African Journal of Biotechnology Vol. 11(7), pp. 1545-1558, 24 January, 2012
Available online at http://www.academicjournals.org/AJB
DOI: 10.5897/AJB11.2316
ISSN 1684–5315 © 2012 Academic Journals

*Full Length Research Paper*

# Analysis of some conventional *ab initio* gene finders using human and mouse DNA sequences

**Jaber Nasiri[1*#], Aboubakr Moradi[2#], Somayeh Ahangarian Abhari[3] and Ali Nasiri[4]**

[1]Islamic Azad University of Iran, Khorasgan Branch, Division of Plant Genetics Research, Khorasgan, Isfahan, Iran.
[2]Department of Plant Biotechnology, College of Agriculture, University of Zanjan, Q1 Zanjan P.O. Box 313, I.R. Iran.
[3]Department of Genetics and Plant Breeding, College of Agriculture, Zanjan University, Zanjan- Iran.
[4]Department of Health, Treatment and Medical education, Isfahan University of Medical Science, Isfahan, Iran.

Accepted 4 November, 2011

**An evaluation of the prediction accuracy of five *ab initio* gene prediction programs (that is, FGENESH, Genscan, HMMgene, GeneMark.hmm and FGENES) was conducted by the use of 110 human and mouse orthologous sequences. As expected, all programs presented different predictions with various ranges of accuracy. According to our results, FGENESH and Genscan generally had the maximum power to produce more reliable results in both nucleotide and exon levels than others. Although, both FGENES and GenMark.hmm predicted the highest number of exons (966 and 946 exons, respectively), when exon sensitivity (*ESn*), exon specificity (*ESp*) and (*ESn*+*ESp*/2) were considered, their overall accurate performance descended and was clustered in the lowest positions. It was also determined that all programs have lower power in predicting initial and terminal exons, as compared to internal exons, which suggested that such programs cannot accurately determine translational start sites (TSS) and translational stop codons (TSC) as internal exons, whose boundaries are highlighted by acceptor and donor sites. Apart from the species difference, it was finally recognized that the programs, FGENESH and GeneMark.hmm, presented much more sensitivity in detecting genes with low guanine-cytosine (GC) content.**

**Key words:** *Ab initio* gene prediction programs, human, mouse, orthologous sequences.

## INTRODUCTION

Over the last 20 years, despite some difficulties observed in discovering eukaryotic genes which clearly result from low gene density as well as large spacers found between adjacent genes (Taher et al., 2004; Wang et al., 2004; Stanke et al., 2004; Do and Choi, 2005; Irimia et al., 2009), there has been a great explosion in genomic sequence data with plentiful genomes of both eukaryotes and prokaryotes in different phases of sequencing and annotation. In fact, eukaryotic genomes are being sequenced at an ever-increasing rate (Do and Choi, 2005; Abeel et al., 2008; Schweikert et al., 2009) and nearly 180 complete genomes of both eukaryotes and prokaryotes are publicly available (http://en.wikipedia.org/wiki/List_of_sequenced_eukaryotic_genomes).

Annotation gene structures are therefore invariably the first step after the completion of the genome DNA sequence (Harrow et al., 2009; Schweikert et al., 2009). In view of that, developing quick, reliable and accurate methods for the prediction and annotation of gene structure is essential. However, two basic appro-aches have been generally developed for computational gene-finding: intrinsic and extrinsic. Intrinsic (*ab initio* or *de novo*) methods deal strictly with DNA sequences and extract information regarding gene locations using statistical patterns inside and outside of gene regions, as well as those patterns typical of gene boundaries. They are actually the programs of choice in the absence of known transcript or protein sequences, or phylogenetically related genomes (Harrow et al., 2009).

Pioneering studies using intrinsic statistical approaches were conducted in the early 1980s (Fickett, 1982; Gribskov et al., 1984; Staden, 1984). Since then, a com-

*Corresponding author. E-mail: nassiri.j@gmail.com. Tel: +989133220399. Fax: +983115354060.

#These authors contributed equally to this work.

prehensive comparative analysis of a number of gene structure prediction programs was performed by the use of vertebrate genomic DNA (Burset and Guigo, 1996). Five years later, another study in this case was accomplished and a significant improvement in developing new gene finders was reported. Even though it was the same as that of previous studies, the accuracy of the seven used programs was systematically lower than that of those originally found (Rogic et al., 2001). To increase the performance ability of gene finders, a number of researches were then proposed by the use of different algorithms. In each one, some positive features were mentioned and reviewed by Mathe et al. (2002), Vladimir (2002), Wang et al. (2004), Brent and Guigo (2004), Do and Choi (2006), Zhu et al. (2007) and Harrow et al. (2009). Recently, in the study of Kwan et al. (2009), two *ab initio* gene prediction programs (GeneMark.hmm-ES 3.0 and GreenGenie2) were examined using a total of 140 experimental sequences of *Chlamydomonas reinhardtii*. In all gene, exon and nucleotide levels, Green Genie2 had the maximum sensitivity and specificity which, on the first two levels, were statistically significant. Nonetheless, the value of specificity and sensitivity was somehow in agreement with our results, which suggested that we still need more accurate programs to verify our current genomic sequence data. In addition, another novel gene prediction algorithm called Multivariate Entropy Distance (MED) was developed to improve and facilitate the comparative studies of prokaryotic genomes (Zhu et al., 2009). The program MED 2.0, as compared to the five current best prokaryotic gene finders could achieve a competitive high performance in gene prediction for both 5' and 3' end matches. On the other hand, if related genome sequences are available, the intrinsic information can be combined with patterns of genomic sequence conservation using programs often referred to as comparative (or dual- or multi-genome) gene finders. With these programs, maximum resolution is achieved when the compared genomes are at a phylogenetic distance such that there is maximum separation between conservation in coding and non-coding regions. These approaches, nevertheless, are highly dependent upon the quantity and quality of pre-existing sequence data (Hong Yao et al., 2005). Although, some investigations have indicated that the prediction accuracy is based on these programs, this method is more reliable than *ab initio* based programs with no employed-similarity (Salamov and Solovyev, 2000; Guigo and Wiehe, 2003; Flicek et al., 2003; Parra et al., 2003; Knapp and Chen, 2006; Nasiri et al., 2011). Since the genomes of many organisms are yet to be sequenced entirely, *ab initio* gene prediction programs are still important annotation tools and the evaluation of these programs could be necessary for their improvement (Zhang, 2002; Lomsadze et al., 2005; Li et al., 2005; Stanke et al., 2008; Nasiri et al., 2011). Likewise, they are not very useful when the expected homology between the gene searched for and the known sequences is low. Lastly, these software cannot detect possible changes in nucleotide sequences due to RNA editing mechanisms (Fassetti et al., 2010).

Even though quite a number of studies have been performed either to introduce a new gene finder with employing a novel algorithm or to evaluate the ability of various gene finding programs (Fickett, 1996; Stormo, 2000; Zhang, 2002; Mathe et al., 2002; Vladimir, 2002; Wang et al., 2004; Brent and Guigo, 2004; Do and Choi, 2006; Zhu et al., 2007; Harrow et al., 2009; Kwan et al., 2009; Liang et al., 2009; Fassetti et al., 2010) regarding higher eukaryotic organisms, the accuracy of the current available gene prediction programs using ortologous genes is limited (Flicek et al., 2003; Parra et al., 2003; Knapp and Chen, 2006). However, these three research groups recommended TWINSCAN, SGP-2 and TWINSCAN programs as the most reliable programs, respectively. At present, it is noticeable that although, there are some new programs including mSplicer (Ratsch et al., 2007), Craig (Bernal et al., 2007), Conrad (DeCaprio et al., 2007) and Contrast (Gross and Brent, 2006), there is no easy-to-use web application available. To employ these tools, the respective packages have to be downloaded and installed, which in some cases requires substantial programming knowledge as well as the accessibility of sufficient computational power for each user. In contrast, the conventional *ab initio* gene finders are not only available as online and easy-to-use, but also majority of them according to the previous and current studies seem to be able to generate prediction with acceptable accuracy.

In this study, an effort was made accordingly to assess the performance ability of five conventional *de novo* gene prediction programs on account of predicting different parts of a given protein coding sequence so that the users be able to choose the best program(s) in accordance with their research goals.

## MATERIALS AND METHODS

### Sequence data set

In assessing five *ab initio* gene prediction programs, a data set consisting of 110 known orthologous genes of human and mouse were employed. This data collection, in both organisms, consisted of three genes with no introns in the open reading frame (commonly referred to as a 'single exon gene') and the rest were multi-exon genes. The number exons per gene vary from two to 30 with an average number of 8.37 for human and 8.50 for mouse. Likewise, in both genomes, around 927 coding exons (totally, 152488 bp) with a mean length of 164.5 base pairs were detected in a real experimental data. As the last point, our data is composed of 1,224,136 nucleotides (nt) over 110 sequences with a mean sequence length of 11,128.5 bases.

