Hypothesis

Function annotation of peptides generated from the non-coding regions of *D. melanogaster* genome

Varughese Deepthi1*, Vineetha V. I. Nair3, Vipin Thomas1, Navya Raj1, Shidhi P. Ramakrishnan1, Juveria Khan2, Monika Kaushik2, Pawan K. Dhar1, 2, Achuthsankar S. Nair1

1Department of Computational Biology and Bioinformatics, University of Kerala, Kariyavattom, Trivandrum; 2School of Biotechnology, Jawaharlal Nehru University, New Delhi 110067, India; 3Indian Institute of Information Technology and Management, Kerala, Technopark, Trivandrum; Deepthi Varughese – E-mail: deepthidcb@keralauniversity.ac.in; Tel: 0471-3216730; *Corresponding author

Received April 20, 2016; Revised May 25, 2016; Accepted May 25, 2016; Published June 15, 2016

Abstract

*De novo* emergence of genes is the most fundamental form of genetic diversity that is attracting the attention of the scientific community. Identification of short open reading frames (sORFs) from the non-coding regions of different genomes has been leading this thought recently. The coding potential of these newly identified sORFs have been investigated through experimental and computational approaches in recent studies. In the present work we have tried to make peptides from intergenic sequences of *D. melanogaster* genome leading to therapeutic applications. Towards this goal of making novel peptides from non-coding genome, we have found strong computational evidence of 145 peptides with conformational stability from the intergenic sequences of *D. melanogaster*. The structure of these completely unique peptides was predicted using *ab initio* method. The function annotation of these peptides was carried out using this structural information. The newly generated proteins were categorised as DNA/Protein/ion binding proteins, electron transporters and a very few as enzymes too. Experimental studies can certainly provide validations to these preliminary findings. This work provides further evidence of untapped potential of non-coding genome.

Keywords: Non-coding; junk DNA; *de novo* peptides; short ORFs; antimicrobial peptides

Background:

The term ‘junk DNA’ for non-genic DNA was coined by Ohno *et al.* in 1970s and has caught the attention of scientists for many years since now due to its increased perception as regulators of gene expression. However, it is still not clear on whether the non-coding DNA merely exists for giving structural support to the folding DNA or are they expressed on demand and used as a dormant expression reservoir. Deep transcriptome sequencing has shown up transcripts that lack long or conserved open reading frames (ORFs) and there is growing proof that these transcripts translate to make novel peptides. The coding potential of short peptides has been explored widely using different computational and experimental approaches [1-3]. Recently, evidence of short novel genes or *de novo* genes which function in male reproduction has been reported in *Drosophila* non-coding genome [4, 5].

*D. melanogaster* is one of the extensively used model organisms for the investigation of many developmental and cellular processes in higher eukaryotes. Since last few years, scientists have been exploring the possibility of novel genes from the originally known non-coding regions of the organisms. Studies indicate that almost 12% of the newly emerged genes in *D. melanogaster* subgroup may have arisen from the non-coding DNA [7]. A common feature of *de novo* genes is that they are translated from short ORFs (<100 codons length ORF) originated from introns and are not observed in the coding region. The intergenic region of *D. melanogaster* could be considered as a
potential repository for *de novo* genes and many of them might be functional too. The identification and functional annotation of these genes is still in its infant stage. One way of finding the potential function of the non-coding region is to study if the DNA shows some characteristics of other sequences known to be functional. But if the novel genes are 'orphans' which lacks homologs, a real challenge in characterizing such protein coding genes exists. The present study was carried out to predict the potential peptides from the intergenic sequence of *D. melanogaster* and to understand its functional significance. The study is an extension of an earlier work where novel and functional proteins were non-natural proteins have been successfully synthesized and characterized from the non-coding regions of *Escherichia coli* K-12 (strain MG1655) genome [6]. In this study we have proposed an *in silico* approach in identifying potential peptides from the non-coding genome.

