Trends in Bioinformatics

ISSN 1994-7941

science alert

ANSInet
an open access publisher
http://ansinet.com

Trends in Bioinformatics 4 (1): 10-22, 2011
ISSN 1994-7941 / DOI: 10.3923/tb.2011.10.22

Evaluation of Prediction Accuracy of Genefinders Using Mouse Genomic DNA

[1]J. Nasiri, [2]A. Haghnazari and [1]M. Alavi
[1]Department of Plant Breeding, College of Agriculture, Zanjan University, Zanjan, Iran
[2]Department of Genetics and Plant Breeding, College of Agriculture, Zanjan University, Zanjan, Iran

Corresponding Author: Jaber Nasiri, Department of Plant Breeding, College of Agriculture, Zanjan University, Zanjan, Iran Tel: +982415152465/+989133220399 Fax: +982415152546

ABSTRACT

Six gene-finding programs i.e., Genscan, GeneMark hmm., HMMgene, GenView2, FGENESH and FGENESH+ were evaluated using 24 well defined mouse single- and multiexon genes to predict the structure of protein coding genes. Our analyses indicated that different methods often produce different and sometimes contradictory-results. In the nucleotide level, the highest correlation coefficient (0.87) and approximate correlation (0.86) values and also the lowest correlation coefficient (0.67) and approximate correlation (0.67) values were detected only for FGENESH+ and GenView 2 programs, respectively. Furthermore, at the exon level, similar results were obtained. In general, our results at either the nucleotide or exon levels showed that FGENESH+ (HMM plus sequence similarity programs), provide a level of improvement over the *ab initio* gene prediction methods such as Genview 2 and suggested that, probably, FGENESH+ and also Genscan can be more helpful than the others. Meanwhile, based on phylogenetic tree, all *ab initio* genefinders, excepted of GeneMarkhmm., were placed in the same group and FGENESH+ with GeneMarkhmm. programs assigned in another one. Moreover, based on our results, we realized that the accuracy of these programs, is strongly dependent on GC content. At last, on the basis of whole known sequences it was concluded that predictive accuracy of these programs is lower than actual.

Key words: GC content, gene prediction, mouse genome, protein

INTRODUCTION

As soon as the genome of an organism is sequenced and assembled, one of the first ensuing tasks is to locate all of the protein-coding genes hidden within the genomic sequence (Blanco and Guigo, 2005). This is a necessary step toward understanding better the functional content of the genome. Stanke *et al.* (2004) have suggested that the prediction of all gene structures in a given genomic sequence is the first step in genome annotation. The development of gene-finding methods is, therefore, an important field in biological sequence analysis. The genes of most eukaryotic organisms are neither continuous nor contiguous (Rogic *et al.*, 2001). In eukaryotic organisms identification of genes is more difficult than prokaryotes, because of the split nature of eukaryotic genes and because of the often large spacers found between adjacent genes (Stanke *et al.*, 2004).

In total, the mouse genome is a sequence of approximately 2.6 billion nucleotides assembled in 21 chromosomes numbered 1-19 plus the X and Y sex chromosomes. The chromosomes are numbered mostly in descending order, from the largest, chromosome 1, containing 195 million

nucleotides to the smallest with, chromosome 19, containing 60 million nucleotides. Chromosome Y is smaller with 48 milhon nucleotides, while chromosome X falls in between chromosomes 2 and 3 with 160 million nucleotides. There are approximately 27 thousand known mouse genes, which code 33 thousand known transcripts (Hubbard *et al.*, 2005). Of these, the sequences for approximately 20 thousand transcripts are definitely known (Pruitt *et al.*, 2005).

There are several methods introduced for the experimental discovery of genes, but they are time-consuming and costly. Accordingly, for the last 15 years researchers have been developing computational methods for gene-finding that can automate, or facilitate, the identification of genes (Rogic *et al.*, 2001). Two basic approaches have been generally established for computational gene-finding: the sequence similarity search, look-up or extrinsic method (including sequence similarity gene prediction and comparative gene prediction) and *ab initio* gene prediction methods (including searching by signal and searching by content) (Fickett, 1996). The latter method is also commonly referred to as intrinsic or template gene prediction. So far, different studies were applied about of gene finding methods and gene prediction programs in various organisms. For instance, Snyder and Stormo (1995) analyzed three gene-finding programs on rather limited test sets containing 28 and 34 sequences. A more comprehensive evaluation of gene structure prediction programs was done by Burset and Guigo (1996), that they evaluated nine programs, using a set of 570 vertebrate single-gene sequences. Salamov and Solovyev (2000) analyzed some *ab initio* gene finding programs in *Drosophila* genomic DNA. Rogic *et al.* (2001) published a comparative analysis of seven gene prediction programs, using a set of 195 single-gene sequences of Human and rodent species. Parra *et al.* (2003) published a report of a number of comparative gene prediction programs in Human and Mouse genome. Guigo and Wiehe (2003) evaluated accuracy of a number of comparative and *ab initio* gene finder programs on Human chromosome 22. Stanke *et al.* (2004) measured accuracy of AUGUSTUS, GeneID and Genie programs on the *Drosophila Adh* region and also they compared the accuracy of AUGUSTUS and GENSCAN programs on 178 human single-gene sequences. Yao *et al.* (2005) were used five *ab initio* gene prediction programs for the discovery of maize genes. Stanke *et al.* (2006) published a report on the accuracy values of variants of AUGUSTUS on the ENCODE test set with 296 genes and an average of 2.2 transcripts per gene.

