A BRIEF INTRODUCTION TO METHODS OF GENE PREDICTION

Jakub Filák
Doctoral Degree Programme (1), FIT BUT
E-mail: xfilak01@stud.fit.vutbr.cz

Supervised by: Jaroslav Zendulka
E-mail: zendulka@fit.vutbr.cz

ABSTRACT

Recent research has shown that we can get a full genome sequence in a few days and for a little amount of money. However, the raw genomic sequence is not very useful, so we need to know what does the sequence means. There is a place for many types of sequence analysis, one of which is the gene prediction. This paper reviews the existing approaches to predict eukaryotic genes and, finally, several possible improvements are encountered.

1 INTRODUCTION

Cells in living organisms are mostly formed from proteins. Coding instruction how to build the proteins are stored in Deoxyribonucleic acid (DNA). However, not all DNA codes for the proteins. DNA is a sequence of nucleic acids and only specific regions of the sequence called genes are used as blueprints for the proteins synthesis. Proteins are not only used for forming the cell body but they are used in almost all processes in cells and living organisms.

Genes are special regions in genomic DNA and they have own specific structural organisation. The structure is common for prokaryotic and eukaryotic cells in basics. Every gene in both cell types have preceding sequence called promoter, which facilitates transformation of genes to proteins. Other common property is a start and stop codon. The codon is a unit of a genetic code responding to amino acid in a protein chain. However, differences between genes in both cell types are more significant than similarity. A prokaryotic gene sequence is completely transformed to protein but eukaryotic gene sequences contain spliced regions called introns. Eukaryotic genes have much more complex structure in general, which is complication in the gene prediction.

We know a lot about processes in organism, when genes and all of their products are revealed. Lots of genes have more variations called allels. Each of the allels has different behaviour because they code for different proteins. Sometimes they code for the protein with an inappropriate function, which can cause disease. The current state gives us opportunity for a treatment of genetic diseases by providing appropriate proteins or even changing the bad gene allel with the seamless allel.

In this review, we start with the description of basic principles of gene prediction, without going

too deeply. After this, we briefly describe several main algorithms and we take a look at a few possible improvements.

2 APPROACHES

For simplicity, we consider only the eukaryotic genes. The eukaryotic genes have more complex structure than the prokaryotic genes and algorithms for their search have to be more sophisticated. The prokaryotic gene has a stable structure which is in all cases flanked by the start and stop codon. If the probability of the codon occurrence has a uniform distribution the genes can be found by encountering all of Open Reading Frames (ORF, sequence between the start and stop codon) longer than an appropriate constant [7].

Essentially, the gene prediction approaches can be divided into two main classes. The first class is a content sensor that classify a DNA region into types, e.g. coding versus non-coding. The second class is a signal sensor that try to detect presence of functional sites specific to the genes. At last, there are computer programs combining both classes [5].

2.1 CONTENT SENSORS

Methods of the gene prediction using the content sensors can be divided into two main categories. The first category is called extrinsic. The extrinsic methods can be described as methods of comparative genetics, because the prediction of genes compares the examined sequence with another sequences. In the second category, there are methods using various statistical models and algorithms. Methods of second group try to determine relevance of each nucleotide to some of the gene structures. The second group is also called intrinsic or *ab initio* prediction.

Extrinsic sensors The natural approach is a simple local alignment of the examined sequence with another DNA sequence or protein. The basic tools are Smith-Water algorithm or fast heuristic algorithms such as FASTA or BLAST. The alignment with the sequences from an other organism assume that such important sequences (e.g. genes) are well evolutionary conserved. Assuming the preceding condition, the results of the alignment denotes the genes.

Another sources of the alignment are the sequences obtained from the current organism in form of cDNA or Expressed Sequence Tags (EST). Sources of the organism provide accurate results for the gene structure. However, they are obtained after a post translation modification and contain a large number of errors. Even, we can use the proteins as the sources but an advantage is same as in the last mentioned sources. By this approach, about 50% of genes can be revealed.