![Diagram showing in silico strategy for identifying potential peptides from non-coding DNA](image)

*Figure 1: In silico strategy designed for identifying potential peptides from non-coding DNA*
Methodology:

Refining the study sample
The non-coding DNA of *D. melanogaster* was used in this study. *D. melanogaster* is a well characterised model organism for eukaryotes and the genome size of *Drosophila* is 168.7Mb out of which 80% is non-coding. The preliminary dataset used for this study consists of 3500 intergenic sequences from the Flybase database version FB2012/04 [8] and were matched against the non-redundant (NR) protein database to verify their uniqueness. Those sequences which possess homologs in protein databases were then omitted. The whole idea was to extract the completely novel sequences and this formed the sample data for this study.

Novel Peptide Prediction
The intergenic sequences were translated into six reading frames and the peptides generated made the novel peptide dataset which do not resemble existing sequences from NR protein databases. The physiochemical characteristics of the peptides such as molecular weight, isoelectric point, instability index were then computed using the Expasy Protparam tool [9]. The structural and functional importance of these newly generated peptides were investigated in various steps as discussed in the workflow chart (Figure 1).

Function prediction based on the peptide sequence
The molecular function of the peptides was predicted using the standard ProtFun 2.2 Server [10] which made *ab initio* predictions of protein function from sequence. The program counted on the sequence derived protein features such as predicted post translational modifications (PTMs), protein sorting signals and physical/chemical properties calculated from the amino acid composition. These features used are integrated into final predictions of the cellular role, enzyme class (if any), and selected Gene Ontology categories of the submitted sequence. Search for the presence of sequence patterns helped in identifying the biologically meaningful sites in the sequence and thus in determining the function of a protein. Protein or proteins of a particular family shares common attributes which could be derived from a common ancestor. Prosite database [11] was used to search for the functionally significant regions such as profiles, domains and motifs from these novel sequences.

Antimicrobial Peptide Prediction
From the EKA dataset, peptides with size <40 AA were further considered to investigate on its therapeutic property. This was carried out with a sequence matching of each of the novel peptide against all natural antimicrobial peptides documented in the Antimicrobial Peptide Database (APD) [12]. APD consists of all natural antimicrobial peptides from bacteria, plants and animals with antiviral, antifungal, anticancer activities. The tool was used to calculate properties of the input peptide such as net charge, peptide length, hydrophobic residue (in %) and other amino acid compositions. These details were then used to traverse the database and to list those peptides which are most similar to the input given.

Peptide Structure Prediction and Evaluation
The function annotation of these novel potential peptides would be more reliable if the 3D structure information are available. The structure prediction of these peptides was done using *ab initio* method as they lack sequence homologs. I-Tasser webserver [13] was used to predict the protein structures with the combined approach of *ab initio* and fold recognition or threading alignment. The accuracy of the predicted structures was estimated based on the C-score (confidence score) of I-Tasser and is the score is typically in the range [−5, 2], wherein a higher score reflects a model of better quality.

Antimicrobial Peptide Database (APD) consists of all natural antimicrobial peptides from bacteria, plants and animals with antiviral, antifungal, anticancer activities. The tool was used to calculate properties of the input peptide such as net charge, peptide length, hydrophobic residue (in %) and other amino acid compositions. These details were then used to traverse the database and to list those peptides which are most similar to the input given.

Peptide Structure Prediction and Evaluation
The function annotation of these novel potential peptides would be more reliable if the 3D structure information are available. The structure prediction of these peptides was done using *ab initio* method as they lack sequence homologs. I-Tasser webserver [13] was used to predict the protein structures with the combined approach of *ab initio* and fold recognition or threading alignment. The accuracy of the predicted structures was estimated based on the C-score (confidence score) of I-Tasser and is the score is typically in the range [−5, 2], wherein a higher score reflects a model of better quality.