In present study, we used six gene finding programs using a set of 24 single- and multi-exon gene sequences of Mouse genome to obtain following objectives:

- Do different programs and different methods predict same results or not?
- If for different programs detected different results, which of them is in agreement with the actual gene structures?
- Since in the previous studies related of genefinders, we saw no data about FGNENSH+ program using mouse sequences, thus for the first time, we examined the accuracy of FGNENSH+ program and compared its results with data obtained from the other programs based on *ab initio* method

MATERIALS AND METHODS

Current study was performed to realize the potential of six recently developed gene finding programs i.e., FGENESH (Salamov and Solovyev, 2000), Genview 2 (Milanesi *et al.*, 1993), Genscan (Burge and Karlin, 1997), HMMgene (Krogh, 1997), GeneMark hmm. (Lukashin and Borodovsky, 1998) and FGENESH+ (Salamov and Solovyev, 2000). The programs discussed here

Table 1: List of gene prediction programs used for this study

Program	Trained on	Available at
HMMgene	Vertebrates and C.elegans	http://l25.itba.mi.cnr.it/~webgene/wwwgene.html
Genemark.hmm	Human, mouse, Drosophila, Gallus gallus, Arabidopsis, rice, maize, Chlamydomonas reinhardtii, C.elegans, barley and wheat	http://opal.biology.gatech.edu/GeneMark/eukhmm.ci
Genscan	Vertebrates, Arabidopsis and maize	http://genes.mit.edu/GENSCAN.html
GenView2	Human, mouse and diptera	http://l25.itba.mi.cnr.it/~webgene/wwwgene.html
FGENESH	Human, mouse, Drosophila, Gallus gallus, Arabidopsis, rice, maize, Chlamydomonas reinhardtii, C.elegans, barley, wheat and so on	http://www.softberry.com/berry.phtml?topic=gfind-file
FGENESH+	Human, mouse, Drosophila, Gallus gallus, Arabidopsis, rice, maize, Chlamydomonas reinhardtii, C.elegans, barley, wheat and so on	http://www.softberry.com/berry.phtml?topic=gfind-file

can be accessed through easy-to-use Web frontends (Table 1). Among these programs, FGENESH+ is only based on sequence similarity. In fact, the program is a variant of FGENESH that takes into account some information about similar proteins (Salamov and Solovyev, 2000). As compared FGENESH program, FGENESH+ requires an additional file with protein homolog and aligns all predicted potential exons with that protein using the Smith Waterman algorithm, as implemented in the sim program (Huang and Miller, 1991). The program can be used if the protein sequence similar with protein which is encoded by the gene to be available. Furthermore, to analyze real data with these programs, 24 single genes (containing single- or multi-exon) of Mouse genome were used and every character of any gene (i.e., TATA box, length of any gene (bp), number of exons, number of introns, start and stop points of any exon and length of amino acid sequence of any gene) were obtained from National Center for Biotechnology Information (NCBI) website. Of 24 genes, only one gene (AF010405.2 accession) consists of no introns in the open reading frame (commonly referred to as single exon gene) and 23 remained genes were multi-exon genes, from 2 to 26 exons and the mean number of coding exons per gene was 6.0. In Total, in all genes, there were 144 coding exons.

For all programs, the accuracy of the predictions were measured at two different levels: coding nucleotide sequence and exonic structure. Moreover, we examined the programs accuracy based on CG content (Burset and Guigo, 1996). Note that, only the exons predicted on the forward strand (predictions for the reverse strand were ignored) were compared to the actual coding exons.

Statistics: The accuracy of gene prediction programs is usually determined using controlled, defined data set, comparing the prediction made by a method with the actual gene structure, determined experimentally (Blanco and Guigo, 2005). Accordingly, twenty four Mouse single- or multiexon genes were used as a defined data sets and each program was run by them and totally 144 predictions were obtained. An output of six gene prediction programs as well as actual structure of two single-gene sequences is exemplified in Fig. 1. This Fig. 1 was produced by using the gff2ps software (Abril and Guigo, 2000).

At the nucleotide level, we measured the accuracy of a prediction on a test sequence by comparing the predicted coding value (coding or noncoding) with the true coding value for each nucleotide along the test sequence. This has been one of the most widely used approaches in evaluating the accuracy of coding region identification and gene structure prediction methods (Burset and Guigo, 1996).