Intrinsic sensors This approach uses a knowledge about the gene structure and a genome sequence rules in order to classify sequence regions. Example of this approach is the encountering of all ORFs in prokaryote, mentioned above. Several other measures were introduced. The simplest measure is a nucleotide composition and its variation with more consequence nucleotides. The most effective measure is a hexamer frequency, which is a frequency of six consequence nucleotides [7]. More generally, the measure based on the frequency of the sequence of length k is Markov chain of order k, which is described in section 3 Algorithms.

2.2 SIGNAL SENSORS

Methods based on a signal detection are similar to the *ab initio* category methods. Methods create statistical models describing shorter stretches of DNA sequences. In the introductory chapter sequences surrounding both prokaryotic and eukaryotic genes were mentioned. These sequences can be considered as signals for cellular enzymes, providing information about where the gene is located, whether it can be converted into a protein, or how to implement a cut. We can use the signals as enzymes to determine the location of the genes and their whole structure.

Theoretically, the sequence of each signal is always the same and never occurs in a place different to where it functions as the signal. In fact, the sequences of signals vary among the genes in a single organism and the sequence may occur at different places, making it difficult to identify the useful signals from random sequences. The situation is not as desperate as it might seem. The sequences of signals, although different, often occur in several variants and the presence of the signal usually indicates the presence of the other signals.

The essence of this approach is to create a model for each known signal to identify the genes. This approach has to rely on a fact that there are known enough sequences to each signal and the parameter estimation of the model is precise enough. The natural approach is to find good signal sequences and their alignment, for example using a method CLUSTAL, and to derive the model parameters from the results. The problem is how to create the alignment of the sequences. The signal sequences are very short and it is therefore necessary to use special alignment algorithms.

Some signals can have similar sequences across majority of related organisms and smaller amount of data is sufficient to estimate the model parameters. However, other signals may be very variable and large amounts of data is needed. If it is not possible to obtain enough sequences, it is necessary to choose suitable simplification of the model.

3 ALGORITHMS

Now we take a look at several main algorithms used in the gene prediction. The most frequently used algorithms use Markov chains and their variations such as Hidden Markov Model or Inhomogenous Periodic Markov Model [2].

Markov chain is a statistician model with discreet set of states which can be changed in discrete steps. In this model, DNA is a consequence of random variables $X_1, X_2, .., X_n$ and each variable X_i responds to the position i in the sequence. The probability of the variable X_i depends on value of variable X_{i-1}. This model is very simple and higher order Markov chains are used. In Markov chains of order k the probability of the random variable X_i depends on the random variables X_{i-o} where o is from 1 to $k+1$[3].

Probability of nucleotide occurrence is given by a transitional table. Current nucleotides are in rows of the transitional table and coming nucleotides are in columns and the probability of transition is in the table body. The transitional table remains unchanged for each sequence position in regular Markov chains. However, the transitional table in the genes vary. Introns may be modeled as regular Markov chains but for exons modeling Inhomogenous Markov chains and especial Inhomogenous Periodic Markov chains are used. The transitional table vary at each step in Inhomogenous Markov chains. In Periodic Markov chains the transitional table is constant in a period of length k [6].

The mostly used variant of Markov chains is Hidden Markov Model (HMM). HMM is more

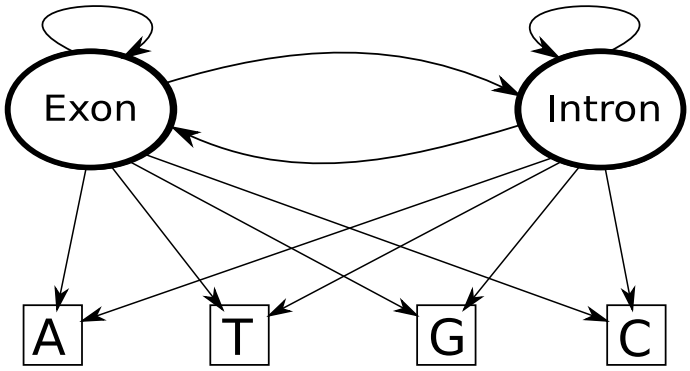

Figure 1: The simplest Hidden Markov Model of eukaryotic gene. (no probability of the edges for clarity)