Stability of the predicted protein/peptide was evaluated using a consensus approach. Ramanathan et al. (2010) [14] has developed a strategy for evaluating the peptide structures based on the intra-molecular interactions and the same was adopted for the study. The stability centres (SCs) of the protein/peptide were predicted using the SCide program [15]. The stability centres are residues that are involved in cooperative long-range contacts, which are important in maintaining the stability of the protein. The total energy of a protein was calculated based on the bonds, angles, torsions, non-bonded and electrostatic constraints. The GROMOS force field implemented in Deep View [16] was used to compute the energy of the peptide and to confirm that the modelled protein structures are energetically stable. Chemically specific interactions (hydrogen bonds, ionic interactions) determine the globular structure of a protein. Hydrophobic interactions, ionic interactions and Disulphide bridges were computed using the webserver PIC, Protein Interaction Calculator [17]. Hydrogen bonds and Salt bridges in the peptides

**Figure 2: Predicted protein patterns from the novel peptides / proteins**

Stability of the predicted protein/peptide was evaluated using a consensus approach. Ramanathan et al. (2010) [14] has developed a strategy for evaluating the peptide structures based on the intra-molecular interactions and the same was adopted for the study. The stability centres (SCs) of the protein/peptide were predicted using the SCide program [15]. The stability centres are residues that are involved in cooperative long-range contacts, which are important in maintaining the stability of the protein. The total energy of a protein was calculated based on the bonds, angles, torsions, non-bonded and electrostatic constraints. The GROMOS force field implemented in Deep View [16] was used to compute the energy of the peptide and to confirm that the modelled protein structures are energetically stable. Chemically specific interactions (hydrogen bonds, ionic interactions) determine the globular structure of a protein. Hydrophobic interactions, ionic interactions and Disulphide bridges were computed using the webserver PIC, Protein Interaction Calculator [17]. Hydrogen bonds and Salt bridges in the peptides
were computed using the WHAT-IF server [18]. Apart from these interactions, the non-canonical interactions (C–H···π, C–H···O and N–H···π interactions) were also computed using the HBAT program [19].

Table 1: Sequence profiles predicted from EKA sequence dataset (145 peptides)

| Protein sequence Profiles                              | Frequency |
|-------------------------------------------------------|-----------|
| Microbodies C-terminal targeting signal               | 7         |
| Histidine-rich region profile                         | 1         |
| Cysteine-rich region profile                          | 2         |
| Phenylalanine rich region profile                     | 1         |
| Leucine zipper pattern                                | 8         |
| Bipartite nuclear localization signal profile         | 1         |
| Prenyl group binding site (CAAX box)                  | 4         |
| Cell attachment sequence                              | 3         |

Function Prediction based on the Peptide Structure

Considering the fact that similar fold may have similar function too, the functional significance of these novel peptides was inferred by a homology-based function prediction approach. COFACTOR webserver [20] was used to match the modelled protein/peptide structures with all the template proteins from the PDB libraries with known gene ontology (GO) vocabulary, and ligand-binding sites. Structural homologous for these newly generated proteins was searched from PDB using Dali server [21] based on the RMSD and Zscore values.

Table 2: EKA sequences showing sequence similarity with antimicrobial peptides

| Peptide ID | AMPs       | Activity                      |
|------------|------------|-------------------------------|
| EKA-26     | WAM2       | Gram+ & Gram-                 |
| EKA-31     | Latarcin 1 | Gram+ & Gram-, Fungi, Mammalian cells |
| EKA-35     | BMAP-27    | Gram+ & Gram-, Virus, Fungi, Parasites |
| EKA-56     | RANATU ERIN 2 | Gram+ & Gram-    |
| EKA-66     | Hyfl G     | Unknown                       |
| EKA-80     | Cliotide T1 | Mammalian cells, Cancer cells |