### Programs tested

The research was conducted to realize the potential of five *ab initio* gene finding programs, that is, FGENESH (Salamov and Solovyev, 2000), Genscan (Burge and Karlin, 1997), HMMgene (Krogh,

**Table 1.** List of the five *ab inito* gene prediction programs used for this study.

| Program | Organism | Employed algorithm | Available at |
|---|---|---|---|
| *FGENES* | Human | DP | http://genomic.sanger.ac.uk/gf/ |
| *FGENESH* | Human, mouse, *Drosophila*, rice, , other | HMM | http://www.softberry.com/berry.phtml?topic=gfind-file |
| *Genscan* | Vertebrates, *Arabidopsis* and maize | GHMM | http://genes.mit.edu/GENSCAN.html |
| *Genemark.hmm* | Human, mouse, *Drosophila*, other | GHMM | http://opal.biology.gatech.edu/GeneMark/eukhmm.ci |
| *HMMgene* | Vertebrates and *C. elegans* | GHMM | http://l25.itba.mi.cnr.it/~webgene/wwwgene.html |

DP, Dynamic programming; HMM, hidden Markov model; GHMM, generalized HMM.

1997), GeneMark. hmm (Lukashin and Borodovsky, 1998) and FGENES (V. Solovyev, unpublished data) (Table 1).

**Accuracy measurement**

Prediction accuracy of all five *ab initio* programs was measured at two different levels: coding nucleotide sequence and exonic structure. Furthermore, we examined precision based on guanine-cytosine (GC) content. Note that the exons predicted on the forward strand containing known genic sequences were only analyzed (predictions for the reverse strand were not considered, because all prediction results were compared with the known actual gene structures in our test data sets, all of which were identified in forward strand in the NCBI) and compared to the actual coding exons.

**Nucleotide level statistics**

**Consistent with Burset and Guigo (1996), the following four metrics were calculated:**

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.

As the second step, both the nucleotide sensitivity *Sn*, (that is, the proportion of coding nucleotides that are correctly predicted as coding) and nucleotide specificity *Sp*, (that is, the proportion of nucleotides predicted as coding that are actually coding) values were estimated using the following formulas:

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

It has been demonstrated that high *Sn* can be achieved with little *Sp* and vice versa (Burset and Guigo, 1996). Accordingly, an additional parameter was defined called *correlation coefficient* (*CC*), reflecting both *Sn* and *Sp*. *Correlation coefficient* is actually defined as:

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

In order to assess the global performance of any program, approximate correlation (*AC*) was also defined (Burset and Guigo, 1996);

$$AC = (ACP - 0.5) \times 2$$

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

**Exon level statistics**

It has been demonstrated that the prediction precision at the exon level is important when designing primers or probes (Li et al., 2005). In this regard, exon specificity (*ESp*) is defined as the proportion of exons that are actually coded, whereas, exon sensitivity (*ESn*) is the proportion of actual exons in the test sequence that is correctly predicted (Burset and Guigo, 1996):

$$\mathrm{ESp} = \frac{\mathrm{TE}}{\mathrm{PE}}, \quad \mathrm{ESn} = \frac{\mathrm{TE}}{\mathrm{AE}}$$

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.

In general, during a prediction process through related programs, annotated exons can be divided into exons that are exactly predicted, partially predicted, overlapped, missed or wrong (not overlapped with any predicted exon), while an actual exon is counted as a *missing exon*, and if it does not have a single base predicted, the term *wrong exon* is applied when no single predicted base is present in the actual exons. The following formulas have made the measurement of both items possible:

$$ME = \left(\frac{\text{No. of missing exons}}{\text{No. of actual exons}}\right)$$

$$WE = \left(\frac{\text{No. of wrong exons}}{\text{No. of predicted exons}}\right)$$

Finally, to explore the rate of the performance of each program in predicting various exon classes, all exons were divided into four classes: 5′ exons (or initial exon), internal exons, 3′ exons (or terminal exon) and intronless exons (or, simply, intronless genes) and further subdivided into 13 subclasses, according to their coding content (Zhang, 2002).

**Table 2.** The relative nucleotide and exon level precision of the five *Ab initio* gene finding programs.

| Parameter | Nucleotide level | | | | Exon level | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Sn | Sp | CC | AC | ESn | ESp | (ESn+ESp)/2 | ME% | WE% |
| **Human** | | | | | | | | | |
| *FGENES* | 0.95 | 0.93 | 0.92 | 0.93 | 0.80 | 0.84 | 0.86 | 6.00 | 9.90 |
| *FGENESH* | 0.95 | 0.93 | 0.93 | 0.93 | 0.85 | 0.86 | 0.84 | 6.90 | 6.70 |
| *Genscan* | 0.95 | 0.90 | 0.91 | 0.91 | 0.80 | 0.79 | 0.75 | 5.50 | 9.00 |
| *Genemark.hmm* | 0.89 | 0.92 | 0.89 | 0.89 | 0.73 | 0.75 | 0.74 | 11.50 | 11.10 |
| *HMMgene* | 0.92 | 0.93 | 0.91 | 0.91 | 0.81 | 0.83 | 0.82 | 8.80 | 6.30 |
| **Mouse** | | | | | | | | | |
| *FGENES* | 0.90 | 0.88 | 0.87 | 0.87 | 0.80 | 0.76 | 0.78 | 7.70 | 12.9 |
| *FGENESH* | 0.98 | 0.93 | 0.94 | 0.95 | 0.90 | 0.88 | 0.88 | 3.80 | 6.00 |
| *Genscan* | 0.98 | 0.91 | 0.92 | 0.93 | 0.83 | 0.81 | 0.79 | 5.10 | 9.50 |
| *Genemark.hmm* | 0.95 | 0.88 | 0.89 | 0.90 | 0.80 | 0.74 | 0.75 | 8.00 | 14.60 |
| *HMMgene* | 0.92 | 0.96 | 0.93 | 0.93 | 0.84 | 0.87 | 0.85 | 6.90 | 4.10 |
| **Whole data** | | | | | | | | | |
| *FGENES* | 0.93 | 0.90 | 0.90 | 0.90 | 0.80 | 0.80 | 0.82 | 6.90 | 11.40 |
| *FGENESH* | 0.96 | 0.93 | 0.930 | 0.94 | 0.88 | 0.87 | 0.86 | 5.40 | 6.35 |
| *Genscan* | 0.96 | 0.90 | 0.92 | 0.92 | 0.82 | 0.80 | 0.77 | 5.30 | 9.25 |
| *Genemark.hmm* | 0.92 | 0.90 | 0.89 | 0.89 | 0.77 | 0.75 | 0.75 | 9.80 | 12.90 |
| *HMMgene* | 0.92 | 0.94 | 0.92 | 0.92 | 0.83 | 0.85 | 0.83 | 7.90 | 5.20 |

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

## RESULTS

All sequences were analyzed using each program. To verify the correct annotation, only results from the positive strands were considered and observed for a total of 550 predictions. Each predicted sequence was then compared to its coding sequences (CDS) annotation of GenBank entry. The prediction accuracy at both nucleotide and exon levels are shown in Table 2.

### Nucleotide level precision

At the nucleotide level, Genscan had the highest sensitivity (0.96), while GeneMark.hmm had the least (0.89). Further, both FGENESH and HMMgene, with the specificity (0.93) were determined as the most accurate programs, whereas the minimum *specificity* was detected only for Genscan (0.90). Surprisingly, when the values of *CC* and *AC* were calculated, the maximum *CC* (0.93) and *AC* (0.93) values were detected only for FGENESH program, whereas the reverse was true for GeneMark.hmm, with the lowest seen for *CC* (0.88) and *AC* (0.89), suggesting that the prediction of FGENESH in the case of predicting human genes could be more helpful, at least, when compared with the other four programs. Regarding mouse genome, FGNENS was the feeblest program, but unexpectedly in both *Sn* and *Sp,* it was 0.90 and 0.87, respectively. On the other hand, FGENESH and HMMgene, in the same order, experienced the greatest increase in both *Sn* (0.98) and *Sp* (0.96). Nevertheless, when the following parameters, *CC* and *AC*, were calculated, it was observed that although, similar to the previous situations, FGENESH was the leader with 0.94 and 0.95, respectively, the lowest, *CC* (0.87) and *AC* (0.87), were identified only for FGENES. We notified that apart from some differences among these programs, such observations, in particular regarding Genscan and HMMgene, apparently are not statistically significant, suggesting that users can use them as alternative programs. In order to confirm our results, we also constructed a data collection, known as *whole data*, and as could be seen, FGENESH had more value of *AC* (0.94) and *CC* (0.93) than GenMark.hmm, which again emerged as the weakest program.

