Ab initio gene identification: prokaryote genome annotation with GeneScan and GLIMMER

GAUTAM AGGARWAL* and RAMAKRISHNA RAMASWAMY[†]
School of Physical Sciences, Jawaharlal Nehru University, New Delhi 110 067, India
**Present address: Seattle Biomedical Research Institute, Seattle, WA 98109, USA*

[†]*Corresponding author (Fax, 91-11-6194137; Email, rama@vsnl.com).*

We compare the annotation of three complete genomes using the *ab initio* methods of gene identification GeneScan and GLIMMER. The annotation given in GenBank, the standard against which these are compared, has been made using GeneMark. We find a number of novel genes which are predicted by both methods used here, as well as a number of genes that are predicted by GeneMark, but are not identified by either of the nonconsensus methods that we have used. The three organisms studied here are all prokaryotic species with fairly compact genomes. The Fourier measure forms the basis for an efficient non-consensus method for gene prediction, and the algorithm GeneScan exploits this measure. We have bench-marked this program as well as GLIMMER using 3 complete prokaryotic genomes. An effort has also been made to study the limitations of these techniques for complete genome analysis. GeneScan and GLIMMER are of comparable accuracy insofar as gene-identification is concerned, with sensitivities and specificities typically greater than 0·9. The number of false predictions (both positive and negative) is higher for GeneScan as compared to GLIMMER, but in a significant number of cases, similar results are provided by the two techniques. This suggests that there could be some as-yet unidentified additional genes in these three genomes, and also that some of the putative identifications made hitherto might require re-evaluation. All these cases are discussed in detail.

[Aggarwal G and Ramaswamy R 2002 *Ab initio* gene identification: prokaryote genome annotation with GeneScan and GLIMMER; *J. Biosci*. (*Suppl. 1*) **27** 7–14]

1. Introduction

The increased effort in genome sequencing has led to a drastic reduction in the amount of time required to determine the complete DNA sequence of a given organism. The first draft of the Human genome has already been released, as have the complete sequences for human chromosome 22 (Dunham *et al* 1999) and 21 (Hattori *et al* 2000). A large number of smaller genomes have been completely sequenced, ranging from the first example, *Mycoplasma genitalium* which has a genome of size 0·58 Mb, to *Caenorhabditis elegans*, with a size of 100·1 Mb, and *Drosophila melanogaster* which is ~ 137 Mb in size[1].

The initial step in studying a given genome is its annotation, namely the identification of coding regions. This is usually carried out via computational means and there are now several methods of computational or *in silico* gene identification (Claverie 1997; Fickett 1996; Guigo 1999) which can be routinely employed to determine the different genes on the genome. These methods are based on a variety of different properties of protein coding regions, and use a variety of different mathematical techniques, including neural networks [GRAIL (Xu and Uberbacher 1997), GENEID (Parra *et al* 2000), Markov models (GeneMark) (Borodovsky and McIninch 1993), GENSCAN (Burge and Karlin 1997), GLIMMER (Delcher *et al* 1999)], Fourier transforms (GeneScan), etc.

These methods can be differentiated as *ab initio* (or non-consensus) methods and consensus methods, depending

[1] A list of completely sequenced genomes is available on the website, http://linkage.rockefeller.edu/wli/seq/.

Keywords. GLIMMER; GeneScan

on whether they need to be trained on a set of genes in order to assess whether or not a query sequence is coding or not. Consensus methods are indeed very successful in identifying genes in organisms which are similar to those on which the methods have been trained or optimized. These methods are less successful in identifying novel organisms or identifying novel genes. Part of the problem can be ascribed to the fact that the existing genes in GenBank, on which these consensus methods are trained, are biased towards the highly expressed genes in extensively studied organisms.

