Mining *Plutella xylostella*’s expressed sequence tags (EST) for a functionally annotated candidate gene index

Dinakar R. Ampasala, M. Pritika, Mohane Selvaraj Coumar and Amouda Venkatsan*

Centre for Bioinformatics, Pondicherry University, Kalapet, Pondicherry-605 014, India.

Received: 29-11-2014 Accepted: 27-07-2015 DOI: 10.18805/ijare.v50i1.8584

**ABSTRACT**

Diamondback moths, *Plutella xylostella* (Linnaeus) are one of the major pests of cruciferous plants such as cabbage and cauliflower (*Brassica oleracea*) in India. These insects show resistance against commonly used pest management practices such as, chemical insecticides and bacterial pathogens, including *Bacillus thuringiensis* (Bt). To overcome the resistance problem, develop new pesticides that can act through specific drug targets. To identify such drug targets specific to this insect, data mining and annotation of Expressed Sequence Tags (ESTs) were performed in this study. Expressed Sequence Tags (ESTs, 37,915) of the insect obtained from GenBank were clustered and consensus sequences (4224) were constructed using mining tool (CAP3). Out of it 256 sequences were functionally annotated using three Gene Ontologies (GO) terms, molecular function, biological process and cellular component using similarity search methods (BLASTOGO). By mapping the candidate genes to KEGG Pathway, 38 insect metabolic pathways, inclusive of xenobiotic metabolism by Cytochrome P450 were generated. One of the mapped candidate gene codes for aldehyde dehydrogenase enzyme, which is potentially involved in xenobiotic detoxification of synthetic insecticides, and play a role in the development of resistance to pesticides. Data mining and functional annotation helped to narrow down the choices for potential drug targets, which could aid in the development of new pesticides to overcome resistance in Diamondback moths. This methodology can be extended to other agriculturally important pests to identify the drug targets using EST data.

**Key words:** Data mining, EST, Functional annotation, Gene ontology, KEGG.

**INTRODUCTION**

Diamondback moths, *Plutella xylostella* (Linnaeus) of the Order Lepidoptera, feed heavily on cruciferous plants (*Brassica oleracea*) such as cabbage and cauliflower (Patil and Pokharkar, 1971; Singh and Singh, 1982). Extensive damage to the crop is caused by them during their larval stage. They migrate long distances and thrive in temperate and tropical climate, hence making them one of the most widely distributed pests around the world, including India (Harcourt, 1957).

Currently, use of synthetic insecticides such as pyrethroids (Saxena *et al*., 1989) and organophosphates are ineffective against this insect, as they have gained resistance (Regupathy; Chawla and Joia, 1992). Moreover, widespread uses of broad spectrum insecticides have also led to decrease in population of pest’s natural predators and parasitoids. In addition, the moths are also showing resistance against biological control methods such as bacterial insecticide, *Bacillus thuringiensis* (Bt) (Kirsch and Schmutterer, 1988). This has made the diamondback moth a formidable pest to deal with.

To tackle the pest problem, novel drug target identification from the insect could lead to the development of new pesticides which can overcome the resistance (Pang *et al*., 2012). Use of these pesticides can further be incorporated with Integrated Pest Management (IPM) practices such as inter-cropping, spraying of jaggery (sugar) solution to encourage the activity of predatory ants and the removal of the old leaves where the majority of pupation occurs (Chelliah and Srinivasan, 1986).

To carry out drug target identification, knowledge of essential genes or proteins of the organism is necessary. The Arthropod Genome Consortium (www.arthropodgenomes.org) contains a collection of genomic resources for Arthropods including Lepidoptera. Here, the abundantly available genomic data listed for diamondback moths were found to be gene expression data in the form of Expressed Sequence Tags (ESTs). ESTs which represent portions of the genes expressed in a tissue are short sequenced portions of complementary DNA (cDNA) obtained by reverse transcription of mRNA (Adams *et al*., 1991; Hatey *et al*., 1998). Most importantly, they are used in creating gene indices of organisms with little or no gene data.

Gene indices are the inventories of genes and their corresponding functions. Data mining technique is used to create gene indices from ESTs data, which uses different techniques to cluster and classify large amounts of
data. Using clustering technique, the large data set of ESTs from the insect *Plutella xylostella* were converted to candidate genes for functional annotation using Gene Ontology (GO) terms to create a gene index. Using this index, potential drug targets can be selected, modelled and studied to design new pesticides which can act against the diamondback moth. Aldehyde dehydrogenase from Cytochrome P450 metabolic pathway is identified as a potential drug target which is involved in the development of resistance to pesticides. Identification and development of potent chemical pesticides using this target protein may overcome the resistance problem.