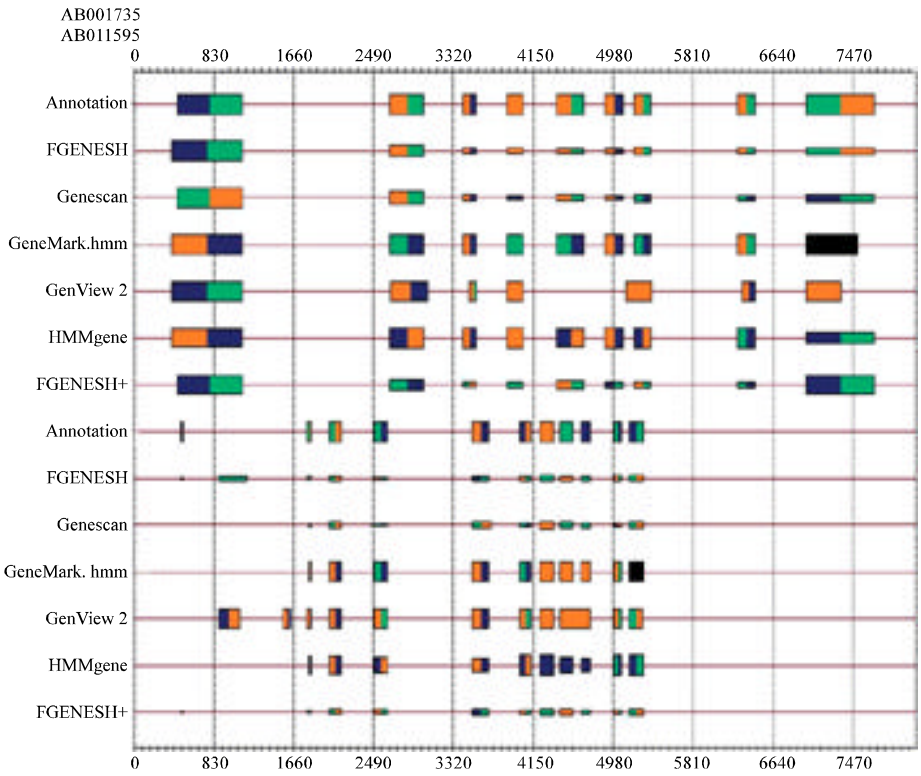

Fig. 1: Comparison of gene prediction by six genefinders for AB001735 and AB011595 accessions (Total size: 9248 and 6190 bp, respectively) of mouse using gff2ps program (Abril and Guigo, 2000). The annotation line shows the structure of the actual genes, This plot has been obtained using gff2ps. The most recent version of gff2ps is freely available at http://www.imim.es/softcols/GFF2PS.html. Copy right 1999 by Josep F. Abril and Roderic Guigo

Based on Burset and Guigo (1996), if we define the values of the TP (true positives), TN (true negatives), FP (false positives) and FN (false negatives) as follows:

TP = The number of coding nucleotides predicted as coding
TN = The number of noncoding nucleotides predicted as noncoding
FP = The number of noncoding nucleotides predicted as coding
FN = The number of coding nucleotides predicted as noncoding

Then, the sensitivity (i.e., the proportion of coding nucleotides that are correctly predicted as coding) and specificity (i.e., the proportion of nucleotides predicted as coding that are actually coding) values can be calculated by the following formulas:

$$Sn = \frac{TP}{TP + FN} \quad \text{and} \quad Sp = \frac{TP}{TP + FP}$$

These are widely used measurements of accuracy for gene prediction programs (Rogic *et al.*, 2001). In general, Sn and Sp can take on values from zero to one; for a perfect prediction, Sn = 1

and Sp = 1. Neither Sn nor Sp alone provide a good measure of global accuracy, because high Sn can be achieved with little Sp and vice versa (Burset and Guigo, 1996). An easier to understand measure that combines the Sn and Sp values is called the Correlation Coefficient (CC). From the above 2×2 table, CC is defined as:

$$CC = \frac{(TP \times TN) - (FN \times FP)}{\sqrt{TP + FN)(TN + FP)(TP + FP)(TN + FN)}}$$

The CC ranges from -1 to 1, were a value of 1 corresponds to a perfect prediction: a value of -1 indicates that every coding region has been predicted as non-coding and vice versa. To evaluate gene structure prediction programs, this parameter has been widely used (Guigo *et al*., 1992; Solovyev *et al*., 1994; Snyder and Stormo, 1995).

A measure with similar characteristics, but defined under any circumstance, is the approximate correlation (AC), introduced in Burset and Guigo (1996), defined as:

$$AC = (ACP\text{-}0.5)\times 2$$

where, ACP is the average conditional probability defined as:

$$ACP = \frac{1}{4}\left(\frac{TP}{TP + FN} + \frac{TP}{TP + FP} + \frac{TN}{TN + FP} + \frac{TN}{TN + FN}\right)$$

Since, at least two of the conditional probabilities in this formula are always defined, ACP can always be calculated as the average of the one defined. CC and AC range over [-1, 1] and usually are close to each other whenever CC is defined. Nucleotide level accuracy measures indicate how good the search by content element of the program is, but they don't tell us much about the search by signal component (Rogic *et al*., 2001).