Just locating all the genes can itself be quite revealing: Human chromosomes 21 and 22, have approximately 770 genes between them, leading to an estimate of about 40,000 for the entire human genome, a downward revision from the 100,000 genes presumed to be present. Finding the genes is the first step. More complete understanding requires the DNA to be also characterized in terms of promoters, regulatory elements, intergenic regions etc. Determining what the genes do is a more complicated task, and is usually determined by comparative or homology analysis. In spite of this, a large fraction of genes (up to 1/3 in *Saccharomyces cerevisiae* and about 20% in *C. elegans*) must be presently classed as of “unknown” function (Vukimirovic and Tilghman 2000). In the case of prokaryotic organisms, Uberbacher *et al* (1996) estimate that up to 50% of the newly discovered genes have no prior homologues in existing sequence databases.

In contrast, there are relatively few non-consensus methods of gene prediction, namely methods that are based either on “universal” measures for differentiating between coding and noncoding DNA, or on some self-consistent model of gene structure. In the former category, there are two methods that are based on the correlation properties of DNA sequences, GeneScan (Tiwari *et al* 1997) and the Coding Region Finder (Ossadnik *et al* 1994) methods. The latter category includes SELFID (Audic and Claverie 1998) and GLIMMER (Delcher *et al* 1999), which are interpolated Markov model based techniques.

In this paper we compare the predictions of two of the nonconsensus methods, namely GeneScan and GLIMMER with annotation of three completely sequenced genomes of the organisms *Haemophilus influenzae*, *Helicobacter pylori*, and *Campylobacter jejuni*. All these organisms have been annotated previously using the extremely successful program GeneMark.hmm, which is a (hidden) Markov model based algorithm. Our motivation is two-fold. Firstly, we wish to undertake an extensive comparison of the predictions of GeneScan, a gene identification tool that has been developed in this laboratory, against the standard annotation. Our second motivation is to compare the two nonconsensus methods against each other. There has been some controversy recently between GLIMMER predictions versus those of other methods (Pertea *et al* 2000).

Although in the majority of cases, the genes identified by the present two techniques are the same as those identified by the earlier used methods, by using two independent nonconsensual methods we hope to locate novel genes which may have been missed. Furthermore, there are also cases where the predictions of these non-consensus methods is at variance with the earlier annotation. In the case of putative genes, this can be indicative of open reading frames (ORFs) which are ultimately noncoding.

One of the objectives of computational genomics is to develop algorithms with 100% specificity and sensitivity. Many individual programs are in the 80% or greater range (Burset and Guigo 1996). The current practice in gene identification is to use the criterion that a given sequence should be independently identified as a gene by two or more (usually consensual) algorithms. The focus has shifted from improving any given individual program to devising algorithms which are based on a combination of techniques. This study, which stringently examines the role that *ab initio* methods can play in this regard, is a step in that direction.

2. Methods

The basic algorithm used in GeneScan[2] has already been described in detail (Tiwari *et al* 1997). The essential feature of this program is its detection of the 3-base periodicity which is shown by coding regions to identify putative genes. Briefly, this is done as follows. A given symbolic sequence (ATGTAGCA . . ., for instance) is converted into four binary signals using projection operators U_a, a = A, T, G and C which replace the symbol sequence by a digital signal containing 1 in those positions where the base is a and 0 elsewhere. These signals are then Fourier analysed to obtain the power spectrum,

$$S(f) = \sum_a S_a(f) = \sum_a \frac{1}{N^2} \left| \sum_{j=1}^{N} U_a(x_j) \exp(2pifj) \right|^2 . \tag{1}$$

The length of the spectrum is N, and the discrete frequency $f = k/N$; $k = 1 \ldots N/2$. The power spectrum for protein coding regions regardless of the organism, typically shows a distinctive peak at frequency $f = 1/3$, and this peak is absent (or lost in the noise) for

[2] A commercial version of the program GeneScan can be purchased from Unitech International Ltd., 491 Udyog Vihar, Gurgaon 122 015, India.

