A Study on Dicodon-oriented Gene Finding using Self-Identification Learning

Chungfan Kim

School of Knowledge Science,
Japan Advanced Institute of Science and Technology

February 15, 2000

Keywords: gene finding, HMM, self-identification, dicodon.

Abstract

1 Objective

Dicodon usage measure has been well known as one of those which can discriminate protein coding and non-coding regions out in genomic sequence data. However, there has been no objective and quantitative analysis on the dicodon usage measure thus its biological semantics remain undeclared until now. In order to push the envelope of gene finding out and to develop more efficient coding measure, we need to uncover how the dicodon usage measure can distinguish protein coding/non-coding regions. This paper aims to give an objective criteria for dicodon usage measure.

2 Background

The global on-going research activities which aim to reveal genomic sequence of many species including *homo-sapiens* (human) are propelled by intensive effort of many researchers. Recent technological revolution of experiment equipments and computer performance helped largely to push the research activities further. Databases for genomic sequence data are growing faster and bigger on daily basis.

Such enormous genomic sequence data can be exploited for functional analysis of genes and proteins. The analysis requires to find protein coding regions out of plain genomic sequence data which look like merely a lengthy series of A, C, G, T letters.

It is usually very difficult to distinguish protein coding regions and non-coding regions because gene structure in higher eukaryotic genomic sequence such as homo-sapiens is

very complicated. Besides, such analysis simply requires computational methods because the genomic sequence data are too large to deal with human work load. Gene finding is a method to predict where those protein coding regions reside in a genomic sequence data with aid of computational methods.

Gene finding has been under intensive research effort nearly a couple of decades. There has been a lot of different approaches proposed. The approaches are roughly divided into two categories [1]. One is to find protein coding regions based on sequence similarities. Another is to find protein coding regions based on stochastic and probabilistic coding measures. The dicodon usage measure is included in the latter and it evaluates potential bias of dicodon (a pair of codons) usage between the coding and non-coding regions in order to find protein coding regions. Fickett and Tung indicated that the dicodon usage measure is one of the superior coding measures [2]. The dicodon usage measure can be reduced to several basic molecular biological elements such as codon usage and pair amino-acid. Therefore determination of weights for these elements can clarify which element is the most important and what makes the dicodon usage measure such efficient.

This paper focuses on a quantitative analysis of the dicodon usage measure and on clarification to know which reduced element play more important role to the dicodon.

3 Examination and Result

Our preliminary analysis for the dicodon usage measure using "dicodon-oriented Hidden Markov Model" with "Self-identification learning method" indicated that the dicodon usage measure is redundant as a gene finding measure.

The "dicodon-oriented Hidden Markov Model" gives a simple but powerful probabilistic description of gene structure with relatively small size of parameters. The "Self-identification learning method" offers novel learning scheme that does not require prepared training data.

This result induces a hypothesis; there is another measure that has smaller size of parameters and performs as good as the dicodon usage measure. Such measure would make the gene finding more self-identification learning friendly. Hence it will facilitate complete automation of gene finding that is demanded by next generation of genome sequencing projects.

In order to evaluate the redundancy and accuracy of dicodon usage measure, we took "divide and conquer" approach. Based on the compositions of the dicodon usage measure such as codon usage, pair amino-acid, and C+G content, we defined six different model to emulate the dicodon usage measure by the compositions with smaller parameter size than the dicodon usage measure. We evaluated the six models and the dicodon usage measure in aspect of sensitivity, specificity, and approximation error. Our evaluation result shows that the dicodon model outperforms the six emulators in terms of sensitivity as well as specificity. This result indicates that the dicodon model can not be represented by a combination of the pair amino-acid, the codon usage, and the G+C content.

4 Summary and Future Work

In this paper, our preliminary examination indicated that the dicodon usage measure is redundant for gene finding. Furthermore, the quantitative analysis using six emulators of dicodon usage measure with smaller parameter size gives a negative result that denies implicit belief that is "dicodon usage measure = codon usage + pair amino-acid + C+G content." According to our result of analysis, it should be "dicodon usage measure = codon usage + pair amino-acid + C+G content + α." The alpha is not re-veiled in this paper yet. However, results of our examination imply that C+G content does not give sufficient granularity for coding/non-coding discrimination because the C+G content does not distinguish C/G and A/T at third nucleotide position in a codon and it should be considered of reasonable bias among A, T, C, and G respectively. Therefore every nucleotide should be dealt as it is.

We are under investigation for finding the peculiar bias of A/T/C/G at the third nucleotide position. and performing gene clustering based on a A/T/C/G bias for several amino-acids. The clustering is effective because every coding region has different dicodon profiles thus gene finding with several sets of dicodon usage measure will perform better than that with single set of dicodon usage measure. The investigation will bring us a clear view of the α in the future.

References

[1] Burge, C.
Identification of Genes in Human Genomic DNA (Doctoral Thesis).
Stanford University March 1997

[2] Fickett J. W., Tung C. S.
Assessment of protein coding measures.
Neucleic Acid Research, 1992, Vol. 20, No. 24:6441-50

Publications

- Kim, C., Konagaya, A., Asai, K.
  A Gene Finding using a di-codon oriented Hidden Markov Model (in Japanese),
  Proc. of the 13th Annual Conference of JSAI, 1999., pp.330-331.

- Kim, C., Konagaya, A., Asai, K.
  A Generic Criterion for Gene Recognition in Genomic Sequences,
  Proc. of Genome Informatics Workshop 1999, pp.13-22.