# CDS Annotation in cDNA Sequence of Caenorhabditis_elegans: Relevance to Atherosclerosis

Pramod W Ramteke, Amit Dubey*, Satendra Singh

**Abstract**— To interpret the effects of genetic variation and phenotypic changes between individuals is still very complex in human. We are developing various strategies for Predicting and understanding the downstream effects of genetic variation using computational methods are becoming increasingly important for the identification of novel genes codons and their target and impact over protein-coding genes which is also known as coding sequences (CDS) and their prediction is an beneficial step over functional annotation of genes. CDS prediction for human genes from genomic sequence is convoluted by the vast affluence of intergenic sequence in the genome, and delivers little information about how contrasting parts of potential CDS regions are expressed. In contrast, *Caenorhabditis_elegans* gene CDS prediction from cDNA sequence offers obvious advantages, yet confrontation a different set of intricacy when observed on high-throughput cDNA sequences, are an increasingly relevant tool for genetic and biomedical research. Bioinformaticians have become superior at functional assumption methods derived from functional and structural genomics. Examination of CDS that change a rarely used codon in to a frequently used one or vice versa may help in predicting their phenotypic effect on the individual carrying the change. The code is not either overlapping or punctuated, but has mRNA sequences read in successive triplet codons till to reaching a stop codon. The perfect genetic code uses triplet bases for every amino acid. The efficiency of the genetic code can be significantly increased if the requirement for a fixed codon length is discarded so that the more common amino acids have briefer codon lengths and rare amino acids have expanded codon lengths. It would be more challenging for the system of transcription to deal with a variable codon length. This study also provide significant insights over new approaches to reveal and understand chymase inhibitors through systems-based approach to analyzing conversion of Angiotensin-I converting enzyme (ACE) converting angiotensin I to angiotensin II or for inactivating bradykinin and miRNA which provides application to evaluate diverse expansion of atherosclerosis vulnerabilities which is the basis of new medical informatics approaches in terms of their further drug development.



**Index Terms**— cDNA, Annotation, CDS Prediction, Atherosclerosis, Compuatational Analysis, Angiotensin Converting Enzyme(ACE), Start Codons, Stop Codons,Phenotypic changes

## 1 INTRODUCTION

Caenorhabditis_elegans genome contains an ACE-like and several neprilysin-like sequences, but similar sequences have not been detected in unicellular eukaryotes, suggesting that an ACE/neprilysin ancestor arose during the Cambrian radiation about 530 million years ago. After divergence from neprilysin, ACE acquired a unique chloride activation mechanism that has been identified in invertebrates through to mammals [1]. Early studies in Caenorhabditis_elegans determined that miR-mediated regulation was post-transcriptional, because there were large effects on protein expression and no discernible effects on mRNA abundance [2]. In other systems, modest effects on the amounts of the mRNA target were seen in addition to substantial degrees of regulation at the protein level [3]. Furthermore, miRs have been associated with inflammation, oxidative stress, impaired adipogenesis and insulin signaling, and apoptosis and angiogenesis in relation to obesity. All of these processes contribute to the development of type 2 diabetes, atherosclerosis, and associated cardiovascular disorders [4-6]

Bioinformatics analysis of the complete genome sequence of C. elegans by the WormBase consortium initially revealed over 19000 coding genes [7]. When the genome of the closely related species C. briggsae was sequenced and a comparative analysis was performed between the two species, 6% more coding genes were predicted (20261) [8]. Since the bioinformatics annotation pipeline from the WormBase consortium is constantly evolving new protein-coding genes are being predicted and this number is increasing. The latest version of the C. elegans genome sequence (WS228) predicts 24610 coding genes. [9] Considering that twice the number of new genes has been predicted using gene prediction algorithms, novel approaches that explore different search spaces may reveal even more protein-coding genes. Indeed, evidence suggests that

*Pramod w. Rameke, Department of Biological Sciences, Allahabad Agriculture Institute, PH-09415124985.*
*E-mail: pwramteke@yahoo.com*
*Amit Dubey*, Department of Computational Biology and Bioinforrmatics, Allahabad Agriculture Institute, PH- 09198278069.*
*E-mail: ameetbioinfo@gmail.com*
*Satendra Singh, Department of Computational Biology and Bioinforrmatics, Allahabad Agriculture Institute, PH-09565448367,*
*E-mail: satendralike@gmail.com*
*Corresponding Author (*)*

more protein may exist in C. elegans in the case of old protein fold families that evolved a long time ago from divergent (or convergent) evolution [10]. Such protein family members are renowned to be difficult to identify by conventional sequence alignment software since they share very little sequence identity. The OB-fold is one example [11].