noncoding sequences. The measure used is the peak to noise ratio for the spectral peak at this frequency, namely

$$P_N = \frac{S(f = 1/3)}{\bar{S}}, \tag{2}$$

normalized to unit power[3]. The average of the entire spectrum is

$$\bar{S} = \frac{2}{N}\sum_{k=1}^{N/2} S(K/N) = \frac{1}{N}\left(1 + \frac{1}{N} - \sum_a r_a^2\right), \tag{3}$$

r_a being the frequency of nucleotide a. The result of an extensive study of coding and noncoding DNA sequences has shown that P_N exceeds the value 4 for 95% of all coding sequences, and below the same value for about 90% of all noncoding sequences. By calculating the value of this measure in a window of length M, centered on nucleotide j, and sliding this window along the sequence, one can devise a simple method to distinguish coding and noncoding regions in a plot of $P_M(j)$ versus j.

The algorithm employed in GLIMMER has also been described in detail (Delcher *et al*, 1999). The method employs a self-consistent Markov model which is "trained on the job", namely, one identifies ORFs of sufficient length which are most likely to be coding to give an initial model of the coding regions of the organism. This information is subsequently used in a Markov chain in order to locate all other coding sequences. The version of GLIMMER2 that was used here was obtained from the TIGR website[4].

3. Comparative analysis

The complete genomes that we have analysed here are those of *H. influenzae*, *H. pylori*, and *C. jejuni*. The genes on these genomes all consist of single exonic regions, namely they are coding ORFs.

The training set used for the GLIMMER results consisted of all ORFs of length greater than 1000 bp. The resulting output file of predictions is parsed into sense and antisense predictions for comparison with documented CDS (predicted using GeneMark.hmm) as obtained from the header of the GeneBank file of the sequence under analysis. For purposes of standardization, we take this annotation as correct.

False Negatives denoted by FN, therefore, are coding regions that are present in the GenBank annotation which are not predicted to be coding by the algorithm being used. Correspondingly, False Positives denoted by FP, are coding regions identified by a given algorithm which are not present in the standard annotation. TP or True Positives are those genes that are correctly predicted by the algorithm and also exist in the GenBank annotation.

The accuracy of a given method can be judged by a number of different criteria, and here we use the simplest, namely the Sensitivity (S_N) and Specificity (S_P), defined as

$$S_N = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{4}$$

$$S_P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}. \tag{5}$$

The sensitivity of a given method is strongly dependent on number of False Negatives.

GeneScan predictions are made as follows. Probable coding regions are identified as described earlier, and the sequence is scanned in order to identify coding ORFs. For ORF's longer than 300 bp, the P_N measure is computed, and based on the nature of the Fourier spectrum, the ORF is predicted as coding or non-coding (Bhattacharya *et al* 1999, 2000; Ramachandran and Ramakrishna 1999).

4. Results

In table 1 we present the results of comparative study of the two *ab initio* methods. The three genomes used in this

Table 1. Results of comparative analysis of *ab initio* methods GeneScan and GLIMMER against the GenBank annotations (predicted using GeneMark) in three complete genomes of prokaryotic organisms.

	C. jejuni	*H. influenzae*	*H. pylori*
Genome length (~ Mbp)	1·64	1·83	1·67
G + C content (%)	30·5	38·2	38·9
Total number of predicted CDS	1654	1709	1566
Number of CDS larger than 300 bp	1502	1543	1390
FN (Glimmer)	10	3	6
FN (GeneScan*)	46 (2**)	25 (8**)	42 (18**)
Common FN	9	1	4
FP (Glimmer)	19	54	39
FP (GeneScan)	45	69	56
Common FP	7	29	15
S_N (GeneScan)	96·97	98.38	96·98
S_N (Glimmer)	99·33	99·80	99·57
S_P (GeneScan)	96·00	95·65	96·01
S_P (Glimmer)	98·74	96·61	97·26

*In case of GeneScan, FN predictions have contribution from two accounts: ORF totally missed and matching ORF found but correlation value less than discriminator.