**MATERIALS AND METHODS**

**Retrieval of EST sequences:** EST sequences of *Plutella xylostella* were retrieved from NCBI’s dbEST (http://www.ncbi.nlm.nih.gov/dbEST/). A total of 37915 sequences were taken from the database in FASTA format.

**Retrieval of vector sequences:** For sequencing experiments, nucleotide sequences of vectors were taken from NCBI’s UniVec database (ftp://ftp.ncbi.nih.gov/pub/UniVec/). A total of 4383 vector sequences were downloaded in the FASTA format and saved in a separate file.

**Pre-processing EST sequences:** To remove contaminating vector regions, the EST sequences were cleaned using Crossmatch data mining tool (http://www.crossmatch.com/). Where EST sequences are compared with vector sequences using Smith Waterman algorithm for pairwise alignment. Here a banded search was performed between the set of input sequences to find pairs of exactly matching sub sequences, also called word matches. Vector matches found in the EST sequences were masked by replacement of the bases in the EST sequence with an ‘X’. The scoring parameters were set to default with minimum word / band match of 10 and minimum Smith Waterman matrix score of 20.

**Clustering and assembling of EST’s to generate consensus sequences:** CAP3 program (http://pbil.univ-lyon1.fr/cap3.php) was used to cluster and assemble the cleaned EST sequences into consensus sequences. The following steps are followed by the algorithm for clustering:

1. Low quality 5’ and 3’ regions of the cleaned EST reads were clipped.
2. Clustering the pairs of reads based on computation of overlaps between them. An overlap between two reads is global alignment of the reads with maximum similarity score.
3. Each pair of overlapping reads was then joined to construct individual contigs.
4. Multiple sequence alignment of the contigs was done to generate consensus sequences.

**Annotation of the consensus sequences based on GO terms:** The set of consensus sequences obtained from the previous step represent portion of the genes which are transcribed in the organism. To understand the role of these genes, the consensus sequences are functionally annotated using BLAST2GO tool, version 2.6.6 (http://www.blast2go.com/b2ghome). For this purpose, the sequences which are homologous to the query (input consensus sequence) were identified from NCBI’s nr databases using BLASTx alignment program. The query sequences are translated from nucleotide to protein sequences and search for homolog proteins in the database. To retrieve the significantly similar sequences, e-value threshold of 1.0E-3 was assigned and hit number threshold of 20 were considered.

Protein IDs of hits are identified from the database to link or to map the different IDs with corresponding IDs of Gene Ontology database (http://geneontology.org/). Since each query sequence is associated with several GO annotations, an Annotation Rule (AR) is used to calculate an Annotation Score (AS) by BLAST2GO. The highest scoring GO term was used as definitive functional annotation of the query sequence. To annotate the queries with high e-values the threshold was set at 1.0E-6. Sequences homologous to humans, cabbage, cauliflower and its natural predators including ants and yellow wagtail were not included in functional annotation. Figure 1 shows the overall methodology used for functional annotation of EST sequence.

**RESULTS AND DISCUSSION**

Out of 37,915 ESTs of *Plutella xylostella*, 4224 consensus sequences were constructed using CAP3 and 256 of them were functionally annotated. Different levels of abstraction were represented pictorially using Directed Acyclic Graphs (DAGs). These DAGs were generated for the main GO terms molecular function, biological process and cellular component. The DAGs are tree like structures
containing nodes which represent specific GO terms. In the DAGs generated for molecular function (Figure 2), biological process (Figure 3), cellular component (Figure 4), the parent nodes are the GO terms.

Based on the number of sequences at each node of DAG, a cut off value of five per node was assigned to generate overall consensus sequence annotation based on molecular function, biological process and cellular component. The maximum annotations associated with GO molecular function were nucleotide binding, protein binding, transporter activity, translation factor activity, structural molecule activity, receptor activity, protein kinase activity, peptidase activity, calcium ion binding, DNA binding, enzyme regulator activity and lipid binding (Figure 5).

