditional systematic homology search, is a valuable approach, enabling us to break new ground in undeveloped invertebrate EST analysis. Although the BLAST homology search program is equipped to deduce nucleotide sequences to amino acid sequences automatically, it has no ability to refer to functional motifs. In addition, the conventional automated EST annotation programs are not incorporated with a functional motif search program.[17] Molecular biologists thus have a difficult task to discover functional motifs from large numbers of EST sequences. A novel form of automated EST annotation system, which incorporates our fmEST algrithm, is proposed to help molecular biologist investigating a wide variety of invertebrate taxa which are involved in familiar human sciences such as agriculture, fisheries, and ecology.

Acknowledgments The authors thank Prof. Ikuo Yoshihara, Department of Computer Science and Systems, Miyazaki University, for his valuable support in manuscript preparation. This work was supported in part by a grant from the Agriculture, Forestry and Fisheries Research Council of Japan to FA.

References

1. Asamizu E, Nakamura Y (2004) System for expressed sequence tags (EST) analysis (in Japanese). Tanpakushitsu Kakusan Koso 49:1847–1852
2. Angus AC, Ong ST, Chew FT (2004) Sequence tag catalogs of dust-mite-expressed genomes: utility in allergen and acarologic studies. Am J Pharmacogenomics 4:357–369
3. Samuels AK, Weisrock DW, Smith JJ, et al. (2005) Transcriptional and phylogenetic analysis of five complete ambystomatid salamander mitochondrial genomes. Gene 349:43–53
4. Evans DL, Kaur H, Leary J III, et al. (2005) Molecular characterization of a novel pattern recognition protein from nonspecific cytotoxic cells. Dev Comp Immunol 29:1049–1064
5. Faria-Campos AC, Cerqueira GC, Anacleto C, et al. (2003) Mining microorganism EST databases in the quest for new proteins. Genet Mol Res 31:169–177
6. Gross PS, Bartlett TC, Browdy CL, et al. (2001) Immune gene discovery by expressed sequence tag analysis of hemocytes and hepatopancreas in the Pacific white shrimp, *Litopenaeus vannamei*, and the Atlantic white shrimp, *L. setiferus*. Dev Comp Immunol 25:565–577
7. Lehnert SA, Wilson KJ, Byrne K, et al. (1999) Tissue-specific expressed sequence tags from the black tiger shrimp *Penaeus monodon*. Mar Biotechnol 1:465–476
8. Supungul P, Klinbunga S, Pichyangkura R, et al. (2004) Antimicrobial peptides discovered in the black tiger shrimp *Penaeus monodon* using the EST approach. Dis Aquat Organ 61:123–135
9. Jenny MJ, Ringwood AH, Lacy ER, et al. (2002) Potential indicators of stress response identified by expressed sequence tag analysis of hemocytes and embryos from the American oyster, *Crassostrea virginica*. Mar Biotechnol 4:81–93
10. Lee YH, Huang GM, Cameron RA, et al. (1999) EST analysis of gene expression in early cleavage-stage sea urchin embryos. Development 126:3857–3867
11. Ljunggren EL, Nilsson D, Mattsson JG (2003) Expressed sequence tag analysis of *Sarcoptes scabiei*. Parasitology 127:139–145
12. Ohkubo M, Aranishi F (2004) What does invertebrate EST analysis acquire from bioinformatics? In: Kim J-H (ed) Simulated evolution and learning. KAIST, Daejeon, Korea, Genome Informatics, pp 1–5
13. Whisstock JC, Lesk AM (2003) Prediction of protein function from protein sequence and structure. Q Rev Biophys 36:307–340
14. Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–4680
15. Pertea G, Huang X, Liang F, et al. (2003) TIGR gene indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. Bioinformatics 19:651–652
16. Boutros M, Perrimon N (2000) Drosophila genome takes flight. Nat Cell Biol 2:E53–E54
17. Hotz-Wagenblatt A, Hankeln T, Ernst P, et al. (2003) EST annotator: a tool for high throughput EST annotation. Nucleic Acids Res 31:3716–3719