## 2 MATERIAL AND METHODS

### 2.2 Data Analysis

Following are the steps for data analysis.

**Step1 Step1: Data download from site: Caenorhabditis_elegans.WBcel215.ncrna.fa**

The dataset, consisting of the cDNA protein coding Sequence samples with information of known chromosome, gene biotype and transcript biotype at 47 development stages. The data was obtained from public one from the Ensembl genome browser 72-*Caenorhabditis_elegans* database.

**Step2: Sequence retrieval:**

The sequences of genes containing the novel genes codons were retrieved from the publicaly available ENSEMBL genome browser with the help of the free available Biomart tool using the list of ENSEMBL gene Id and Parameters selected for this: Chromosome name, Ensembl gene id, Coding sequence

**Step3. Than, finding all the codons for each sequence:**

**Algorithmic Logic**

In the nucleotide sequence of (ATGC) we have to search ATG (start codon) and stop codon (TAG, TGA, and TAA)
So, we have to select the small sequences which start with start codon and end with stop codon.

1. Our original sequence is AAAAATGGTGGTGCGCCGAACCTTGGGTTATTAGTAG-CAGCAGCAGATGACAGATGAATGATGACGATGACAG-TAGCAA

2. Then our small subsequence will be ATGGTGGTGCGCCGAACCTTGGGTTATTAG and ATGACAGATGAATGA (length of small subsequence should be divisible by 3)

3. Now we have to search all pairs of codon in these two small subsequence and right according to that

['AAA', 'AAG', 'AAC', 'AAT', 'AGA', 'AGG', 'AGC', 'AGT', 'ACA', 'ACG', 'ACC', 'ACT', 'ATA', 'ATG', 'ATC', 'ATT', 'GAA', 'GAG', 'GAC', 'GAT', 'GGA', 'GGG', 'GGC', 'GGT', 'GCA', 'GCG','GCC','GCT',

1 2 1 3 4 5 3 and so on

'GTA', 'GTG', 'GTC', 'GTT', 'CAA', 'CAG', 'CAC', 'CAT', 'CGA', 'CGG', 'CGC', 'CGT', 'CCA', 'CCG', 'CCC', 'CCT', 'CTA', 'CTG', 'CTC', 'CTT', 'TAA', 'TAG', 'TAC', 'TAT', 'TGA', 'TGG', 'TGC', 'TGT',

1 2 1 3 4 5 3 and so on

'TCA', 'TCG', 'TCC', 'TCT', 'TTA', 'TTG', 'TTC', 'TTT']

1 2 1 3 4 5 3 and so on

above one is for Gene1

Now for Gene2

1 2 1 3 4 5 3 and so on...

1 2 1 3 4 5 3 and so on...

1 2 1 3 4 5 3 and so on....

whatever count will come instantly we have to show down there.....

Prediction of CDS which may have phenotypic effects involves identifying those with one or more of the following properties –

I. Changes a rare codon to a frequent one or frequent codon to a rare one.
II. Involves creation of CpG dinucleotide (most likely due to a change to C in the third position of the Affected codon (GC3))
III. Changes a rare codon coding for amino acids occurring in turns, loops or domains linkers to a frequent one.

## 3 RESULTS

A separate list of the data containing the CDS that change a rare codon to a frequently used one, or vice versa, was created. Other fields listed in the data include associated gene, Ensembl gene id, Ensembl transcript id as well as coding sequence start position which will help in finding the phenotypic changes caused due to CDS and we have also predicted perfect codons by searching start codon (ATG) and next stop (TAA, TAG, TGA) codon but in some case we found start codon. This result is persistent with the fact of genetic variation and phenotypic changes at mRNA level which is having association towards conversion of Angiotensin-I converting enzyme (ACE) converting angiotensin I to angiotensin II at the differentiating component causing of atherosclerosis disease.