WAM2 = wallaby antimicrobial 2

Figure 3: EKA peptide structures predicted using I-TASSER. (A) EKA-27, (B) EKA-33, (C) EKA-36 with the structural homolog GCN4 superimposed. Red shade is given for GCN4 monomer and blue shade for EKA_36 peptide.
Among our 53 selected structurally stable novel peptides, 12 peptides showed a very significant sequence similarity with the antimicrobial peptides from APD database. These peptides were having a length in the range of 33-40 AA which is comparable to that of the known antimicrobial peptides. Few among these peptides (EKA-26, EKA-31, EKA-36 and EKA-80) showed a significant sequence similarity (>35%) to the antimicrobial peptides deposited in APD database with activity against Gram +, Gram -, fungi and some viruses (Table 2).

### Peptide Structure prediction and stability evaluation

Tertiary structure of the peptide was predicted using the I-Tasser Server and the C-score for the model structures are given in Table 3. The C-score range for the peptides was reported from -4 to 2.5. Top 53 best hits out of the 145 peptides were selected based on the C-score and structure stability. Table 3 gives report on the selected peptides (10 peptides) with the I-Tasser C-score, total energy, instability index, stability centers and the number of different intra-molecular interactions such as hydrogen bonds and hydrophobic interactions. Figure 3(A, B) shows a few of the predicted EKA peptide structures.

### Structure based Function Prediction

These homology based approach was used to capture the information regarding the possible functions that these proteins may acquire if they were expressed inside the cell. The gene ontology function annotation and the cellular localization prediction of these stable structures were done using Cofactor webserver [20]. Cofactor analysis reported the molecular and biological features of the selected proteins with reference to their structural homologs. These novel proteins were seemed to possess binding affinity to molecules, which were predicted with the Gene Ontology term prediction using the I-Tasser webserver (Table 4). Protein with DNA binding, lipid binding, protein binding, ADP binding, calcium ion binding, electron transporting properties etc. were remarkably noted. Some of these proteins

### Table 3: Predicted structural properties of the novel EKA peptides (best results shown)

| Peptide ID | I-Tasser (C-score) | Total Energy | No. of Instability centres | Non covalent interaction | Non canonical interaction |
|------------|-------------------|--------------|----------------------------|--------------------------|--------------------------|
| EKA-1      | 2.646             | -1780.711    | 6                          | 37.9                     | 84                       | 12                       |
| EKA-7      | -2.435            | -1219.2      | 5                          | 11.63                    | 78                       | 115                      |
| EKA-55     | -1.498            | -955.090     | 6                          | 38.62                    | 100                      | 13                       |
| EKA-95     | -2.048            | -1494.830    | 6                          | 29.48                    | 84                       | 16                       |
| EKA-109    | -2.422            | -1731.463    | 7                          | 25.81                    | 44                       | 20                       |
| EKA-143    | -2.18             | -1953.243    | 8                          | 29.91                    | 35                       | 10                       |
| EKA-149    | 1.55              | -5624.295    | 15                         | 29.53                    | 264                      | 378                      |

### Result and Discussion

The initial dataset consists of 3500 intergenic sequences from the 3rd chromosome of D. melanogaster. These sequences were then subjected to sequence similarity checking against NR protein database. The sequences with no homologs were considered as unique intergenic sequences were then subjected to six frame translation. We were able to generate 145 novel peptides which we named as “EKA peptides” (after it was named by Dhar et al. 2009) [6] and were deposited in the in-house database - EKA knowledgebase. The sequences were investigated to predict the possible functions of the novel peptides. Profunc server reported and categorised the biological function of the novel proteins into energy metabolism, transcription-translation regulation, cell-envelop, transport and binding etc. Gene Ontology prediction categorised the novel peptides into different classes such as transporter, growth factor, receptor, immune response and transcription regulation.

Important functional sites such as motifs, domains or patterns or profiles of the novel proteins were then identified using the Prosite webserver. The pie chart given below (Figure 2) describes the profiles and domains predicted from the novel proteins. Many phosphorylation sites, N-glycosylation sites, N-myristoylation and amidation sites were identified. Apart from that many important sequence profiles like Histidine rich, Cysteine rich regions were identified (Table 1). Leucine zipper patterns which are unique for DNA binding protein were predicted in 8 novel proteins which could be considered as significant findings.