The maximum annotations associated with GO term biological process were cell differentiation, cellular protein modification process, translation, signal transduction, secondary metabolic processes, response to stress and external stimuli, reproduction, protein transport, lipid metabolic process, ion transport, growth, generation of precursor metabolites and energy, embryo development, cytoskeleton development, cellular protein modification process, DNA metabolic process, anatomical structure morphogenesis, behaviour, carbohydrate metabolic process, catabolic process, cell cycle, cell death, cell proliferation, cell-cell signalling and cellular homeostasis (Figure 6). The maximum annotations associated with GO term cellular component were protein complex, vacuole, plasma membrane, nucleoplasm, chromosome, cytoskeleton, cytosol, endoplasmic reticulum, lipid particle, mitochondrion and nuclear envelope (Figure 7).

The parent node - molecular function, biological process and cellular component - represents the first level of abstraction. The GO term ‘binding’ represent separate node under the parent node ‘molecular function’ which
Functional annotation of the diamondback moth’s EST consensus sequences was done by using the GO annotation of top hit homologs in the nr database. A total of 256 (6%) sequences out of 4224 were annotated for three GO domains and mapped to KEGG pathway, resulting in the identification of 38 metabolic pathways. One of the KEGG maps generated was the metabolism of xenobiotic by Cytochrome P450 containing a single consensus sequence from the candidate gene list (Fig 8). This consensus sequence was mapped to the enzyme aldehyde dehydrogenase which is potentially involved in xenobiotic detoxification including synthetic insecticides. It is reported in *Aedes aegypti* that the aldehyde dehydrogenase is actively involved in detoxification of pyrethroid and development of insecticide resistance. Hence it is inferred that aldehyde dehydrogenase could lead to the development of resistance in diamondback moth. Therefore, targeting aldehyde dehydrogenase with small molecule insecticides may overcome the resistance problem in this moth.
FIG 8: KEGG pathway for Xenobiotic metabolism by Cytochrome P450

As, majority of the consensus sequences were annotated based on their similarity with Lepidopteran non-pest species such as Danaus plexippus (Monarch butterfly), Papilio xuthus (Asian Swallowtail) and Bombyx mori (Silk moth), the selection of potential drug target from the candidate gene list for pesticide targeting in the diamondback moth may have the propensity to kill these non-target organisms leading to a disturbance in their ecosystem. However, in case of aldehyde dehydrogenase the BLASTx e-values for Danaus plexippus and Bombyx mori were found to be zero and the hits did not contain any of Papilio xuthus sequences. Therefore choosing this enzyme as a target reduces the chances of affecting non-target organisms considerably.

A caveat of using this methodology for functional annotation is the assumption that similar sequences mean similar function. This does not always hold true and hence require structural studies involving protein interactions for validation. Also, information of which tissue and stage of moth’s life cycle, the candidate genes are expressed is lost during EST assembly. Moreover, the final set of functionally
annotated ESTs is not a complete representation of the genes present in the moth. The candidate gene list though can be increased by taking into account the singlets produced by CAP3 in which overlaps were absent or not matching threshold values. The percentage of annotated sequences can be increased by adjustment of threshold values in the BLASTx and annotation step. However, e-value of the hits in the range of 1.0E-3 to 1.0E-6 has been used for valid annotation of assembled ESTs.

CONCLUSION

Recently, the whole genome sequence of the diamondback moth has been sequenced (You et al., 2013; Dai et al., 2014). The clustered and assembled ESTs can therefore be aligned with this sequence for genome mapping and gene discovery. They can be used for identification of molecular markers such as SNPs (Single Nucleotide Polymorphisms) to study the genetic basis of insect diversity and mapping of important genes such as those corresponding to insecticide resistance. The candidate gene sequences can also be used to design probes for microarray experiments to infer and validate gene expression based on tissue and stage of the moth. Phylogenetic analysis of the moth with other Lepidopteran species for genomic studies in adaptive radiation and speciation can be done further with these annotated sequences.

To facilitate the identification of drug targets for Diamondback moths from the candidate gene index, sequences homologous to humans, cabbage, cauliflower and its natural predators including ants and the yellow wagtail were removed. By using data mining approach and functional annotation of EST sequences, aldehyde dehydrogenase was identified as a possible target to overcome the insecticide resistance in this moth. Finally, data mining methods can be combined with traditional drug target identification protocols to identify target specific compounds as bio-rational insecticides. Further, this methodology can be extended to other agriculturally important pests to identify the organism specific drug targets.

REFERENCES