In addition, We measured both sensitivity and specificity to evaluate accuracy at the exonic structure using the following formulas:

$$ESp = \frac{TE}{PE} \quad \text{and} \quad ESn = \frac{TE}{AE}$$

where, TE (true exons) is the number of correctly predicted exons, AE (actual exons) is the number of annotated exons and PE (predicted exons) is the number of predicted exons (Rogic *et al*., 2001). Thus, Sensitivity is the proportion of actual exons in the test sequence that are correctly predicted and Specificity is the proportion of predicted exons that are correctly predicted methods (Burset and Guigo, 1996). Finally, based on Burset and Guigo (1996), for all programs, we calculated and compared first the proportion of true exons without overlap to predicted exons-the Missing Exons (ME)- and the proportion of predicted exons without overlap to actual exons- the Wrong Exons (WE)- using the following formulas:

$$ME = \left(\frac{\text{No. of missing exons}}{\text{No. of actual exons}}\right) \quad \text{and} \quad WE = \left(\frac{\text{No. of wrong exons}}{\text{No. of predicted exons}}\right)$$

RESULTS

Nucleotide level: The accuracy measures of six programs at the both nucleotide and exon levels have been shown in Table 2. In the nucleotide level, the highest Sn was detected for both FGENESH+ (0.98) and FGENESH (0.98) programs and the lowest Sn (0.82) were detected only for GenView2 program. In the meantime, the maximum and minimum Sp were detected for Genscan (0.79) and GenView 2 (0.65) programs, respectively. Therefore, we can not decidedly say that which programs have the highest or the lowest potential to predict structure of a given gene. Conversely, when we calculated the CC and AC values, it was identified that the maximum CC (0.87) and AC (0.86) values were detected only for programs based on sequence similarity (i.e., FGENESH+ program) whereas the minimum CC (0.67) and AC (0.67) were found for programs based on *ab initio* method (i.e., GenView 2).

Exon level: At the exon level, maximum ESn and ESp were detected only for FGENESH+ program (0.72 and 0.70, respectively) and minimum ESn (0.31) and ESp (0.41) were obtained only for GenView 2 program. Consequently, the highest and lowest mean average of ESn and ESp were detected for FGENESH+ (0.71) and GenView2 (0.36) programs, respectively (Table 2). Moreover, the maximum wrong exon (WE) and missing exon (ME) were detected for Genemark.hmm and GenView 2 programs (0.13 and 0.29, respectively), whereas, the minimum Wrong Exon (WE) and Missing Exon (ME) were detected only for FGENESH+ program (0.00 and 0.07, respectively) (Table 2).

On the other hand, for the all programs, we measured the number of initial, internal and terminal exons to identify the power of any programs to predict the exon classes (Table 3). Based

Table 2: The relative nucleotide and exon level accuracy of *ab initio* and sequence similarity gene finding programs

	Nucleotide level				Exon level				
Programs*	Sn	Sp	CC	AC	Esp	Esn	Sp+Sn/2	WE	ME
HMMgene	0.95	0.78	0.84	0.84	0.64	0.60	0.62±0.36	0.04	0.15
Genemark.hmm	0.95	0.74	0.79	0.79	0.52	0.52	0.53±0.26	0.13	0.12
Genscan	0.97	0.79	0.86	0.85	0.71	0.67	0.70±0.29	0.05	0.12
GenView2	0.82	0.65	0.67	0.67	0.41	0.31	0.36±0.31	0.08	0.29
FGENESH	0.98	0.75	0.84	0.83	0.70	0.68	0.69±0.32	0.01	0.08
FGENESH+	0.98	0.78	0.87	0.86	0.72	0.70	0.71±0.34	0.00	0.07

*For each sequence, the exons predicted on the forward (+) strand were compared to the annotated exons. The standard measures of predictive accuracy on nucleotide and exon level were calculated for each sequence and averaged over all sequences for which they were defined. This was done separately for each of the programs tested

Table 3: Predicted No. of exons in each class on 23 multiexon genes. The data given in the table are the TE/PE*

Class	Initial	Internal	Terminal	Total
AE	23	97	23	143
HMMgene	9/18	83/91	17/19	109/128
Genemark.hmm	12/18	85/112	13/25	110/155
Genscan	11/15	88/98	17/20	116/133
GenView 2	8/16	45/73	9/14	62/103
FGENESH	11/18	89/93	16/20	116/131
FGENESH+	14/23	88/92	16/20	113/135

*TE = Trne exons; AE = Actual exons and PE = Predicted exons

Table 4: Sensitivity and Specificity of predictions for various classes of exons on 23 multiexon genes

	Initial		Internal		Terminal		Total	
Programs	ESn	ESp	ESn	ESp	ESn	ESp	ESn	Esp
HMMgene	0.39	0.50	0.86	0.91	0.74	0.89	0.76	0.85
Genemark.hmm	0.52	0.67	0.88	0.76	0.57	0.52	0.76	0.71
Genscan	0.48	0.73	0.91	0.90	0.74	0.85	0.81	0.87
GenView 2	0.35	0.50	0.46	0.62	0.39	0.64	0.43	0.60
FGENESH	0.48	0.61	0.92	0.96	0.70	0.80	0.81	0.89
FGENESH+	0.61	0.61	0.91	0.96	0.70	0.80	0.79	0.84

on our results, the maximum number of correctly initial and internal exons were obtained for FGENESH+ and FGENESH programs and both Genscan and HMMgene programs had the highest values of terminal exon.

