ISSN: 2394-7659
IMPACT FACTOR- 2.785

International Journal of Engineering Researches and Management Studies

DEVELOPMENT OF COMPUTATIONAL METHOD FOR GENE PREDICTION IN EUKARYOTIC GENOMES

Kumar Sahaj[*1] and Mahendra Singh Charan[2]
[*1]Department of Biochemical Engineering and Biotechnology, IIT Delhi
[2]Department of Chemical Engineering, IIT Delhi

ABSTRACT

As sequence data continues to be generated at a logarithmic rate our dependence on effective gene prediction methods is also increasing. So, I review the current state of eukaryote gene prediction methods, their advantages and future prospective.

Keywords:- *Computational method, Gene Prediction, Eukaryotic genomes.*

I. CONTENT & SIGNALS SENSORS

Gene annotation strategies rely on sensors within the DNA sequence to allow accurate presentation of gene structure and organization.

Two types of sensor are routinely used to locate genes in the genomic sequence:

1) **Content Sensors**- classify DNA into **coding** regions and **non-coding** regions (introns, intergenic regions and un-translated regions (UTR's)).
Content sensors can be further divided into **extrinsic** and **intrinsic** sensors. Here it is assumed that coding sequences are **more conserved** than non-coding sequences.

   a) **Extrinsic sensors**- It exploit homology searches to identify highly conserved exons.

   Limitation -A significant failing of extrinsic approaches is that they are limited to homologies within the database; if no homologs exist no data can be extracted.

   b) **Intrinsic sensors**- It focuses on specific innate characteristics of the DNA sequence itself, which help to predict the likelihood of whether the sequence in question **"codes" for a protein or not**. The most obvious indicator of coding versus non-coding sequence identified to date is **hexamer frequency, nucleotide composition, codon** usage and **base occurrence periodicity.**

2) **Signal Sensors**- It detect the presence of **functional sites** specific to a gene. Signal relating to **transcription, translation** and **splicing** have all been employed to be used in gene identification and structure prediction.

   a) **TSS(Transcriptional Signal Sensors)**- It include **initiator** or **cap signal** located at transcriptional start site and the upstream of TATA box promoter signal and **polyadenylation** signal.

   b) **Translation signal**-"Kozak Signal" located immediately upstream of start codon.

   c) **Splice Site Signal**- Higher eukaryotic genes have multiple exons. So, prediction of genes heavily rely on these signals

ISSN: 2394-7659
IMPACT FACTOR- 2.785

IJERMS

International Journal of Engineering Researches and Management Studies

II. GENE TRDICTOR METHODS(A background)

Emperical Method

It works on **similarity, homology** based gene finding system thus, it uses extrinsic method . Target genome is searched for sequence that are similar to extrinsic evidence in form of **ESTs**(Expressed Sequence Tags),**mRNA** ,**Protein Tags** and **homologous** and **orthologous** sequences.

Given a protein sequence and mRNA a family of possible coding DNA sequences can be derived **by reverse translation and reverse transcription** of the genetic code. Once candidate DNA sequences have been determined it is easy to find target sequence in **database** using the method of Local alignment etc.

Limitation-

a) **Accuracy** of the **database** where we are looking for the target sequence matches.
b) Requires extensive sequencing of mRNA and protein products(expensive and cumbersome) .
c) only a subset of all genes in the organism's genome are expressed at any given time, meaning that extrinsic evidence for many genes is not readily accessible in any **single cell culture**. So, we need hundreds and thousands of **cell types** to collect these evidence.

Major Challenges Envolved for Future Prospective-

Sequencing errors in raw data, dependence on the quality of sequence assembled, handling short reads, frameshift mutation, overlapping etc.

important factor underused in current gene detection tools is existence of gene clusters **– operons**. Most gene detectors treat each gene in isolation, independent of others which is not true as it is found that some gene affects each other or they are interdependent.

As, these detection are done with help of Sensors. We can make the Sensors more sensitive or we can find more sensors to predict genes. New Transcriptome sequencing technologies such as RNA-Seq and ChIP-sequencing incorporate additional extrinsic evidence into gene prediction and validation.

III. AB INITIO METHOD

It is an intrinsic method based on gene content and signal detection.

Genomic DNA sequence alone is systematically searched for certain regions/sign of protein coding gene. These sign are broadly categorised as signals that indicate the presence of gene nearby, or content, statistical properties of protein coding sequence itself.

Signals like promoter and other regulatory signals in these genome are more complex(so we can increase its sensitivity). Two classical examples of signal identified by eukaryotic gene finders are CpG islands and binding sites for a poly(A) tail.

Secondly, splicing sites are themselves another signal used in eukaryotic gene finders.

Advantages-

a) Difficulty in obtaining **extrinsic evidence** for many genes is not in the case of Ab initio method.
b) This method more accurately characterized as gene prediction, since extrinsic evidence is generally required to establish that a given gene is functional.