### Exon level accuracy

In this situation, concerning human sequences, the descending order of the programs based on the values of (*ESn* + *ESp*)/2 were: FGENES (0.86), FGENESH (0.84), HMMgene (0.82), Genscan (0.75) and GeneMark.hmm (0.74). In addition, FGENES together with FGENESH,

**Table 3.** Predicted number of exons in each class on multi-exon genes in three different data. The data given in the table are the TE/PE.*

| Class | Human | | | | Mouse | | | | Whole data | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Initial | Internal | Terminal | Total | Initial | Internal | Terminal | Total | Initial | Internal | Terminal | Total |
| AE | 53 | 354 | 53 | 460 | 54 | 358 | 54 | 466 | 107 | 712 | 107 | 926 |
| *FGENES* | 38/57 | 317/363 | 45/57 | 400/477 | 44/61 | 298/368 | 32/60 | 374/489 | 82/118 | 615/731 | 77/117 | 774/966 |
| *FGENESH* | 36/48 | 323/360 | 37/44 | 433/452 | 37/50 | 321/374 | 45/52 | 403/476 | 73/98 | 644/734 | 82/96 | 836/928 |
| *Genscan* | 30/45 | 321/386 | 37/45 | 388/476 | 32/46 | 335/392 | 41/50 | 408/488 | 62/91 | 656/778 | 78/95 | 796/964 |
| *Genemark.hmm* | 15/31 | 304/356 | 28/37 | 347/424 | 20/40 | 323/434 | 38/51 | 381/525 | 35/71 | 627/790 | 66/88 | 728/949 |
| *HMMgene* | 35/48 | 292/345 | 40/49 | 367/442 | 37/48 | 295/350 | 44/47 | 376/445 | 72/96 | 587/690 | 84/96 | 743/887 |

*TE, True exons; AE, actual exons; PE, predicted exons.

and only GeneMark.hmm had the lowest and highest percentage of the missing and wrong exons. Regarding mouse data, FGENESH again had the best performance in terms of *ESn*, *ESp* and *CC*. Although, both GeneMark.hmm and FGENES (as the weakest programs) had equal *ESn* (0.80), the efficiency of the second one moved up a little more when its *ESp* and *CC* were taken into account. Furthermore, the average proportion of the missing exons (*ME*) was 3.8 and 8.0% for FGENESH and Genemark.hmm, respectively, which was lower than that of humans. While regarding the second item as wrong exons (*WE*), a three-fold growth from 4 to 13% could be observed, nearly the same as that of humans. Therefore, these programs seem to have more power in predicting many more exons correctly when they are applied to mouse sequences.

Eventually, when all sequences were considered, programs such as FGENESH and HMMgene were identified as the first and second top programs and again GeneMark.hmm emerged as a program with the lowest prediction accuracy of the exons boundaries. Regarding *ME* and *WE*, GeneMark.hmm had the maximum values of the *ME* and *WE* with 9.80 and 12.90%, respectively.

## Recognition power of programs in distinguishing various exon classes

Briefly, to predict the number of initial exons precisely, the programs such as FGENESH and FGENES had more potential when they were loaded by human and mouse sequence data, while both HMMgene and FGENESH were detected as the best programs when total data was employed (Table 3). On the other hand, GeneMark.hmm had the lowest accuracy in this position for all the three mentioned categories. At the second exon class (internal exons), Genscan was concluded as the best program concerning mouse sequences, whereas FGENESH showed the greatest increase when both human and whole sequence data were taken into account. In addition, though FGENES program for anticipating terminal exon can be a reliable source, its potential is dubious as the mouse and whole data is taken into account. Since plenty of gene prediction tools are now available freely, it is accordingly advisable to utilize other powerful programs instead of FGENES, such as Genscan, HMMgene or FGENESH in acquiring more reliable results at least here.

## GC content

GC-rich regions include many genes with short introns, while GC-poor regions are essentially deserts of genes (Galtier et al., 2001). Moreover, it has been suggested that the distribution of GC content in mammals could have some functional relevance related to genes (Mouchiroud et al., 1991; Duret et al., 1995; Jabari and Bernardi, 1998; Galtiera et al., 2001).

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. Instead, the mouse genome appears to have fewer CpG islands than the human genome (that is, 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 the 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. In the human genome, 2.7% of the

**Table 4.** Sensitivity and specificity of predictions for various classes of exons in three different genomic data.

| Programs | Initial | | Internal | | Terminal | | Total | |
|---|---|---|---|---|---|---|---|---|
| | ESn | ESp | ESn | ESp | ESn | ESp | ESn | ESp |
| **Human** | | | | | | | | |
| FGENES | 0.72 | 0.66 | 0.9 | 0.87 | 0.85 | 0.79 | 0.82 | 0.77 |
| FGENESH | 0.68 | 0.75 | 0.92 | 0.9 | 0.7 | 0.84 | 0.77 | 0.83 |
| Genscan | 0.57 | 0.67 | 0.91 | 0.83 | 0.7 | 0.82 | 0.73 | 0.77 |
| Genemark.hmm | 0.28 | 0.5 | 0.86 | 0.85 | 0.53 | 0.76 | 0.56 | 0.70 |
| HMMgene | 0.67 | 0.73 | 0.83 | 0.84 | 0.75 | 0.81 | 0.75 | 0.79 |
| **Mouse** | | | | | | | | |
| FGENES | 0.81 | 0.72 | 0.83 | 0.81 | 0.59 | 0.53 | 0.74 | 0.69 |
| FGENESH | 0.68 | 0.74 | 0.90 | 0.86 | 0.83 | 0.87 | 0.80 | 0.82 |
| Genscan | 0.59 | 0.7 | 0.94 | 0.85 | 0.76 | 0.82 | 0.76 | 0.79 |
| Genemark.hmm | 0.37 | 0.5 | 0.9 | 0.74 | 0.7 | 0.75 | 0.66 | 0.66 |
| HMMgene | 0.68 | 0.77 | 0.82 | 0.84 | 0.81 | 0.93 | 0.77 | 0.85 |
| **Whole data** | | | | | | | | |
| FGENES | 0.77 | 0.69 | 0.87 | 0.84 | 0.72 | 0.66 | 0.79 | 0.73 |
| FGENESH | 0.68 | 0.75 | 0.91 | 0.88 | 0.76 | 0.86 | 0.78 | 0.83 |
| Genscan | 0.58 | 0.69 | 0.93 | 0.84 | 0.73 | 0.82 | 0.75 | 0.78 |
| Genemark.hmm | 0.33 | 0.50 | 0.88 | 0.80 | 0.62 | 0.76 | 0.61 | 0.69 |
| HMMgene | 0.68 | 0.75 | 0.83 | 0.84 | 0.78 | 0.87 | 0.76 | 0.82 |

20 kb segments have GC content greater than 56% or less than 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 in other vertebrates (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 to 80% of the 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 it is significantly less extreme in the GC-content than the human genome (Waterston et al., 2002).Basically, *ab initio* gene prediction methods rely on two types of sequence information: searching by signal and searching by content (Wang et al., 2004; Blanco and Guigo, 2005). In order to discriminate protein-coding regions from non-coding regions, a number of content-based measures which are also known as *coding statistics* can be used (Ficket and Tung, 1992; Gelfand, 1995; Guigo, 1999). Among the numerous methods for the computation of content-based measures, hexamer frequency, usually in the form of codon position-dependent fifth-order Markov models (Borodovsky and McIninch, 1993), seems to have maximal discriminative power; and surprisingly, it has been demonstrated that *coding statistics* used by gene-finding programs (codon, dicodon and hexamer frequency) are strongly dependent on GC content (Guigo and Fickett, 1995). Moreover, by a brief look at the previous studies in the case of gene prediction programs, it is obvious that not only the performance of *ab intio* based methods could be affected by GC content, but also other available gene finding programs could be affected as well on the basis of sequence similarity or alignment (Snyder and Stormo, 1995; Burset and Guigo, 1996; Rogic et al., 2001; Yao et al., 2005; Li et al., 2005). This has largely been due to the fact that GC-rich regions include a large number of genes with short introns, whereas GC-poor regions are essentially deserts of genes (Xu et al., 1994; Lopez et al., 1994; Snyder and Stormo, 1995; Burset and Guigo, 1996; Rogic et al., 2001). Anyway, the question is: to what extent could such parameters be significant and which programs are more sensitive against this parameter?

In order to assess these issues, all employed genes in accordance with their GC content were divided into three parts: lower than 47% (27% of all sequences), between 48 and 52% (35%) and finally higher than 53%, containing 45 accessions (48%). The GC content of the both genomes varied from 34 to 65%. Table 4 presents the programs' accuracy measures on the sequences with different GC contents. Consistent with the observations made in Burset and Guigo (1996) and Rogic et al. (2001), some programs were sensitive to the GC content of asequence, and performed better when the sequence is GC-rich. The programs that exhibited this trend were FGENESH and GeneMark.hmm on both levels, and HMMgene on the exon level. Among programs that are

known to use different parameter sets for different GC content, Genscan and FGENES's prediction accuracy is relatively independent of the base composition.

## DISCUSSION

In this study, a test data set including 110 known orthologous genes of human and mouse were employed in order to examine which conventional de novo gene prediction programs have more power to anticipate different parts of human and mouse protein coding sequences. Unlike previous studies (Burset and Guigo, 1996; Rogic et al., 2001; Yao et al., 2005; Li et al., 2005; Knapp and chen, 2006), in this study, each program had equal chance of being loaded with all the used accessions. Predictably, all programs had the tendency to produce different and occasionally contradictory results.