Nevertheless, the calculated values of ESn and ESp for various classes of exons were relatively different than the calculated number of them, so that for initial exon, the highest and lowest ESn were detected for FGENESH+ and GenView 2 programs, respectively. Indeed, for internal exon, the highest and lowest *ESn* were detected for FGENESH and GenView 2 programs and finally for terminal exon, the highest *ESn* was detected for both HMMgene and Genscan programs and lowest ESn was detected only for GenView 2 program (Table 4). On the whole, contrary to nucleotide level, the highest and lowest ESn, ESp and also the average of ESn and ESp were found out only for two programs (i.e., FGENESH+ and Genview 2 programs, respectively).

G + C content: GC content (or guanine-cytosine content), in molecular biology, is the percentage of nitrogenous bases on a DNA molecule which are either guanine or cytosine. In general, GC-content percentage is calculated as following formula:

$$\text{GC content} = \frac{G+C}{G+C+T+A} \times 100$$

In this case, it has been pointed that the GC content is correlated with various genomic features comprised of repeat element distribution, methylation pattern (Jabbari and Bernardi, 1998) and most remarkably, gene density (Mouchiroud *et al*., 1991; Duret *et al*., 1995). GC-rich regions include many genes with short introns while GC-poor regions are essentially deserts of genes. Moreover, Galtiera *et al*. (2001) have suggested that the distribution of GC content in mammals could have some functional relevance, raising the issue of its origin and evolution (Galtiera *et al*., 2001). Since in some studies have been mentioned that the GC content has an important role in the accuracy of some gene prediction programs (Snyder and Stormo, 1995; Lopez *et al*., 1994; Burset and Guigo, 1996; Rogic *et al*., 2001), accordingly, here, we investigated the effect of this item on accuracy of gene prediction programs.

Although, the overall GC content of the mouse genome is slightly higher than that of human (42 vs. 41%), the human genome exhibits a much greater variability, when measured using non-overlapping 20 kb windows. In the human genome, 2.7% of the 20 kb segments have GC content of greater then 56% or less then 33%; this kind of variability is virtually absent in the mouse genome (Waterston *et al*., 2002). While the correlation between gene distribution and GC content has been shown in humans (Zoubak and Bernardi, 1996), as well as other vertebrates

Table 5: Accuracy versus G + C content

	<50% (13)		≥50% (11)	
Programs*	AC	Esp +ESn /2	AC	ESp +ESn /2
HMMgene	0.81	0.61	0.86	0.64
Genemark.hmm	0.74	0.41	0.85	0.67
Geuscan	0.81	0.66	0.89	0.73
GenView 2	0.66	0.32	0.68	0.40
FGENESH	0.80	0.59	0.86	0.80
FGENESH+	0.83	0.60	0.90	0.85

*The all dataset was partitioned according to the G + C% content of the sequences. The number in parenthesis in the header of each column represents the number of sequences belonging to each partition. For each program, AC and (ESn+ESp)/2 are averaged over all sequences belonging to the particular partition for which they are defined

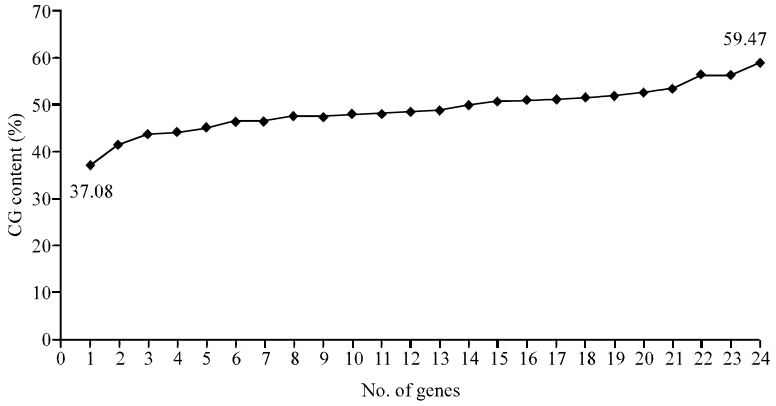

Fig. 2: Distribution of GC content of 24 mouse genes applying for this study

(Bernardi *et al*., 1985), the mouse genome sequencing project demonstrated that gene distribution in both mouse and human genomes correlates well with relative rather than absolute GC content. For example, 75-80% of genes of both species reside in the GC-richest half of the genome. Thus, the mouse genome demonstrates the same trends in gene density while being significantly less extreme in GC-content than the human genome (Waterston *et al*., 2002). The mouse genome appears to have fewer CpG islands than the human genome (i.e., 15,000 vs. 23,000) (Waterstonet *et al*., 2002). However, this could be an artifact resulting from the mouse genome having significantly less variability in GC content than the human genome. Thus, if the same parameters are used to scan both genomes a requirement to get comparable results, it is expected that mouse will have fewer CpG islands, since it has fewer segments with extremely high GC content.

Anyway, the distribution of GC content of 24 mouse single genes has been shown in Fig. 2. Of these, 11 genes (46%) have high GC content (= 50%) and the rest 13 genes (54%) have low GC content (<50%). Table 5 presents the programs' accuracy measures on the sequences with different G + C content. In General, for the genes with low GC content, the maximum AC value (0.83) was detected for FGENESH+ but the maximum ESp +ESn /2 was detected for Genscan program. Moreover, for genes with high GC content, the maximum AC value (0.90) and ESp +ESn /2 (0.85), were detected only for FGENESH+. For both low and high GC contents, the minimum AC value and ESp +E Sn /2 were calculated only for GenView 2 program.