**Number of genes missed due to ORF having values less than threshold.

+ S_N and S_P are the Sensitivity and Specificity respectively (see §3).

[3] Thus P_N is further divided by a factor $N\bar{S}$ to make it independent of the choice of window length. This was not explicitly clarified in Tiwari *et al* (1997).

[4] The TIGR website http://www.tigr.org/softlab/glimmer/glimmer.html has information on distribution of the program.

study are prokaryotic species, with G + C content between 30–40%. [The present (default) version of GeneScan uses the standard genetic code and does not allow for alternate start or stop codons.]

Although both the present methods are based on very different algorithms, the predictions of the two are in consonance for the three genomes studied here. Both methods have a high degree of accuracy, typically correctly identifying over 95% of the genes on the genomes. On an average, the sensitivity of GLIMMER is higher (99%) than that of GeneScan (96%), though their specificity is comparable (96%). For both techniques, a decrease in the number of FN predictions is accompanied by an increase in the number of FP predictions. For the individual genomes, the comparative results are as follows. We largely focus on FPs or FNs which are predicted by both GeneScan and GLIMMER since these are more likely to be significant in contrast with FPs or FNs identified by a single program.

In *C. jejuni* the two programs find 9 common False Negatives which are tabulated in table 2 which also gives the description of the coding region from the GenBank annotation. Most false negatives appear to be pseudogenes with no other homologues in GenBank. A detailed study of the Fourier spectrum of these putative coding regions showed that the spectrum was either very grassy, or that there were multiple peaks, leading to the classification by GeneScan as noncoding; this is annotated as ‘not a gene’ and ‘probably not a gene’, respectively in table 2. Two cases, namely CJ0444 and CJ0672 are of particular interest as they appear to be FNs surrounded by FPs. These two cases are discussed in detail in the next section.

There are 7 False Positives found in *C. jejuni*; these are listed in table 3. Five of these are homologous to other portions of *C. jejuni* and may be copies of other genes. One putative coding region on the antisense strand between positions 1448554–1446341 does not have homology to any known gene and is possibly a novel gene that requires experimental verification and study. The seventh FP has partial matches to the kdpA gene in *C. jejuni*.

In *H. influenzae*, there is only one False Negative. This CDS is annotated as a hypothetical protein. The similarity search shows significant alignment to many sections of

Table 2. Results of study of common FN predictions by both GeneScan and GLIMMER methods. Similarity search was done using BLASTN* of NCBI. The predictions are classified depending upon the nature of power spectrum in GeneScan.

Sequence identity	Similarity search results	Comments
C. jejuni false negative		
Cj0046 (Psuedogene)	Protein match in *C. jejuni*	Not a gene
Cj0223 (–do–)	–do–	–do–
Cj0444 (–do–)	–do–	See discussion
Cj0565 (–do–)	–do–	Probably not a gene
Cj0672 (putative periplasmic protein)	Protein match in *C. jejuni* and kdpA gene in *C. jejuni*	See discussion
Cj0752 (Psuedogene)	Protein match in *C. jejuni* and similar to parts of *H. pylori*	Probably not a gene
Cj0866 (–do–)	Protein match in *C. jejuni*	Not a gene
Cj0654c (–do–)	No significant match	–do–
Cj0072c (–do–)	–do–	Probably not a gene
H. influenzae false negative		
HI0493 (hypothetical protein)	Matches to many sections in *H. influenzae*	Not a gene
H. pylori false negative		
HP0411 (hypothetical protein)	Significant matches to many organisms	Not a gene
HP1585 (identified by sequence similarity)	Two protein match in *H. pylori*	Probably not a gene
HP0985 (hypothetical protein)	No significant match	Not a gene
HP0536 (identified by sequence similarity)	No significant match	Not a gene

*BLASTN is available at http://www.ncbi.nlm.nih.gov/BLAST.

Table 3. Results of study of common FP predictions by both GeneScan and GLIMMER methods. Similarity search was done using BLASTN of NCBI.