general, and hence more flexible, allowing us to model phenomena that we cannot model sufficiently well with regular Markov chain model. In this model we add emission of letter in visited states. Here states respond to structural features and letters from a sequence are emitted in each state with different probability [3]. Structural features are signals, thus HMM are used as a primary algorithm in signals sensors. A graphic representation of eukaryotic gene HMM is in Figure 1.

Several methods from machine learning were used. Decision trees are used as primary algorithm and for joining results of other algorithms in order to improve accuracy. The usage of neural networks and support vector machine is same as decision trees. When we use one of the algorithms mentioned as primary algorithm, we often have several measures from intrinsic content sensors and algorithms try to classify DNA regions. Neural networks can be used as a signal model in signal sensors.

Also algorithms using Digital Signal Processing (DSP) methods have been proposed. The mostly used methods are Z-curve and Discrete Fourier Transformation. Their main goal is to find a signal. The genetic code can be considered as this signal. DSP methods are naturally used as intrinsic content sensors. These methods seem to be suitable for both eukaryotic and prokaryotic genomes and their strange is that they do not require prior information.

4 FUTURE WORK

The research of gene prediction field is currently strongly explored. There are many methods to reveal genes, so there is a small place for completely new methods of analysis. Better we need to improve the accuracy of existing methods. One way to achieve this is a combination of several methods and sources. This effort has improved the accuracy of the gene prediction results [1].

Next improvement is to use a physical map of genome in order to improve the predicted gene results. It was discovered that all genes are physically located in clusters [4]. Knowledge of the physical gene location can be used in the models to improvement of their parameters. Even, the gene location can be used for selection of models appropriate for gene families located in a physical cluster near the examined DNA region.

Last but not least possible improvement lies in DNA sequencing process. Modern sequencing devices are fast but not accurate as well. We need to do more research in this field because errors of devices are explored a little. There are known general errors produced by sequencing devices. These errors must be integrated into the gene prediction models. This situation has two possible solutions. The first is the integration of errors mentioned above and the second is elimination of these errors in the sequencing process.

5 CONCLUSION

The prediction of genes and especially those coding for proteins was briefly described. There exists many methods but most of them are based on Markov chains and their variations. Currently, we do not need new methods but we need more accurate results. Several algorithms combining results of other methods was introduced to achieve this goal and this approach seems promising.

ACKNOWLEDGEMENT

This work was partially supported by the BUT FIT grant FIT-S-10-1 and the research plan MSM0021630528.

REFERENCES

[1] Jonathan E. Allen, Mihaela Pertea, and Steven L. Salzberg. Computational gene prediction using multiple sources of evidence. *Genome Research*, pages 142–148, 2004.

[2] J. H. Do and D. K. Choi. Computational approaches to gene prediction. *J Microbiol*, 44(2):137–144, 2006.

[3] Warre J. Ewewns and Gregory R. Grant. *Statistical Methods in Bioinformatics : An Introduction*. Springer, second edition edition, 2005. ISBN-978-0-387-40082-2.

[4] Erez Lieberman-Aiden, Nynke L. van Berkum, Louise Williams, and et. al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science (New York, N.Y.)*, 326(5950):289–293, October 2009.

[5] Catherine Mathe, Marie-France Sagot, Thomas Schiex, and Pierre Rouze. Current methods of gene prediction, their strengths and weaknesses. *Nucl. Acids Res.*, 30(19):4103–4117, 2002.

[6] SL Salzberg, AL Delcher, S Kasif, and O White. Microbial gene identification using interpolated Markov models. *Nucl. Acids Res.*, 26(2):544–548, 1998.

[7] Marketa Zvelebil and Jeremy Baum. *Understanding Bioinformatics*. Garland Science, 1 edition, 2007. ISBN-0815340249.