Table 6: Pairwise correlation coefficient between six genefinders

	FGENESH+	Genescan	Genemarkhmm	HMMgene	GenView 2	FGENESH
FGENESH+	1.00	0.37	0.70	0.66	0.34	0.42
Genscan	0.37	1.00	0.40	0.36	0.46	0.89
Genemarkhmm	0.70	0.40	1.00	0.51	0.17	0.56
HMMgene	0.66	0.36	0.51	1.00	0.62	0.23
GenView 2	0.34	0.46	0.17	0.62	1.00	0.21
FGENESH	0.42	0.89	0.56	0.23	0.21	1.00

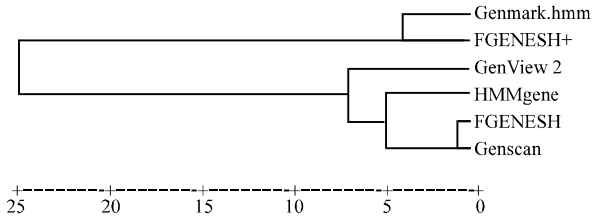

Fig. 3: Dendrogeram for six genefinders was derived from distance matrix using the UPGMA clustering method

Phylogenetic relationships: For each pair of programs, the correlation coefficient CC at the nucleotide level was calculated (Table 6). This is essentially a 6×6 symmetric normalized correlation matrix. Then a normalized distance D between two programs is defined as D = 1-CC. Finally, based on the distance matrix a phylogenetic tree of the programs was generated using UPGMA to show the relations of them (Fig. 3). In general, all programs were clustered in two groups. Interestingly, all *ab initio* genefinders, excepted of GeneMarkhmm., were clustered in the same group and FGENESH+ as well as GeneMarkhmm. program was assigned in another group.

DISCUSSION

In total, our results, as similar as the previous studies, showed that the programs often produce different and sometimes contradictory-results. A question that naturally maybe arisen regards the reliability of any of the gene predictors or of the programs themselves. The answer to this question depends on numerous factors, such as the species under consideration, the sequence context and the existence of experimental evidence (Blanco and Guigo, 2005). In general, a prediction that is supported by spliced ESTs or that shows strong similarity to known coding sequences is more reliable than one having no supporting evidence. Similarity consistent predictions on a given genomic region by different methods should bolster confidence in the prediction (Blanco and Guigo, 2005). However, in the study of Burset and Guigo (1996) the average CC value at nucleotide level for nine programs ranged from 0.65 to 0.80. Later, Rogic *et al.* (2001) published a comparative analysis of seven gene prediction programs, using a set of 195 single-gene sequences of human and rodent species. In their study, the lowest CC (0.66) was detected for MZEF and the highest CC (0.91) was detected for both Genscan and HMMgene programs. In the study of Guigo and Wiehe (2003), the maximum CC (0.78) was calculated for sequence similarity base gene predictions (i.e., ENSEMBL) and the minimum CC was detected for *ab initio* gene finders (i.e., GENSCAN). In this study, the average CC varied from 0.64 for FGENES to 0.91 for FGENESH+ programs.

Moreover, based on present results, we determined that the accuracy of each program is strongly depended on G+C content, so that the calculated AC as well as ESp+ESn /2 values for high

G+C content were higher than low G+C content (Table 5). This findings were in agreement with the findings of the some researchers such as Snyder and Stormo (1995), Burset and Guigo (1996) and Rogic *et al.* (2001) so that all of them have shown that gene-finding programs usually perform worse when the G+C content is low. The proposed reasons for this anomaly are that G+C-rich genes have stronger codon bias that makes them easier to identify and that they are more frequent than the genes in A+T-rich isochores. In addition, Guigo and Fickett (1995), illustrated that coding statistics used by gene-finding programs (codon, dicodon and hexamer frequency) are strongly dependent on G+C content. Moreover, the reason why programs perform better for G+C-rich sequences could also be because they are trained on the sequence subset of GenBank, which is biased towards G+C-rich sequences (Rogic *et al.*, 2001). According to Duret *et al.* (1995), genes from G+C-rich isochores are much more frequently sequenced than those from G+C-poor isochors. Rogic *et al.* (2001) have noticed that if a program has only one set of parameters intended to model gene structure (oligonucleotide frequency, length of coding and intergenic region, exon and intron length and number), it will not be able to perform equally well in both A+T- and G+C-rich sequences due to the significant structural differences between genes in these sequences.

Consistent with the observations made in Burset and Guigo (1996) and Rogic *et al.* (2001), it seems that some programs are sensitive to the G+C content of a sequence, performing better when the sequence is G+C-rich. Here, the programs which exhibited the case, were FGENESH+ on the both levels, Genscan on the nucleotide levels and finally FGENESH on the exon level. In the study of Rogic *et al.* (2001), the programs which exhibited this, were FGENES on the nucleotide level, GeneMark.hmm and Genie on the both levels and HMMgene marginally on the exon level. Among programs that are known to use different parameter sets for different G+C content, the prediction accuracy of Genscan and FGENESH+ are relatively independent of the base composition, but Genview 2 still has very variable results, especially on the exon level, that are not proportional to the G+C content of a sequence.