In all three categories, the preference of FGENESH program, as compared to the other four programs, was irrecusable. Consequently, it could be concluded that FGENESH has enough potential to generate more reliable predictions than the others. This finding was in agreement with the studies of Burge and Karlin (1998), Yao et al. (2005) and Li et al. (2005), all of which reported FGENESH as the best gene prediction program to predict different parts of humans, maize and rice genome sequences, respectively. Instead, at the study of Schweikert et al. (2009), the program mGene was fairly better against programs such as FGENESH, Craig and Augustus in all nucleotide, exon, gene and transcript levels. In the same study, the values of accuracy in nucleotide and exon levels were the first and second maximum amount of validation, while in the case of predicting transcripts and the numbers of genes, these programs appear to require more improvements. Similarly, Nasiri et al. (2011) found that FGENESH+ as compared to FGENESH and other *ab initio* programs could make more reliable results, and again, prediction accuracy at the nucleotide level was the superior. In the current study, all programs, like those of previous studies (Rogic et al., 2001; Schweikert et al., 2009; Nasiri et al., 2011), generated more reliable outputs at the nucleotide versus exon level. To explain this phenomenon, the corresponding formula of both levels should be analyzed. In reality, a little variation in the number of PE, true exon TE and AE can be accompanied by significant differences in the final results (that is, ESp, ESn and CC), while at the nucleotide level, since a large number of nucleotides are examined, the majority is often predicted precisely. Accordingly, each variation can produce slight differences in the final results (that is, Sp, Sn and CC). It is noticeable that in the study of Rogic et al. (2001), both Genscan and HMMgene programs with the highest *CC* (0.91) were marked as the most trustworthy sources, and if FGENESH is ignored, the both programs were located in the first position, suggesting that they can still produce reliable predictions. As could be seen, the programs such as GeneMark.hmm and FGNENS in our ranking were located on the nethermost classes, so their performance is somehow questionable. As a reason, FGENES was originally developed to predict human genes; accordingly, it is advisable to use other available programs in case of anticipating mouse or even human sequences. On the other hand, since orthologs are genes that have a vertical descent relationship from a common ancestor and encode proteins with the same function in different species (Koonin et al., 1996), consequently, just a few rare and insignificant variations could be observed, probably due to different evolutionary agents such as point mutations, insertions, deletions, translocations and/or inversion that have changed the whole structure of gene(s) over a long period of time.

With the exception of some negligible differences, at the exon level, FGENESH with the highest values of (*ESn* + *ESp*)/2 as *CC* and *ESn* appeared again as the most powerful program in all the three classes. On the other hand, GeneMark.hmm had the least significant position, not just because of having the lowest values of *ESn*, *ESp* and *CC*, but also because of its highest percentage of missing and wrong exons. Ignoring some minor discrepancies which are common among different studies, these results are somehow consistent with previous investigations. For instance, in the study of Yao et al. (2005), FGENESH exhibited the maximum level of *ESn*, *ESp* and *CC*, and had correspondently low percentages of both *ME* and *WE*; but contrary to our results in which GeneMark.hmm had the lowest degree of worth, GeneMark.hmm and Genscan were the second and third important programs. Moreover, FGENESH plus BGF, in the study of Li et al. (2005) and all the three programs including FGENES, Genscan and HMMgene in the study of Rogic et al. (2001), were identified as programs with low *ME* and *WE*. In addition, in accordance with Rogic et al. (2001), Li et al. (2005), and to some extent, Burset and Guigo (1996), and Knapp and Chen (2006), for each program, there was no significant variation between the percentage of *ME* and *WE*. Nonetheless, if the report of Yao et al. (2005) is considered, particularly for the case of Genscan and GeneMark.hmm data, a seven- and four-fold trend could be observed, respectively. Interestingly, even though the model parameters of the programs were learnt from the set of human sequences, in some cases, the values for mouse sequences were higher than human sequences. Nonetheless, it looks like these should not be statistically significant and possibly such differences would occur even if the results on two different human sequence sets or any other organism were compared. This hypothesis is also supported by the comparison of the human and mouse grammars constructed by Dong and Searls (1994) and Rogic et al. (2001), and also by different data sets of rice genome (Li et al., 2005) where no significant

**Table 5.** Accuracy versus G + C content measured in three data sets.

| GC content | <48% | | 48-52% | | >52% | |
|---|---|---|---|---|---|---|
| | **AC** | **(ESn+ESp)/2** | **AC** | **(ESn+ESp)/2** | **AC** | **(ESn+ESp)/2** |
| Human | (13) | | (15) | | (27) | |
| *FGENES* | 0.93 | 0.80 | 0.95 | 0.89 | 0.91 | 0.88 |
| *FGENESH* | 0.92 | 0.75 | 0.94 | 0.81 | 0.93 | 0.89 |
| *Genscan* | 0.92 | 0.76 | 0.93 | 0.78 | 0.94 | 0.72 |
| *Genemark.hmm* | 0.92 | 0.69 | 0.85 | 0.70 | 0.9 | 0.79 |
| *HMMgene* | 0.87 | 0.70 | 0.90 | 0.84 | 0.94 | 0.87 |
| Mouse | (14) | | (23) | | (18) | |
| *FGENES* | 0.90 | 0.82 | 0.84 | 0.79 | 0.87 | 0.74 |
| *FGENESH* | 0.93 | 0.91 | 0.96 | 0.89 | 0.96 | 0.82 |
| *Genscan* | 0.93 | 0.75 | 0.95 | 0.87 | 0.93 | 0.73 |
| *Genemark.hmm* | 0.90 | 0.73 | 0.88 | 0.78 | 0.93 | 0.72 |
| *HMMgene* | 0.89 | 0.81 | 0.96 | 0.87 | 0.92 | 0.86 |
| Whole data | (27) | | (38) | | (45) | |
| *FGENES* | 0.92 | 0.80 | 0.88 | 0.84 | 90 | 0.82 |
| *FGENESH* | 0.93 | 0.84 | 0.94 | 0.86 | 0.94 | 0.86 |
| *Genscan* | 0.93 | 0.75 | 0.94 | 0.83 | 0.91 | 0.73 |
| *Genemark.hmm* | 0.89 | 0.71 | 0.91 | 0.75 | 0.90 | 0.76 |
| *HMMgene* | 0.91 | 0.76 | 0.93 | 0.86 | 0.93 | 0.86 |

differences were found.

Having a brief look at the previous studies, for example on vertebrate, Drosophila, vertebrate, maize, rice, and finally human sequences, the programs such as FGENESH, Genscan, Genscan, FGENESH, BGF plus FGENESH, Genzilla plus Genomescan appear to predict internal exons better than initial and terminal exons, respectively (that is, those beginning with start site and ending with a stop codon) (Burge and Karlin, 1998; Salamov and Solovyev, 2000; Rogic et al., 2001; Yao et al., 2005; Li et al., 2005; Knapp and Chen, 2006). This implies that the ability of these programs to detect the correct start and stop codons is probably a little weaker than to identify 5' and 3' splice sites correctly. In this regard, in the current study, FGENESH and HMMgene, as compared to other programs, predicted internal exons much better than the initial and terminal exons (Table 5). Moreover, it is noticeable that although in the study of Yao et al. (2005) on maize, Genscan predicted initial and terminal exons better than it did for internal exons. In this study, such a result was not observed, and it suggested that different programs gave various responses to the species under study or different genes with various genetic features. This demonstrates that the organism under study is one of the most important items in selecting one or more programs to compare their results with laboratory findings obtained from cloning procedures.

