FACULTY OF ENGINEERING AND TECHNOLOGY

Department of Biotechnology

Gene Prediction methods

With the development of genome sequencing for many organisms, more and more raw sequences need to be annotated. Gene prediction by computational methods for finding the location of protein coding regions is one of the essential issues in bioinformatics. Two classes of methods are generally adopted: similarity based searches and *ab initio* prediction. Here, we review the development of gene prediction methods, summarize the measures for evaluating predictor quality, highlight open problems in this area, and discuss future research directions.

XYZ

Since the beginning of the Human Genome Program (HGP) in 1990, databases of human and model organism DNA sequences have been increasing quickly. Computational gene prediction is becoming more and more essential for the automatic analysis and annotation of large uncharacterized genomic sequences. In the past two decades, many gene prediction programs have been developed. Most of them are referenced at the website maintained by Wentian Li (http://www.nslij-genetics.org/gene/).
Gene discovery in prokaryotic genomes is less difficult, due to the higher gene density typical of prokaryotes and the absence of introns in their protein coding regions. DNA sequences that encode proteins are transcribed into mRNA, and the mRNA is usually translated into proteins without significant modification. The longest ORFs (open reading frames) running from the first available start codon on the mRNA to the next stop codon in the same reading frame generally provide a good, but not assured prediction of the protein coding regions. Several methods have been devised that use different types of Markov models in order to capture the compositional differences among coding regions, "shadow" coding regions (coding on the opposite DNA strand), and noncoding DNA. Such methods, including ECOPARSE, the widely used GENMARK, and Glimmer program, appear to be able to identify most protein coding genes with good performance *(1)*.

XYZ

In eukaryotic organisms, it is a quite different problem from that encountered in prokaryotes. Transcription of protein coding regions initiated at specific promoter sequences is followed by removal of noncoding sequences (introns) from pre-mRNA by a splicing mechanism, leaving the protein-encoding exons. Once the introns have been removed and certain other modifications to the mature RNA have been made, the resulting mature mRNA can be translated in the 5′ to 3′ direction, usually from the first start codon to the first stop codon. As a result of the presence of intron sequences in the genomic DNA sequences of eukaryotes, the ORF corresponding to an encoded gene will be interrupted by the presence of introns that usually generate stop codons *(2)*. This review mainly focuses on the more complex problem of gene prediction in eukaryotic sequences.

Gene Prediction Methods

There are two basic problems in gene prediction: prediction of protein coding regions and prediction of the functional sites of genes. A large number of researches working on this subject have accumulated, which can be classified into four generations in summary. The first generation of programs was designed to identify approximate locations of coding regions in genomic DNA. The most widely known programs were probably TestCode *(3)* and GRAIL *(4)*. But they could not accurately predict precise exon locations. The second generation, such as SORFIND *(5)* and Xpound *(6)*, combined splice signal and coding region identification to predict potential exons, but did not attempt to assemble predicted exons into complete genes. The next generation of programs attempted the more difficult task of predicting complete gene structures. A variety of programs have been developed, including GeneID *(7)*, GeneParser 8., 9., GenLang *(10)*, and FGENEH *(11)*. However, the performance of those programs remained rather poor. Moreover, those programs were all based on the assumption that the input sequence contains exactly one complete gene, which is not often the case. To solve this problem and improve accuracy and applicability further, GENSCAN *(12)* and AUGUSTUS *(13)* were developed, which could be classified into the fourth generation.

There are mainly two classes of methods for computational gene prediction. One is based on sequence similarity searches, while the other is gene structure and signal-based searches, which is also referred to as *ab initio* gene finding.

Ab initio gene prediction methods

The second class of methods for the computational identification of genes is to use gene structure as a template to detect genes, which is also called *ab initio* prediction. *Ab initio* gene predictions rely on two types of sequence information: signal sensors and content sensors. Signal sensors refer to short sequence motifs, such as splice sites, branch points, polypyrimidine tracts, start codons and stop codons. Exon detection must rely on the content sensors, which refer to the patterns of codon usage that are unique to a species, and allow coding sequences to be distinguished from the surrounding non-coding sequences by statistical detection algorithms.