## CONCLUSION

Interchangeable variation can have effects of potential pathophysiological and pharmaco- genetic importance. These are the possible points which can be responsible for the phenotypic changes occurring due to CDS.
1. The presence of a rare codon, affects the timing of co-translational folding and insertion of P-gp into the membrane, thereby altering the structure of substrate and inhibitor interaction sites. 2. Some synonymous changes in humans have been shown contribution to the development of type 2 diabetes, atherosclerosis, and associated cardiovascular disorders by exon skipping. 3. GC content of the third position of codons (GC3) has been classically used to predict whether natural selection leverage codon usage the evacuation between the GC3 of a gene and the GC frequency predicted in neighboring non-coding regions which can be used as a measure of gene specific forces (e.g. selection) on gene mutations farther isochoric mutational tendencies. Those genes showing the strongest evacuation from isochoric GC content are the best candidates for the estimation of functional changes resulting from gene mutations. 4. Codons usage can influence the conformational state of the protein and shows that codon precise translation rate may affect the in-vivo protein folding. 5. Presence of rare codon in turns, loops and domains linkers has been predicted. Further studies are required to find the phenotypic changes causes by CDS.

In summary, one of the most exciting new areas of molecular biology is the computational profiling and functional analysis of miRs, which play a central role in how the genome is regulated and how traits are proceed on or wipe-out by environmental and genetic factors. In the last 2 years, scientists have begun to perceive this natural "gene silencing" innards, which is helping them map genomic pathways and review ways of communication between different cell types within the same tissue and between the same and different cell types in remote tissues. However, further improvement in functional analysis and encounter of synergetic actions is needed to uncover connections to diseases and new classes of therapies

## ACKNOWLEDGMENT

We wish to thank my coleuges, friends who gave me their valuable support and guidelines. This work was supported in part by a self grant from me and my guide.

## REFERENCES

[1] Ahsan Husain, Ming Li and Robert M. Graham,"Forming Pathway? Non-ACE, Non-Chymase Angiotensin II Selective Inhibitors Provide Evidence for a-Do Studies with ACE N- and C-Domain," *Circ Res*.; 93, 91-93, 2003.

[2] Olsen, P. H., and Ambros, V,"The lin-4 regulatory RNA controls developmental timing in Caenorhabditis elegans by blocking LIN-14 protein synthesis after the initiation of trans- lation," *Dev. Biol*; 216, 671– 680, 1999.

[3] Brennecke, J., Hipfner, D. R., Stark, A., Russell, R. B., and Cohen, S. M,"bantam encodes a developmentally regulated microRNA that controls cell proliferation and regulates the proapoptotic gene hid in Drosophila," *Cell* 113; 25–36, 2003.

[4] Heneghan, H. M., Miller, N., and Kerin, M. J,"Role of microRNAs in obesity and the metabolic syndrome,"*Obes. Rev*.11; 354 –361, 2010.

[5] Ling, C., and Groop, L,"Epigenetics: a molecular link between environmental factors and type 2 diabetes,"*Diabetes* 58; 2718 –2725, 2009.

[6] Bonauer, A., Boon, R. A., and Dimmeler, S,"vascular microRNAs,"*Curr. Drug Targets* 11; 943–949, 2010.

[7] C. elegans Sequencing Consortium Genome sequence of the nematode C. elegans: A platform for investigating biology. Science 282(5396): 2012–2018, 1998.

[8] Stein LD, Bao Z , Blasiar D, Blumenthal T, Brent MR, et al," The genome sequence of caenorhabditis briggsae: A platform for comparative genomics," *PLoS Biol* 1(2); E45. 10.1371/journal .pbio.000004 5, 2003.

[9]Magrane M,"Consortium U UniProt knowledgebase: A hub of integrated protein data," *Database (Oxford)* 2011; bar009. 10.10 93/database/ bar009, 2011.

[10] Murzin AG,"How far divergent evolution goes in proteins,"*Curr Opin Struct Biol* 8(3); 380–387, 1998.

[11] Murzin AG,"OB (oligonucleo tide/oligosacchari de binding)-fold: Common structural and functional solution for non-homologous sequences,"*EMBOJ* 12(3); 861–867, 1993.