It is noted that in such investigations, all measured parameters are often based on average data (for example, average of 110 data for each program). In other words, if all predictions are examined one by one, undoubtedly, a program with high efficiency (for example, FGENESH) may not predict the structure of a number of genes accurately, while a program with lower value of accuracy (e.g., FGENES) can propose a better prediction for the same number of genes. For instance, when accession U78027 containing 7 exons was loaded by FGENESH, the terminal exon was missed, while as the same accession was run by FGENES, all 7 exons were anticipated precisely (Figure 1). Surprisingly, the same as FGENESH, Genscan (as the second best program) could not predict terminal exon at all. As a result, it is advisable to integrate the results of multiple *ab initio* programs as a scientific solution. Polling *ab initio* and sequence similarity based approaches is another way to improve the accuracy of gene prediction, and is likely to be more widely used as the number of sequenced genomes increases (Mathe et al., 2002). Programs such as Twinscan (Korf et al., 2001), SGP2 (Para et al., 2003), SLAM (Cawley et al., 2003), AGenDA (Taher et al., 2004) and Combiner (Allen et al., 2006) can improve the accuracy of gene predictions through these two approaches. Overall, it seems that introducing only one program as the best one for all the current species is impossible and users should consider this issue as a critical item. Or else, when performing a given study, one may encounter lots of severe incompatible and

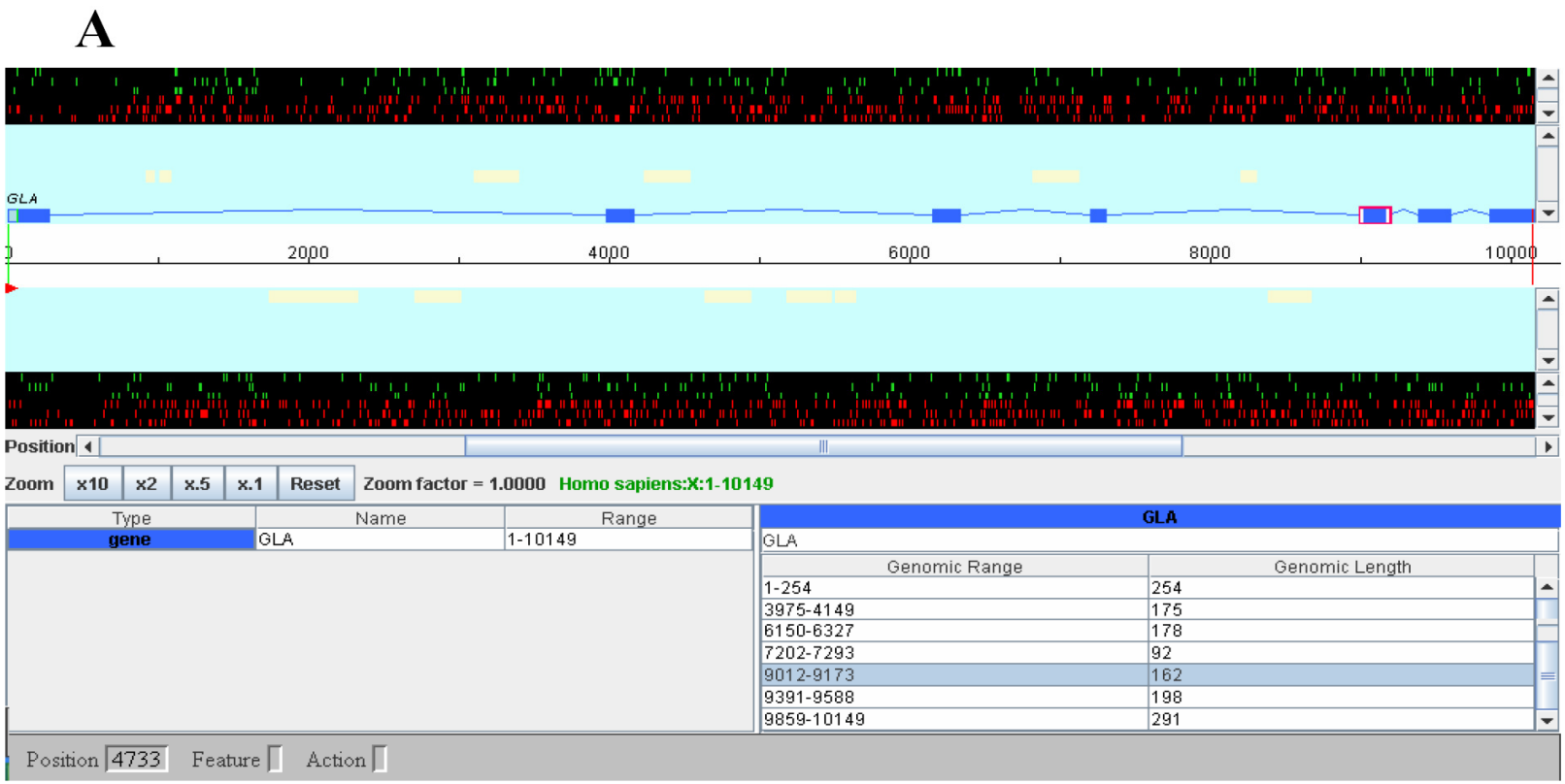


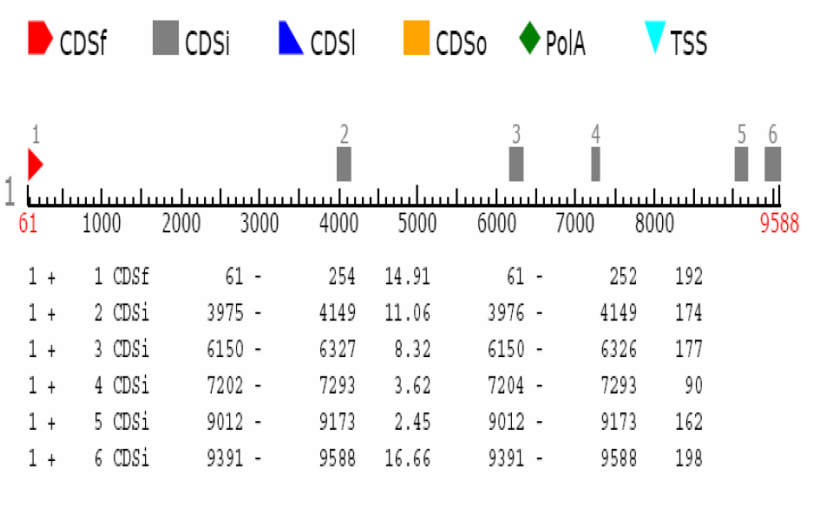

**Figure 1.** (A) Representation of exonic structure of the human FTP3 gene (*U78027*), displayed using the APOLLO graphical interactive interface (Lewis et al., 2003). The gene comprised 7 coding exons on the forward strand. (B) and (C) FGENESH and FGENES outputs in the sequence of the FTP3 gene (*U78027*), respectively.

ambiguous findings; such as, presenting a correct annotation will be somehow complicated. Likewise, since only the genome of a few number of organisms have been sequenced entirely, the users inevitably have to utilize *ab initio* based methods when the relative sequence of a given coding sequence is not accessible.

Regarding transcript prediction, the condition has become worse. In other words, according to the EGASP community experiment organized in 2005 (Guigo et al., 2006), finding the complete transcript structure was more challenging, with the most accurate methods correctly predicting only about 60% of the annotated protein-coding transcripts, although, computational methods were quite accurate in identifying protein-coding exons with an overall accuracy of more than 80% (in terms of both the fraction of real exons correctly identified and the fraction of predicted exons that are real) which were in agreement with our current results. Regarding protein-coding transcripts, the same results were also reported by Schweikert et al. (2009). This indicates that com-putational methods are yet to totally replace human expertise in gene annotation.

Another problem in identifying the protein coding sequences of eukaryotic genomes is the existence of repetitive elements. Actually, contrary to the typically streamlined genomes of prokaryotes, many eukaryotic genomes are

**B**



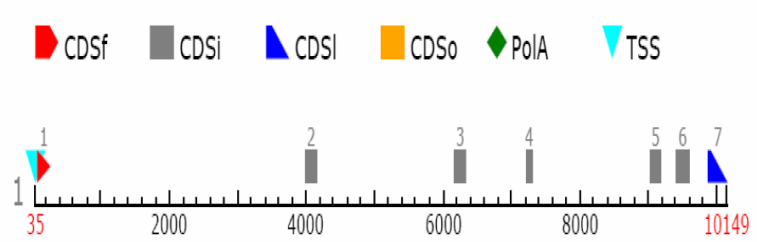

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 1 + | 1 | CDSf | 61 | - | 254 | 14.91 | 61 - 252 | 192 |
| 1 + | 2 | CDSi | 3975 | - | 4149 | 11.06 | 3976 - 4149 | 174 |
| 1 + | 3 | CDSi | 6150 | - | 6327 | 8.32 | 6150 - 6326 | 177 |
| 1 + | 4 | CDSi | 7202 | - | 7293 | 3.62 | 7204 - 7293 | 90 |
| 1 + | 5 | CDSi | 9012 | - | 9173 | 2.45 | 9012 - 9173 | 162 |
| 1 + | 6 | CDSi | 9391 | - | 9588 | 16.66 | 9391 - 9588 | 198 |

```
Seq name:  U78027
Length of sequence: 10149
Number of predicted genes 1: in +chain 1, in -chain 0.
Number of predicted exons 6: in +chain 6, in -chain 0.
Positions of predicted genes and exons: Variant   1 from   1, Score:38.606979
  G Str   Feature   Start        End    Score          ORF           Len

  1 +    1 CDSf       61 -       254   14.91       61 -       252    192
  1 +    2 CDSi     3975 -      4149   11.06     3976 -      4149    174
  1 +    3 CDSi     6150 -      6327    8.32     6150 -      6326    177
  1 +    4 CDSi     7202 -      7293    3.62     7204 -      7293     90
  1 +    5 CDSi     9012 -      9173    2.45     9012 -      9173    162
  1 +    6 CDSi     9391 -      9588   16.66     9391 -      9588    198

Predicted protein(s):
>FGENESH:   1   6 exon (s)     61   -   9588   333 aa, chain +
MQLRNPELHLGCALALRFLALVSWDIPGARALDNGLARTPTMGWLHWERFMCNLDCQEEP
DSCISEKLFMEMAELMVSEGWKDAGYEYLCIDDCWMAPQRDSEGRLQADPQRFPHGIRQL
ANYVHSKGLKLGIYADVGNKTCAGFPGSFGYYDIDAQTFADWGVDLLKFDGCYCDSLENL
ADGYKHMSLALNRTGRSIVYSCEWPLYMWPFQKPNYTEIRQYCNHWRNFADIDDSWKSIK
SILDWTSFNQERIVDVAGPGGWNDPDMLVIGNFGLSWNQQVTQMALWAIMAAPLFMSNDL
RHISPQAKALLQDKDVIAINQDPLGKQGYQLRQ
```