False Positives	No matches	Single matches in the organism itself	Multiple matches
C. jejuni	1448554–1446341(R)	412299–412928, 413504–414433, 1326204–1326605, 1326571–1327107, 1330451–1331008	628193–629062
H. influenzae	1808859–1807963(R), 1806013–1804886(R), 1719912–1718821(R), 1594854–1594339(R), 1526019–1525285(R), 1279935–1279189(R), 1173709–1172942(R), 1107848–1106472(R), 1021200–1018846(R), 854724–854335(R), 848570–848166(R), 655009–654365(R), 279121–276992(R), 240205–239516(R), 170577–169396(R), 130954 -129317(R)	6322–8748, 132222–132959, 201627–202151, 235441–235932, 266944–267378, 928097–929080, 1159701–1160618, 1348011–1348478, 1378066–1378752, 1586980–1587765	235913–238519, 370428–370808, 370801–372912
H. pylori	1616735–1615878(R), 1302953–1302651(R), 954661–953783(R), 799543–798959(R), 780861–779008(R), 7145–5241(R)	–Nil–	774515–776341, 1430856–1432280

(R) stands for complimentary strand of DNA.

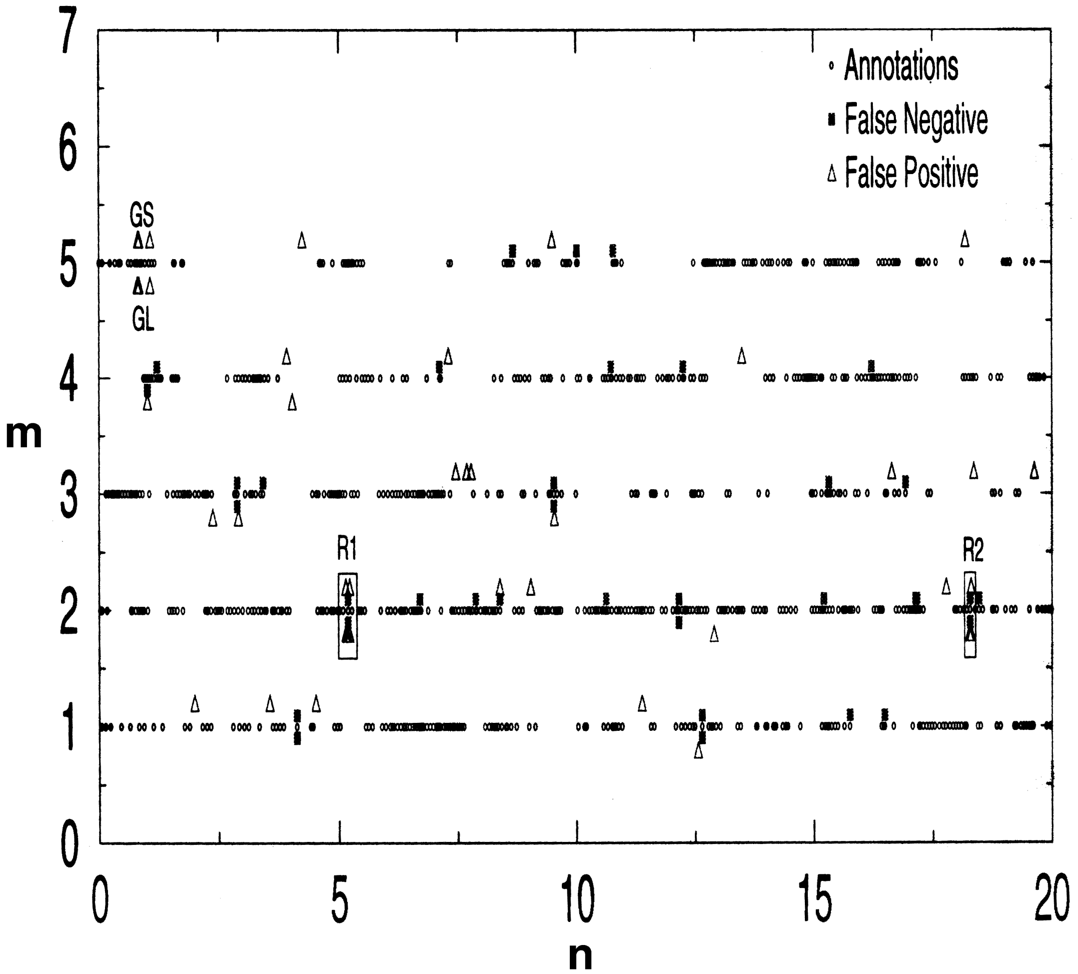

Figure 1. This summarizes results for the sense strand of *C. jejuni*. The entire genome is represented by a line of length 100 which is broken up as 5 lines of length 20, placed horizontally. The annotated genes are indicated along this line by an open circle (at the start position of the CDS). Thus the actual CDS position can be read as (m*n*1·64/100) Mb. Above this line, the GeneScan FP (triangle) and FN (square) predictions are shown, and correspondingly, below this line the FP and FN predictions of GLIMMER. The specific regions of interest discussed in detail in the text are boxed and marked as R1 and R2.

H. influenzae, but the Fourier spectrum of this sequence does not have a sharp peak at frequency $f = 1/3$. Thus GeneScan does not predict it to be a gene. On the other hand, we find a large number (29) of False Positives. Three of these show a significant similarity to genes from other organisms such as *Escherichia coli* and *Neisseria meningitidis,* and the Fourier spectrum for 2 of them is "typical" of genes (Tiwari *et al* 1997). Sixteen of these appear to be possibly novel genes, while the remainder appear to be copies of other genes on the genome. This organism is of significant experimental interest and the fact that both GLIMMER and GeneScan predict novel coding regions should catalyse experimental investigations of these novel putative genes.

In *H. pylori* there are 4 False Negatives, and one of these, HP0411, shows significant matches with coding regions in many organisms. The Fourier transform for this sequence, however, indicates that the strongest