In total, based on the data of Table 5, the following results are presumed: Firstly, all of the programs, at both levels, have the lowest accuracy measures averaged on the sequences with G+C content less than 50%. Secondly, the mean of calculated accuracy of nucleotide level for the all programs, is more than exon level. It is suggested that seemingly, for genes with low GC content, Genescan, especially for exon level has more potential than the other ones, so that in the study of Rogic *et al.* (2001), the highest AC and ESp+ESn/2 were found out only for Genscan program. On the other hand, considering of both similarity matrix and phylogenetic tree, it can be suggested that although the prediction mechanisms of *ab initio* programs is somewhat different from each other (i.e., searching by signal or searching by content), Nevertheless, these programs were placed in the same cluster and it shows that probability they are more different than FGENESH+ program.

In addition, for predicting the structure of coding genes, we point out that the programs which involve sequence similarity, such as FGENESH+, probably has the most potential than others provided that the known coding sequences (i.g., known proteins for both FGENESH+ Genomescan and/or cDNA for FGENESH_C) of the other relative organisms be available. Otherwise, the other programs based on *ab initio* method can be potentially preferred. These results were in agreement with the study of Guigo and Wiehe (2003). In this study which the accuracy of a number of *ab initio* and comparative gene finding programs on human chromosome 22 were analyzed, the maximum CC (0.78) was calculated for sequence similarity based gene predictors (i.e., ENSEMBL) and the minimum CC was detected for *ab initio* gene finders (i.e., GENSCAN). Generally, three noteworthy issues were mentioned: First, the accuracy of *ab initio* programs substantially suffers

when moving up in complexity from single gene sequences to genome-scale sequence data. Note that the CC of GENSCAN drops from 0.91 (Rogic *et al*., 2001) to 0.64 when applied to chromosome 22. Second, dual-genome comparative gene finders, such as *SGP2*, provide a level of improvement over their *ab initio* counterparts. Third, even the more sophisticated gene finders that use known cDNA or RefSeq genes to improve their predictions still fall short of the level needed for the automatic annotation of complex eukaryotic genomes (Guigo and Wiehe, 2003).

Taken together, according to our present findings, it can be suggested that among all programs based on *ab initio* the programs such as FGENESH and Genscan is more efficient to predict the structure of protein coding genes. Indeed, our results showed that the predictive accuracy of the programs analyzed was lower than originally found and suggested that it is essential to improve and arise of the efficiency of these programs to predict gene structures. Nevertheless, with the completion of sequencing of eukaryotic genomes, gene prediction has changed substantially, particularly from the user's standpoint. Users may not need to even run gene prediction programs on sequences from completed genomes, because many genome browsers already contain this information (Blanco and Guigo, 2005). Gene prediction, however, can still be useful, even in completed genomes, because the user may wish to use different parameters, for example, to analyze alternative splicing or to analyze regions apparently devoid of genes. Programs that predict splice signals and suboptimal exons such as GeneID, Genscan and AUGUSTUS (Stanke *et al*., 2006) are particularly useful here (Blanco and Guigo, 2005). Another potential use of gene finders is to build on previous annotations specifically targeting regions that are apparently devoid of genes but where experimental evidence may suggest the presence if a protein-coding region. Programs like GeneID allow users to provide such information, which is used in making the final prediction (Blanco *et al*., 2002).

CONCLUSION

Present results, similar to the previous studies, demonstrate that even programs based on similar approaches often produce significantly different results. For instance, the maximum predicted Terminal exon was detected for both HMMgene and also Genscan programs, whereas, in case of Initial exon, only FGENESH+ had the highest record. Finally, for internal exons, the program such as FGENESH+ as well as Genscan, had the best predictions. Hence, introducing only one program as the best genefinder usually is not possible. Nevertheless, we point out that the programs such as FGENESH+, for predicting the structure of genes, are presumably useful than the others, on condition that the protein sequence similar with protein which is encoded by the gene be available. Otherwise, the other programs based on *ab initio* method can be potentially preferred. Among the latter programs, probably, both programs Genscan and FGENESH are much more better than the others. At last, apart from obvious differences about calculating of both AC and ESp +ESn /2 parameters in the study (Table 5), it was determined that the accuracy of each program is strongly depended on G+C content, so that for the genes with high G+C content the value of the two latter parameters were higher than low G+C content. As a result, it demonstrates that G+C content, for designing programs to predict the structure of protein coding genes, is inevitable case and should be considered as an important signal.

REFERENCES

Abril, J.F. and R. Guigo, 2000. gff2ps: Visualizing genomic annotations. Bioinformatics, 16: 743-744.

Bernardi, G., B. Olofsson, J. Filipski, M. Zerial and J. Salinas *et al.*, 1985. The mosaic genome of warm-blooded vertebrates. Science, 228: 953-958.

Blanco, E., G. Parra and R. Guigo, 2002. Using Geneid to Identify Genes. In: Current Protocols in Bioinformatics, Baxevanis, A., (Ed.). John Wiley and Sons Inc., New York.

Blanco, E. and R. Guigo, 2005. Predictive Methods Using DNA Sequences. In: Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins, Andreas, D.B. and B.F. Ouellette (Eds.). Wiley, John and Sons Inc., New York, ISBN-13: 9780471478782, pp: 116-142.