**C**




| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 + | | TSS | 35 | | | 1.07 | TATA | 9 | wTATA | 17.45 | LDF 0.36 |
| 1 + | 1 | CDSf | 61 | - | 254 | 9.30 | 61 - | 252 | | | |
| 1 + | 2 | CDSi | 3975 | - | 4149 | 3.69 | 3976 - | 4149 | | | |
| 1 + | 3 | CDSi | 6150 | - | 6327 | 1.29 | 6150 - | 6326 | | | |
| 1 + | 4 | CDSi | 7202 | - | 7293 | 2.83 | 7204 - | 7293 | | | |
| 1 + | 5 | CDSi | 9012 | - | 9173 | 2.17 | 9012 - | 9173 | | | |
| 1 + | 6 | CDSi | 9391 | - | 9588 | 4.53 | 9391 - | 9588 | | | |
| 1 + | 7 | CDSl | 9859 | - | 10149 | 3.49 | 9859 - | 10146 | | | |

```
Length of sequence:   10149 GC content: 0.44 Zone: 2
Number of predicted genes:   1 In +chain:   1 In -chain:   0
Number of predicted exons:   7 In +chain:   7 In -chain:   0
Positions of predicted genes and exons:
 G Str Feature  Start       End   Weight  ORF-start ORF-end


 1 +     TSS       35              1.07 TATA      9 wTATA   17.45 LDF   0.36
 1 +   1 CDSf      61 -     254     9.30       61 -     252
 1 +   2 CDSi    3975 -    4149     3.69     3976 -    4149
 1 +   3 CDSi    6150 -    6327     1.29     6150 -    6326
 1 +   4 CDSi    7202 -    7293     2.83     7204 -    7293
 1 +   5 CDSi    9012 -    9173     2.17     9012 -    9173
 1 +   6 CDSi    9391 -    9588     4.53     9391 -    9588
 1 +   7 CDSl    9859 -   10149     3.49     9859 -   10146

redicted proteins:
FGENES 1.6 > test sequence   1 Multiexon gene      61 -   10149     429 a Ch+
QLRNPELHLGCALALRFLALVSWDIPGARALDNGLARTPTMGWLHWERFMCNLDCQEEP
SCISEKLFMEMAELMVSEGWKDAGYEYLCIDDCWMAPQRDSEGRLQADPQRFPHGIRQL
NYVHSKGLKLGIYADVGNKTCAGFPGSFGYYDIDAQTFADWGVDLLKFDGCYCDSLENL
DGYKHMSLALNRTGRSIVYSCEWPLYMWPFQKPNYTEIRQYCNHWRNFADIDDSWKSIK
ILDWTSFNQERIVDVAGPGGWNDPDMLVIGNFGLSWNQQVTQMALWAIMAAPLFMSNDL
HISPQAKALLQDKDVIAINQDPLGKQGYQLRQGDNFEVWERPLSGLAWAVAMINRQEIG
PRSYTIAVASLGKGVACNPACFITQLLPVKRKLGFYEWTSRLRSHINPTGTVLLQLENT
QMSLKDLL
```

**Figure 1.** Contd.

riddled with long intergenic regions, spliceosomal introns, and repetitive elements (Irimia et al., 2009). The presence of repetitive elements is a severe problem in finding protein coding genes, particularly when a sequence with large scale is to be annotated and assembled. In fact, gene prediction programs ignore such stretches in making their predictions. *Repeatmasker* (Smith and Green, unpublished data) is one popular program used to find and mask repetitive elements. Actually, if a given sequence is masked, gene finders tend to predict less false-positive exons, because coding exons tend not to overlap or contain repetitive elements. The majority of genes that contain very short repetitive element, and no bias was observed in the programs' output. By contrast, approximately 40% of the human genome does contain such elements; but when analyzing large genomic sequences, the simple act of masking the sequence can have dramatic effects. For example, Genscan and GeneID predicted 1128 and 1119 genes in the unmasked sequence of human chromosome 22, but when the sequence was masked, the number of predicted genes dropped to 789 and 730, respectively (Blanco and Guigo, 2005). Accordingly, it is extremely clear that the length of a known sequence, its chromosomal position (for example, euchromatin or heterochromatin regions) and also the number of genes of an experimental genomic sequence can affect the prediction accuracy of each program. Finally, it should also be notified that if unmasked data are employed, a number of predicted exons would consequently be false-positive, while some actual exons would possibly be slipped whenever masked data are utilized. This is due to the different categories of repetitive DNA (that is, Satellite DNA, minisatellites and microsatellites known as Tandem-repetitive DNA or transposable elements known as interspersed repeats) which have been reported in both gene-rich regions, such as short terminal inverted repeats (TIRs) (Gierl and Saedler, 1992), or noncoding areas including Satellite DNA which are often located in subtelomeric or centromeric regions. It is accordingly advisable to run the programs with both masked and unmasked sequences as the input.

As an additional finding, it was observed here that the rate of GC content of each gene could play a fundamental role in prediction accuracy of each program. The condition, for instance, concerning the programs including FGENESH and GeneMark.hmm has become worse, as long as they are loaded by low-density GC genes, but FGENES and Genscan appear to perform slightly better in GC-poor sequences. In our study, prediction accuracy of FGENES was correlated with the GC content, while Rogic et al. (2001) clustered this program as a gene finder by means of an independent relation with the GC content. However, observing some unbiased errors which may occur when different programs with different algorithms were utilized was avoided. Accordingly, to predict the structure of a coding sequence, users should also measure the rate of the GC content of a given coding sequence.

In order to find a relation between the number of AE and the PE plus correct exon on the whole sequences, in this study, the correlation coefficient between them was also computed. Surprisingly, regardless of the kind of exon classes, a significant positive linear correlation was computed between the numbers of actual exons and predicted plus correct exon on the whole data. For all five programs, the values of $R^2$ (between the number of actual and predicted exon) ranged from 0.79 (GeneMark.hmm) to 0.96 (FGENESH). This option varied from 0.80 (HMMgene) to 0.95 for both Genscan and FGENESH, but with regards to the number of actual and correct exon (Figure 2). These findings illustrate that whenever FGENESH program is used, the number of both wrong and missing exon will possibly decline; consequently, a few false positive and false negative should be detected as it is shown in our investigation. Meanwhile, about having lower missing exon, Genscan seems to have enough potential as a superseded option. Conversely, if both GeneMark.hmm and HMMgene were used, the rate of wrong and missing exon, and consequently the number of false positive and false negative will move up, respectively. To avoid observing a great deal of false positives, particularly at the exon level, the probability of a predicted exon should be considered. In this case, some programs such as Genscan can compute the probability value (P) of each prediction, and it has been suggested that exons with lower probability ($P<0.50$) should be deemed unreliable, unless the same output is detected using other programs. High-probability predictions ($P>0.99$) can be used for example in the rational design of polymerase chain reaction (PCR) primers or for other purposes, where extremely high confidence is necessary. Anyway, normally all employed programs cannot predict gene structures accurately as it is shown in our results and also other available investigations.