Many algorithms are applied for modeling gene structure, such as Dynamic Programming, linear discriminant analysis, Linguist methods, Hidden Markov Model and Neural Network. Based on these models, a great number of *ab initio* gene prediction programs have been developed. Some of the frequently used ones are shown in Table 1, among which the programs GeneParser, Genie and GRAIL combine similarity searches.

Table 1 *Ab initio* Gene Prediction Programs (Possibly with Homology Integration)

Program	Organism	Algorithm*	Website	Homology
GeneID	Vertebrates, plants	DP	http://www1.imim.es/geneid.html	
FGENESH	Human, mouse, Drosophila, rice	HMM	http://www.softberry.com/berry.phtml?topic=fgenesh&group=programs&subgroup=gfind	
GeneParser	Vertebrates	NN	http://beagle.colorado.edu/~eesnyder/GeneParser.html	EST
Genie	Drosophila, human, other	GHMM	http://www.fruitfly.org/seq_tools/genie.html	protein
GenLang	Vertebrates, Drosophila, dicots	Grammar rule	http://www.cbil.upenn.edu/genlang/genlang_home.html	
GENSCAN	Vertebrates, Arabidopsis, maize	GHMM	http://genes.mit.edu/GENSCAN.html	
GlimmerM	Small eukaryotes, Arabidopsis, rice	IMM	http://www.tigr.org/tdb/glimmerm/glmr_form.html	
GRAIL	Human, mouse, Arabidopsis, Drosophila	NN, DP	http://compbio.ornl.gov/Grail-bin/EmptyGrailForm	EST, cDNA
HMMgene	Vertebrates, *C. elegans*	CHMM	http://www.cbs.dtu.dk/services/HMMgene/	
AUGUSTUS	Human, Arabidopsis	IMM,WWAM	http://augustus.gobics.de/	
MZEF	Human, mouse, Arabidopsis, Fission yeast	Quadratic discriminant analysis	http://rulai.cshl.org/tools/genefinder/	

*DP, dynamic programming; NN, neural network; MM, Markov model; HMM, Hidden Markov model; CHMM, class HMM; GHMM, generalized HMM; IMM, interpolated MM.

Pattern Recognition,

Pattern recognition is the automated recognition of patterns and regularities in data. It has applications in statistical data analysis, signal processing, image analysis, information retrieval, bioinformatics, data compression, computer graphics and machine learning. Pattern recognition has its origins in statistics and engineering; some modern approaches to pattern recognition include the use of machine learning, due to the increased availability of big data and a new abundance of processing power.

A modern definition of pattern recognition is:

The field of pattern recognition is concerned with the automatic discovery of regularities in data through the use of computer algorithms and with the use of these regularities to take actions such as classifying the data into different categories.

This article focuses on machine learning approaches to pattern recognition. Pattern recognition systems are in many cases trained from labeled "training" data (supervised learning), but when no labeled data are available other algorithms can be used to discover previously unknown patterns (unsupervised learning).

Machine learning is strongly related to pattern recognition and originates from artificial intelligence.

In pattern recognition, there may be a higher interest to formalize, explain and visualize the pattern, while machine learning traditionally focuses on maximizing the recognition rates. Yet, all of these domains have evolved substantially from their roots in artificial intelligence, engineering and statistics.

✓In machine learning, pattern recognition is the assignment of a label to a given input value. In statistics, discriminant analysis was introduced for this same purpose in 1936.

✓An example of pattern recognition is classification, which attempts to assign each input value to one of a given set of *classes* (for example, determine whether a given email is "spam" or "non-spam").

✓However, pattern recognition is a more general problem that encompasses other types of output as well. Other examples are regression, which assigns a real-valued output to each input; sequence labeling, which assigns a class to each member of a sequence of values (for example, part of speech tagging, which assigns a part of speech to each word in an input sentence); and parsing, which assigns a parse tree to an input sentence, describing the syntactic structure of the sentence.

XYZ

abc

XYZ

abc

XYZ

abc

XYZ

abc

XYZ

abc

XYZ

abc

MCQs

1. A
2. A
3. A
4. A
5. A
6. A
7. A
8. A
9. A
10. A