Among our 53 selected structurally stable novel peptides, 12 peptides showed a very significant sequence similarity with the...
were showing structural similarity to DNA binding proteins and they were predicted to be located in the nucleus. For instance, our protein (EKA-36) of length 33 AA showed similarity with a general control DNA binding protein (GCN4) (PDB: 4NJ2) with an RMSD of 1.2 Å. The structural homolog search which was carried out with the Dali server came up with the interesting result showing that 5 of our novel proteins are showing considerable similarity with the DNA binding, GCN4 protein at the 3D structural level (Table 5). The structurally superimposed image of the EKA-36 and GCN4 (PDB: 4nj2) is shown in the Figure 3(C).

Table 4: Function Annotation and Localization predictions using Gene Ontology Terms

| Peptide ID | Peptide length | GO-molecular function predicted | GO Location predicted | Structural homologs with known function | RMSD |
|------------|----------------|---------------------------------|-----------------------|----------------------------------------|------|
| EKA-21     | 43             | lipid binding                   | integral to membrane  | GTP binding Protein                    | 1.1  |
| EKA-27     | 33             | voltage gated pottasium channel activity | voltage gated pottasium channel complex | Glutathione - regulated Pottasium - Efflux system | 1.7  |
| EKA-36     | 33             | DNA binding                     | nucleus               | GCN4 (general control protein)         | 0.55 |
| EKA-91     | 35             | adenyl ribonucleotide binding   | plasma membrane       | RecX, DNA binding protein              | 1.5  |
| EKA-97     | 46             | DNA binding                     | nucleus               | Ligand Binding                         | 2.2  |
| EKA-115    | 40             | Ion binding                     | Proteosome accessory complex | Proteosome activator                    | 1.1  |
| EKA-124    | 43             | Ion binding                     | cell                  | RAB1 domain                            | 1.4  |

Table 5: EKA peptide sequences and their structural homologs

| Peptide ID | Structural homolog | ZScore | RMSD | Total Residue | Aligned Residues | Similar molecule                          |
|------------|--------------------|--------|------|---------------|------------------|-------------------------------------------|
| EKA-3      | 2HY6-D             | 3.9    | 1.7  | 32            | 32               | General control protein gcn4              |
| EKA-21     | 3AHA-F             | 4.1    | 1    | 34            | 34               | Transmembrane protein gp41                |
| EKA-36     | 4NJ2-A             | 4.6    | 1.2  | 33            | 33               | General control protein gcn4              |
| EKA-50     | 3W8V-A             | 3.1    | 2.8  | 32            | 30               | Gcn4n coiled coil peptide                 |
| EKA-63     | 4CBJ-G             | 4.8    | 2.1  | 69            | 49               | ATP synthase subunit c                    |
| EKA-81     | 3HTU-H             | 3.8    | 1.8  | 36            | 36               | Vacuolar protein-sorting-associated protein |
| EKA-95     | 2KHH-A             | 4.6    | 2.6  | 57            | 57               | mRNA export factor mex67                  |
| EKA-115    | 1EC5-A             | 5.5    | 1.8  | 48            | 40               | Protein (four-helix bundle model)         |
| EKA-117    | 2R2V-H             | 3.6    | 2.2  | 31            | 30               | Gcn4 leucine zipper                       |
| EKA-145    | 4OJK-C             | 5      | 1    | 37            | 37               | Ras-related protein rab-11b              |