## Conclusion

In conclusion, despite the fact that no gene finder can be definitely recommended as the best, using programs such as FGENESH and Gensacn seems to produce the most noteworthy prediction than others. Nonetheless, since each organism can be accompanied by a unique genetic background, some programs are species-specific and finally the algorithms used in some programs are different than others. Moreover, such occasions should be taken into account by users of gene finding programs. In the interim, to improve gene structure predictions, a combination of gene prediction results from multiple *ab initio* programs and also integration of *ab initio* and sequence similarity based methods together could be useful. At last, since majority of the genes exist in GC-

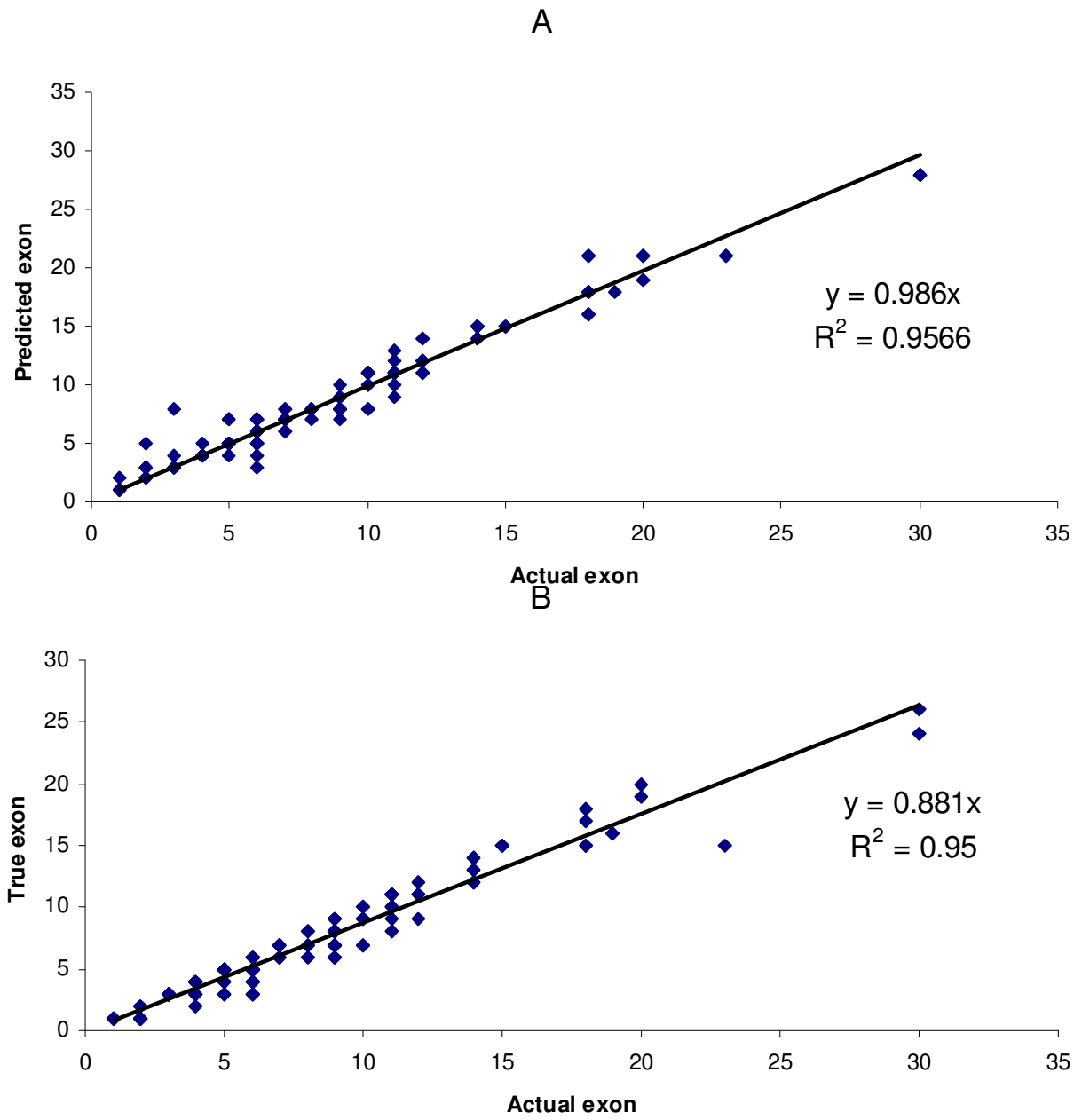



**Figure 2.** An example of calculating correlation coefficient between (A) the number of actual and predicted exons, and (B) the number of actual and true exons of FGENESH program of the whole data.

rich regions, paying attention to the rate of sensitivity value of each program versus this item should be considered either in applying contemporary programs or developing new gene finders with popular or novel algorithms.

## REFERENCES

Abeel T, Saeys Y, Bonnet E, Rouze P, Van de Peer Y (2008). Generic eukaryotic core promoter prediction using structural features of DNA. Genome Res. 18: 310-323.

Allen JE, Majoros HW, Pertea M, Salzberg LS (2006). JIGSAW, GeneZilla, and GlimmerHMM: puzzling out the features of human genes in the ENCODE regions. Genome Biol. 7(1): 9.

Bernal A, Crammer K, Hatzigeorgiou A, Pereira F (2007). Global discriminative learning for higher-accuracy computational gene prediction. PLoS Comput. Biol. 3: p. 54.

Bernardi G, Olofsson J, Filipski M, Zerial J, Salinas G, Cuny M, Rodier F (1985). The mosaic genome of warm-blooded vertebrates. Science, 228: 953-958.

Blanco E, Guigo R (2005). Chapter five: Predictive methods using DNA sequences. In: Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins. Andreas, D.B.and B.F. Francis Ouellette, Edition Number: 3. Wiley, John & Sons, Incorporated. pp. 116-142. ISBN-13: 9780471478782.

Brent MR, Guigo R (2004). Recent advances in gene structure prediction. Curr. Opin. Struct. Biol. 14: 264-272.