Disadvantages/Limitation-

a) Difficulty to **detect periodicity** and other known **content** properties of protein coding DNA

What We Can Do-

We can look for other **signals** other than those **directly detectable** in sequence may improve gene prediction . Like role of secondary structure in the identification of regulator motifs can be used.

Neural networks are computational models that excel at machine learning and pattern recognition can be enhanced.

ISSN: 2394-7659
IMPACT FACTOR- 2.785

IJERMS

International Journal of Engineering Researches and Management Studies

IV. SOME TOOLS & COMPUTATIONAL METHODS USED FOR GENE PREDICTION

Neural Networks

Excels at machine learning and Pattern recognition. It must be trained with example data before being able to generalise for experimental data. It is able to come up with approximate solutions to problems that are hard to solve algorithmically provides sufficient **trained data.**

Limitations- need too many trained data which are difficult to obtained for large genomic search.

Hexamer-coding measures

It is observed that nucleotides are not independent of each other, but tend to occur together as if in a word — of length *k* (*k*-tuples); 6-tuples are called hexamers. It has been used as a powerful way of discriminating coding regions from non-coding regions, as some 'words' are more likely to be present in either type of DNA. A score ***s*** for a hexamer ***w***, such as CAGCAG, can be defined as $s(w) = \log(freq(w))$. Because the frequency of CAGCAG is relatively high in exons, its score in exons will be higher than that of, for example, TAATAA

Limitations- it is assumed that nucleotides occurs together not independent and the whole measure is dependent on its accuracy(is coding nucleotides occurs in some pattern?).

Markov Model

It assumes that the probability of a particular nucleotide occurring at a given position depends only on the k previous nucleotides. There are many categories of MM such as PWM,WAM,IMM and HMM.
These model are defined by the conditional probabilities P(X|k previous nucleotides), where X=A, T, G or C.

Limitations- For a sequence of length *n*, the dynamic programming for finding the best path through a model with *s* states and *e* edges takes memory proportional to *sn* and time proportional to *en*. So, it can't be used for large sequences(practically).

Coding Statistics(codon frequencies)

Assume $S = a_1b_1c_1, a_2b_2c_2, ..., a_{n+1}b_{n+1}c_{n+1}$ is a coding sequence with unknown reading frame. Let f_{abc} denote the appearance frequency of codon abc in a coding sequence. The probabilities p_1, p_2, p_3 of observing the sequence of n codons in the 1st, 2nd and 3rd frame respectively are:

p_1 = fa1b1c1 × fa2b2c2 × ... × fanbncn
p2 = fb1c1a2 × fb2c2a3 × ... × fbncnan+1
p3 = fc1a2b2 × fc2a3b3 × ... × fcnan+1bn+1
The probability Pi of the ith reading frame for being the coding region is (i = 1, 2, 3): Pi = pi / p1 + p2 + p3

Here we can calculate the probability for each reading frame and choose the best for sequencing.

New Method(make use of the above methods)

Here I am assuming CpG Island as a signal to find gene within a genome
Suppose we have a eukaryotic genome piece and we want to annotate it. Here we use the signals that CpG Island are generally found upstream of Promoter region. So, using the above or combination of above methods we will try to drill out the coding part of gene.

Here we are looking for **CG dimer** in the sequence. In starting we can use **Coding statistics** and **Markov Model** to measure probability of dimer with right reading frame.

ISSN: 2394-7659
IMPACT FACTOR- 2.785

IJERMS

International Journal of Engineering Researches and Management Studies

Let a sequence be ….AT**GC**CTG…..

Here we have 3 reading frames

1) …A **TG** CCTG… $P_1 = f_{TG} \times f_{CC} \times f_{TG}$.....

2) …AT **GC** CTG…. $P_2 = f_{AT} \times f_{GC} \times f_{TG}$.....

Here we are interested in $p_2 =$ $P_2 / P_1 + P_2$

Also we can find score s for GC frequency with the help of **Hexamer Coding Measure** as-

$$s(w) = \log(\text{freq}(w))$$

here freq(w) is the frequency of GC occurance in coding sequence

now we are able to get some part of gene(partial exon and introns). We can use **Neural Network Method** as it will be fast and easy (as it require less Trained set examples).

Finally with more refined genes we can use **extrinsic method** for **homology** in database for more accurate results.

REFERENCES

1. *http://www.ch.embnet.org/CoursEMBnet/Zurich04/slides/gene_ho.pdf*
2. *http://www.nature.com/nrg/journal/v3/n9/box/nrg890_BX1.html*
3. *https://en.wikipedia.org/wiki/Gene_prediction#Neural_networks*
4. *https://compbio.soe.ucsc.edu/html_format_papers/tr-94-24/node11.html*