Burge, C. and S. Karlin, 1997. Prediction of complete gene structures in human Genomic DNA. J. Mol. Biol., 268: 78-94.

Burset, M. and R. Guigo, 1996. Evaluation of gene structure prediction programs. Genomics, 34: 353-367.

Duret, L., D. Mouchiroud and C. Gautier, 1995. Statistical analysis of vertebrate sequences reveals that long genes are scarce in GC-rich isochores. J. Mol. Evol., 40: 308-317.

Fickett, J.W., 1996. The gene identification problem: An overview for developers. Comput. Chem., 20: 103-118.

Galtiera, N., G. Piganeaub, D. Mouchiroudb and L. Duretb, 2001. GC-content evolution in mammalian genomes: The biased gene conversion hypothesis. Genetics, 159: 907-911.

Guigo, R., S. Knudsen, N. Drake and T.F. Smith, 1992. Prediction of gene structure. J. Mol. Biol., 226: 141-157.

Guigo, R. and J.W. Fickett, 1995. Distinctive sequence features in protein coding, genic non-coding and intergenic human DNA. J. Mol. Biol., 253: 51-60.

Guigo, R. and T. Wiehe, 2003. Gene Prediction Accuracy in Large DNA Sequences. In: Frontiers in Computational Genomics, Galperin, M.Y. and E.V. Koonin, (Eds.). Caister Academic Press, United Kingdom, pp: 1-33.

Huang, X. and W. Miller, 1991. A time-efficient, linear-space local similarity algorithm. Adv. Appl. Math., 12: 337-357.

Hubbard, T., D. Andrews, M. Caccamo, G. Cameron and Y. Chen *et al.*, 2005. Ensembl 2005. Nucl. Acids Res., 33: D447-D453.

Jabbari, K. and G. Bernardi, 1998. CpG doublets, CpG islands and Alu repeat elements in long human DNA sequences from different isochores families. Gene, 224: 123-128.

Krogh, A., 1997. Two methods for improving performace of a HMM and their application for gene finding. Proc. Int. Conf. Intell. Syst. Mol. Biol., 5: 179-186.

Lopez, R.S., F. Larsen and H. Prydz, 1994. Evaluation of the exons predictions of the GRAIL software. Genomics, 24: 133-136.

Lukashin, A.V. and M. Borodovsky, 1998. GeneMark.hmm: New solutions for gene-finding. Nucl. Acids Res., 26: 1107-1115.

Milanesi, L., N.A. Kolchanov, I.B. Rogozin, I.V. Ischenko and A.E. Kel *et al.*, 1993. Genview: A computing tool for protein-coding regions prediction in nucleotide sequences. Proceedings of the 2nd International Conference on Bioinformatics, Supercomputing and Complex Genome Analysis, (BSCGA' 93), World Scienti®c Publishing, Singapore, pp: 573-588.

Mouchiroud, D., G. D'onofrio, B. Aissani, G. Macaya and C. Gautier *et al.*, 1991. The distribution of genes in the human genome. Gene, 100: 181-187.

Parra, G., P. Agarwal, J.F. Abril, T. Wiehe, J.W. Fickett and R. Guigo, 2003. Comparative gene prediction in human and mouse. Genome Res., 13: 108-117.

Pruitt, K.D., T. Tatusova and D.R. Maglott, 2005. NCBI reference sequence (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. Nucl. Acids Res., 33: D501-D504.

Rogic, S., M.K. Alan and B.F.O. Francis, 2001. Evaluation of gene-finding programs on mammalian sequences. Genome Res., 11: 817-832.

Salamov, A. and V.V. Solovyev, 2000. Ab initio gene finding in Drosophila genomic DNA. Genome Res., 10: 516-522.

Snyder, E.E. and G.D. Stormo, 1995. Identification of protein coding regions in genomic DNA. J. Mol. Biol., 248: 1-18.

Solovyev, V.V., A.A. Salamov and C.B. Lawrence, 1994. Predicting internal exons by oligonucleotide composition and discrimi nant analysis of spliceable open reading frames. Nucl. Acids Res., 22: 5156-5163.

Stanke, M., R. Steinkamp, S. Waack and B. Morgenstern, 2004. AUGUSTUS: A web server for gene finding in eukaryotes. Nucl. Acids Res., 32: 309-312.

Stanke, M., O. Keller, I. Gunduz, A. Hayes, S. Waack and B. Morgenstern, 2006. AUGUSTUS: Ab initio prediction of alternative transcripts. Nucl. Acids Res., 34: 435-439.

Waterston, R.H., K. Lindblad-Toh, E. Birney, J. Rogers and J.F. Abril *et al.*, 2002. Initial sequencing and comparative analysis of the mouse genome. Nature, 420: 520-562.

Yao, H., L. Guo, Y. Fu, L. Borsuk and T.J. Wen *et al.*, 2005. Evaluation of five ab initio gene prediction programs for the discovery of maize genes. Plant Mol. Bio., 57: 445-460.

Zoubak, S., O. Clay and G. Bernardi, 1996. The gene distribution of the human genome. Gene, 174: 95-102.