Burge C, Karlin S (1997). Prediction of complete gene structures in human genomic DNA. J. Mol. Biol. 268: 78-94.
Burge C, Karlin S (1998). Finding the genes in genomic DNA. Curr. Opin. Struct. Biol. 8: 346-354.
Burset M, Guigo R (1996). Evaluation of gene structure prediction programs. Genome, 34: 353-367.
Brodovsky M, McIninch J (1993). Genemark: Parallel Gene Recognition for both DNA Strands. Comput. Chem. 17: 123-133.
Cawley S, Pachter L, Alexandersson M (2003). SLAM web server for comparative gene finding and alignment. Nucleic Acid Res. 31: 3507-3509.
DeCaprio D, Vinson JP, Pearson MD, Montgomery P, Doherty M, Galagan JE (2007). Conrad: gene prediction using conditional random fields. Genome Res. 17: 1389-1398.
Do JH, Choi DK (2006). Computational Approaches to Gene Prediction. J. Micro. 44: 137-144.
Do JH, Choi DK (2005). 'Computational Approaches to Gene Prediction'. J. Micro. 44: 137-144.
Dong S, Searls DB (1994). Gene structure prediction by linguistic methods. Genomics, 23: 540-551.
Fassetti F, Leone O, Palopoli L, Rombo ES, Saiardi A (2010). IP6K gene identification in plant genomes by tag searching. From 6th International Symposium on Bioinformatics Research and Applications (ISBRA'10) BMC Proc. 5(2): p. 1.
Fickett JW (1982). Recognition of protein coding regions in DNA sequences. Nucleic Acids Res. 10: 5303-5318.
Fickett JW (1996). The gene identification problem: An overview for developers. Comput. Chem. 20: 103-118.
Flicek P, Keibler E, Hu P, Korf I, Brent MR (2003). Leveraging the mouse genome for gene prediction in human: from whole-genome shotgun reads to a global synteny map. Genome Res. 13: 46-54.
Galtier N, Piganeau G, Mouchiroud D, Duret L (2001). GC-Content Evolution in Mammalian Genomes: The Biased Gene Conversion Hypothesis. Genet. 159: 907-911.
Gelfand MS (1995). Prediction of function in DNA sequence analysis. J. Comp. Biol. 2: 87-115.
Gierl A, Saedler H (1992). Plant-transposable elements and gene tagging. Plant Mol. Biol. 19: 39-49.
Gribskov M, Devereux J, Burgess RR (1984). The codon preference plot: graphic analysis of protein coding sequences and prediction of gene expression. Nucleic Acids Res. 12: 539-549.
Gross SS, Brent MR (2006). Using multiple alignments to improve gene prediction. Comput. Biol. 13: 379-393.
Guigo R, Fickett JW (1995). Distinctive sequence features in protein coding, genic non-coding, and intergenic human DNA. J. Mol. Biol. 253: 51-60.
Guigo R (1999). DNA composition, codon usage and exon prediction. In Genetic Databases (Academic Press).
Guigo R, Wiehe T (2003). Gene Prediction Accuracy in Large DNA Sequences. In Galperin MY and Koonin EV, editors: 'Frontiers in Computational Genomics. (Functional Genomics Series, Volume 3) Caister Academic Press, United Kingdom, ISBN: 0-9542464-46. 1: 1-33.
Guigo R, Flicek P, Abril JF, Reymond A, Lagarde J, Denoeud F, Antonarakis S, Ashburner M, Bajic VB, Birney E, Castelo R, Eyras E, Ucla C, Gingeras TR, Harrow J, Hubbard T, Lewis SE, Reese MG (2006). EGASP: the human ENCODE Genome Annotation Assessment Project. Genome Biol. 7(1): 2: 1-31.
Harrow J, Nagy A, Reymond A, Alioto T, Patthy L, Antonarakis SE, Guigo R (2009). Identifying protein-coding genes in genomic sequences. Genome Biol. 10: 201.
Irimia M, Roy SW, Neafsey DE, Abril JF, Garcia-Fernandez J, Koonin EV (2009). Complex selection on 59 splice sites in intron-rich organisms. Genome Res. 19: 2021-2027.
Knapp K, Chen PYP (2006). An evaluation of contemporary hidden Markov model genefinders with a predicted exon taxonomy. Nucleic Acid Res. 35: 317-324.
Koonin EV, Mushegian AR, Borr P (1996). Non-orthologous gene displacement Trends Genet. 12: 334-336.
Korf I, Flicek P, Duan D, Brent MR (2001). Integrating Genomic Homology into Gene Structure Prediction. Bioinformation, 1 (1): 1-9.
Krogh A (1997). Two methods for improving performance of an HMM and their application for gene finding. Proc. Int. Conf. Intell. Syst. Mol. Biol. 5: 179-186.
Kwan LA, Li L, Kulp CD, Dutcher KS, Stormo DG (2009). Improving Gene-finding in *Chlamydomonas reinhardtii*: GreenGenie2. BMC Genome, 10: p. 210.
Liang C, Mao L, Ware D, Stein L (2009). Evidence-based gene predictions in plant genomes. Genome Res. 19: 1912-1923.
Lewis SE, Searle SMJ, Harris N, Gibson M, Iyer V, Richter J, Wiel C, Bayraktaroglu L, Birney E, Crosby MA, Kaminker JS, Matthews BB, Prochnik SE, Smith CD, Tupy JL, Rubin GM, Misra S, Mungall CJ, Clamp ME (2003). Apollo: a sequence annotation editor. Genome Biol. 3: 1-14.
Li H, Liu JS, Xu Z, Jin J, Fang L, Gao L, Li YD, Xing ZX, Gao ShG, Liu T, Li HH, Li Y, Fang LJ, Xie HM, Zheng WM, Hao BL (2005). Test data set and evaluation of gene prediction programs on the rice genome. J. Comput. Sci. Tech. 20: 446-453.
Lomsadze A, Ter-Hovhannisyan V, Chernoff OY, Borodovsky M (2005). Gene identification in novel eukaryotic genomes by self-training algorithm. Nucleic Acid Res. 33: 6494-6506.
Lopez R, Larsen F, Prydz H (1994). Evaluation of the exon predictions of the GRAIL software. Genome, 24: 133-136.
Lukashin AV, Borodovsky M (1998). GeneMark.hmm: New solutions for gene-finding. Nucleic Acid Res. 26: 1107-1115.
Mathe C, Sagot MF, Schiex T, Rouze P (2002). Current methods of gene prediction, their strengths and weaknesses. Nucleic Acid Res. 30: 4103-4117.
Mouchiroud D, D'onofrio G, Aissani B, Macaya G, Gautier C, Bernardi G (1991). The distribution of genes in the human genome. Gene. 100: 181-187.
Nasiri J, Haghnazari A, Alavi M (2011). Evaluation of prediction accuracy of genefinders using Mouse genomic DNA. Trends Bioinform. 4: 10-22.
Parra G, Agarwal P, Abril J, Wiehe T, Fickett J, Guigo R (2003). Comparative Gene Prediction in Human and Mouse. Genome Res. 13: 108-117.
Ratsch G, Sonnenburg S, Srinivasan J, Witte H, Muller KR, Sommer R, Scholkopf B (2007). Improving the *C. elegans* genome annotation using machine learning. PLoS Com. Biol. 3: 20.
Rogic S, Alan MK, Francis BFO (2001). Evaluation of gene-finding programs on mammalian sequences. Genome Res. 11: 817-832.
Salamov AA, Solovyev VV (2000). *Ab initio* gene finding in Drosophila genomic DNA. Genome Res. 10: 391-393.
Schweikert G, Behr J, Zien A, Zeller G, Ong CS, Sonnenburg S, Ratsch G (2009). mGene.web: a web service for accurate computational gene finding. Nucleic Acid Res. pp. 312-316.
Snyder EE, Stormo GD (1995). Identification of protein coding regions in genomic DNA. J. Mol. Biol. 248: 1-18.
Staden R (1984). Measurements of the effect that coding for a protein has on DNA sequence and their use for finding genes. Nucleic Acids Res. 12: 551-567.
Stanke M, Steinkamp R, Waack S, Morgenstern B (2004). AUGUSTUS: a web server for gene finding in eukaryote. Nucleic Acid Res. 32: 309-312.
Stanke M, Diekhans M, Baertsch R, Haussler D (2008). Using native and syntenically mapped cDNA alignments to improve *de novo* gene finding. Bioinformation, 24: 637-644.
Stormo GD (2000). Gene-finding approaches for eukaryotes. Genome Res. 10: 394-397.
Taher L, Rinner O, Garg S, Sczyrba A, Morgenstern B (2004). AGenDA: gene prediction by cross-species sequence comparison. Nucleic Acid Res. 32: 305-308.
Vladimir AM (2002). Computer programs for eukaryotic gene prediction. Brief. Bioinform. 3: 195-199.
Wang Z, Chen Y, Li Y (2004). A Brief Review of Computational Gene Prediction Methods. Genome Prot. Bioinform. 4: 216-221.

Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, Antonarakis SE, Attwood J, Baertsch R, Bailey J, Barlow K, Beck S, Berry E, Birren B, Bloom T, Bork P, Botcherby M, Bray N, Brent MR, Brown DG, Brown SD, Bult C, Burton J, Butler J, Campbell RD, Carninci P, Cawley S, Chiaromonte F, Chinwalla AT, Church DM, Clamp M, Clee C, Collins

FS, Cook LL, Copley RR, Coulson A, Couronne O, Cuff J, Curwen V, Cutts T, Daly M, David R, Davies J, Delehaunty KD, Deri J, Dermitzakis ET, Dewey C, Dickens NJ, Diekhans M, Dodge S, Dubchak I, Dunn DM, Eddy SR, Elnitski L, Emes RD, Eswara P, Eyras E, Felsenfeld A, Fewell GA, Flicek P, Foley K, Frankel WN, Fulton LA, Fulton RS, Furey TS, Gage D, Gibbs RA, Glusman G, Gnerre S, Goldman N, Goodstadt L, Grafham D, Graves TA, Green ED, Gregory S, Guigó R, Guyer M, Hardison RC, Haussler D, Hayashizaki Y, Hillier LW, Hinrichs A, Hlavina W, Holzer T, Hsu F, Hua A, Hubbard T, Hunt A, Jackson I, Jaffe DB, Johnson LS, Jones M, Jones TA, Joy A, Kamal M, Karlsson EK, Karolchik D, Kasprzyk A, Kawai J, Keibler E, Kells C, Kent WJ, Kirby A, Kolbe DL, Korf I, Kucherlapati RS, Kulbokas EJ, Kulp D, Landers T, Leger JP, Leonard S, Letunic I, Levine R, Li J, Li M, Lloyd C, Lucas S, Ma B, Maglott DR, Mardis ER, Matthews L, Mauceli E, Mayer JH, McCarthy M, McCombie WR, McLaren S, McLay K, McPherson JD, Meldrim J, Meredith B, Mesirov JP, Miller W, Miner TL, Mongin E, Montgomery KT, Morgan M, Mott R, Mullikin JC, Muzny DM, Nash WE, Nelson JO, Nhan MN, Nicol R, Ning Z, Nusbaum C, O'Connor MJ, Okazaki Y, Oliver K, Overton-Larty E, Pachter L, Parra G, Pepin KH, Peterson J, Pevzner P, Plumb R, Pohl CS, Poliakov A, Ponce TC, Ponting CP, Potter S, Quail M, Reymond A, Roe BA, Roskin KM, Rubin EM, Rust AG, Santos R, Sapojnikov V, Schultz B, Schultz J, Schwartz MS, Schwartz S, Scott C, Seaman S, Searle S, Sharpe T, Sheridan A, Shownkeen R, Sims S, Singer JB, Slater G, Smit A, Smith DR, Spencer B, Stabenau A, Stange-Thomann N, Sugnet C, Suyama M, Tesler G, Thompson J, Torrents D, Trevaskis E, Tromp J, Ucla C, Ureta-Vidal A, Vinson JP, Von Niederhausern AC, Wade CM, Wall M, Weber RJ, Weiss RB, Wendl MC, West AP, Wetterstrand K, Wheeler R, Whelan S, Wierzbowski J, Willey D, Williams S, Wilson RK, Winter E, Worley KC, Wyman D, Yang S, Yang SP, Zdobnov EM, Zody MC, Lander ES (2002). Initial sequencing and comparative analysis of the mouse genome. Nature, 420: 520-562.

Xu Y, Einstein JR, Mural RJ, Shah M, Uberbacher EC (1994). An improved system for exon recognition and gene modeling in human DNA sequences. In Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology, pp. 376-384.

Yao H, Guo L, Fu Y, Borsuk L, Wen TJ, Skibbe D, Cui X, Scheffler B, Cao J, Emrich S, Ashlock D, Schnable P (2005). Evaluation of five *ab initio* gene prediction programs for the discovery of maize genes. Plant Mol. Biol. 57: 445-460.

Zhang MQ (2002). Computational prediction of eukaryotic protein-coding genes. Nat. Rev. Genet. 3: 698-709.

Zhu H, Hu GQ, Yang YF, Wang J, She ZS (2007). MED: a new non-supervised gene prediction algorithm for bacterial and archaeal genomes. BMC Bioinform. 8: p. 97.

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