Conclusion:
The evidence on the huge amount of functionally significant region in the un-expressed ‘junk’ DNA is increasing day by day. Encode project has already claimed that they have assigned biochemical function to 80% of human genome [22]. This indicates that the conventional annotation process might have missed many important functional transcripts which are RNA coding or protein coding. Also, recent reports have come up on the discovery of short peptides or short ORFs [23] from the non-genic region which supports these findings. Peptide mapping
from non-coding DNA region is found to be promising since successful trials have been reported previously by Dhar et al. 2009 [6]. Inspired from all these findings, the present study was planned to develop an in silico approach in predicting potential proteins from the intergenic region of D. melanogaster genome, which may perform some function inside the cell. From the analysis, we were able to create potential novel peptides from intergenic sequences of D. melanogaster genome. The predicted 145 peptides with conformational stability point towards the potentiality of these peptides to be expressed in cell in different conditions. The sequence and the structure of the proteins were used in order to assimilate information regarding the potentiality of the proteins. The in silico functional characterization of these novel peptides revealed the presence of important profiles and patterns such as Histidine-rich region profile, Cysteine-rich region profile, Leucine zipper pattern etc. The prediction of DNA binding capacity of some of these peptides can be further investigated so as to check the role of transcription regulation of these novel proteins. The reliability of the predicted models and its expression capability has to be validated using experimental study. Another set of peptides were found to have similarity with the already known natural anti-microbial peptides. This opens a way to produce novel AMPs of our interest from the intergenic region of D. melanogaster genome. Since the synthesis of peptides from the intergenic region has already been proven [6], there is a good scope for the proposed production of novel AMPs from non-coding DNA. The preliminary findings from the present study open the way for an extensive analysis in re-annotating the non-coding space of D. melanogaster genome.

Conflict of interest:
The authors declare that they have no conflict of interest.

References:
[1] Galindo MI et al. PLoS Biol. 2007 5:106 [PMID: 17439302]
[2] Kondo T et al. Science 2010 329:336 [PMID: 20647469]
[3] Kageyama Y et al. Biochimie. 2011 93:1981 [PMID: 21729735]
[4] Begun DJ et al. Genetics 2007 176: 1131 [PMID: 17435230]
[5] Levine et al. Proc Natl Acad Sci. 2006 26: 9935 [PMID: 16777968]
[6] Dhar PK et al. J Biol Eng 2009 3:2 [PMID: 19187561]
[7] Zhou Q et al. Genome Research 2008 18: 1446 [PMID: 18550802]
[8] Tweedie S et al. Nucleic Acids Research 2009 37:D555 [PMID: 18948289]
[9] Gasteiger E et al. Methods Mol Biol. 1999 112:531 [PMID: 10027275]
[10] Jensen L et al. Bioinformatics 2003 5:635 [PMID: 12651722]
[11] Sigrist CJ et al. Nucleic Acids Res 2013 41:D344 [PMID: 23161676]
[12] Wang G et al. Nucleic Acids Research 2009 37: D933 [PMID: 18957441]
[13] Zhang et al. BMC Biology 2008 9:40 [PMID: 17488521]
[14] Ramanathan K et al. Interdiscip Sci. 2011 3:182 [PMID: 21956740]
[15] Dosztanyi ZS et al. Bioinformatics 2003 19: 899–900 [PMID: 12724305]
[16] Guex N & Petisch MC Electrophoresis 1997 18: 2714 [PMID: 9504803 ]
[17] Tina KG et al. Nucleic Acids Research 2007 35:W473 [PMID: 17584791]
[18] Vriend G et al. J Mol Graph 1990 8:52 [PMID: 2286828]
[19] Tiwari A et al. In Silico Biol, 2007 7:0057 [PMID: 18467777]
[20] Ambrish Roy et al. Nucleic Acids Research. 2012 40:W471 [PMID:22502402]
[21] Holm L & Rosenström P. Nucl. Acids Res. 2010 38:W545 [PMID: 20457744]
[22] ENCODE Project Consortium Nature 2012 489:57 [PMID: 22955616]
[23] Sopko, R., & Andrews, B. Genome Research 2006 16:314 [PMID: 16510897]