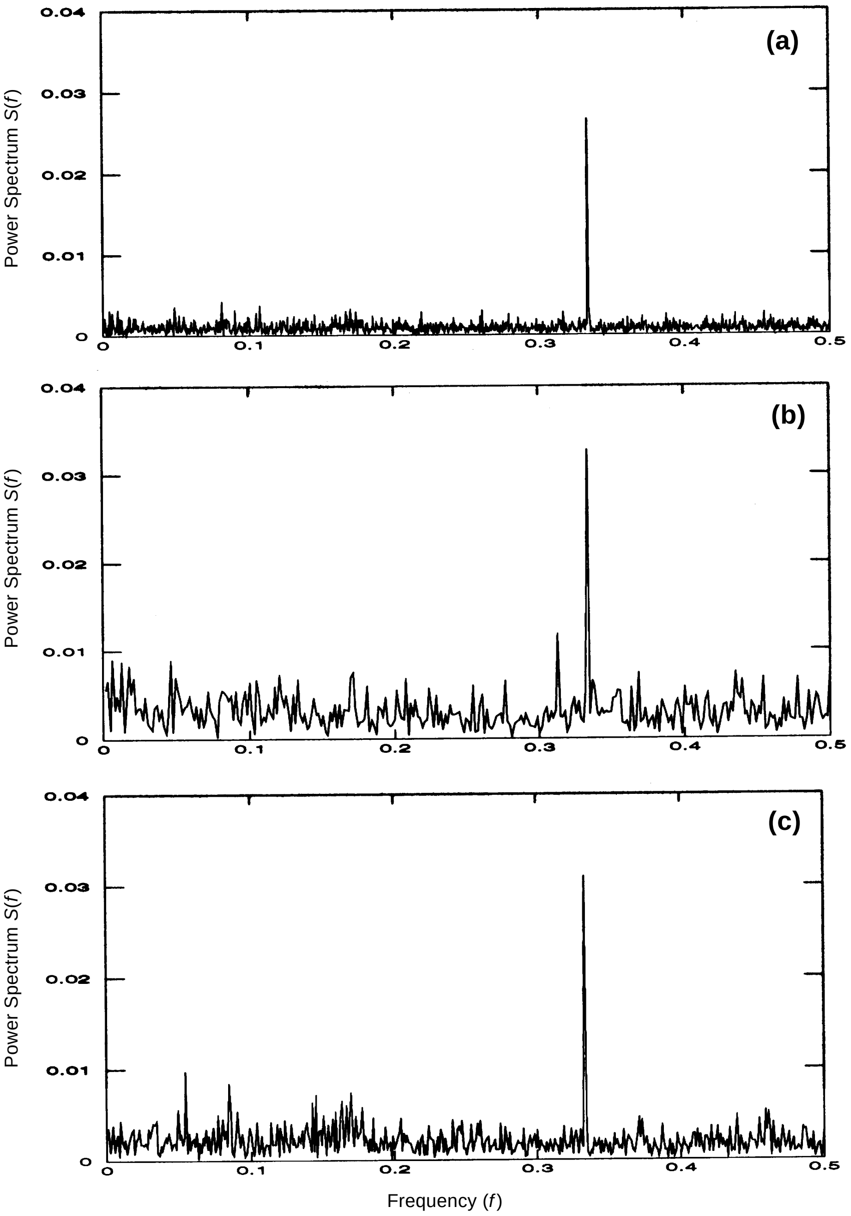

Figure 2. Fourier spectra for the FP and FN predictions by both GeneScan and GLIMMER programs in region R1 (see figure 1) of *C. jejuni.* Spectrum for (**a**) FN (CJ0444, CDS 412359 . . 414430) with $P_N = 10{\cdot}425$, (**b**) FP (412299 . . 412928) with $P_N = 28{\cdot}64$ and (**c**) FP (413504 . . 414433) with $P_N = 40{\cdot}439$.

correlation within the gene is not the 3-base period: there are several other peaks in the spectrum, and thus GeneScan predicts this to be noncoding. Other sequences also show poor correlation statistics and thus the existing prediction for this set of genes requires re-examination.

In this genome the two programs predict 8 False Positives, two of which have a significant alignment to other sections of *H. pylori* itself. The power spectra of these sequences again are typical of genes, so it is very likely that these are copies of other genes in the genome. In addition, there are 6 novel sequences which do not show any significant alignment to any other gene sequence in GenBank.

5. Discussion

Figure 1 summarizes our results for sense strand of *C. jejuni*. The other two genomes studied do not have any coinciding FN and FP predictions and hence are not discussed in similar detail here.

The genes on the genome are indicated as points on a line as described in the figure caption, and the FPs and FNs of the two programs, GeneScan and GLIMMER are respectively indicated above and below this line which represents the genome. Most genes are indeed identified by both the programs, and therefore the TP predictions are not indicated for clarity.

Two regions containing predictions at variance with the existing annotation are of interest. These are at positions 412359 . . 414430 and 627316 . . 629059 on the genome. Both these regions are characterized by different reading frames for start and stop codon. In the first region, the false negative is flanked by false positives. Closer examination of the false negative prediction for the annotated gene CJ0444 (CDS 412359 . . 414430) shows the presence of an alternate stop codon (TTT). Compared to this putative gene, the Fourier spectra of the flanking FPs (412299 . . 412928 and 413504 . . 414433) more clearly indicates the presence of the three-base correlation (cf. figure 2). Furthermore these have homology to known genes in *C. jejuni*, and thus the present results suggest a re-annotation of this portion of the genome.

The other interesting region is in the neighbourhood of CJ0672 (CDS 627316 . . 629059) a putative periplasmic

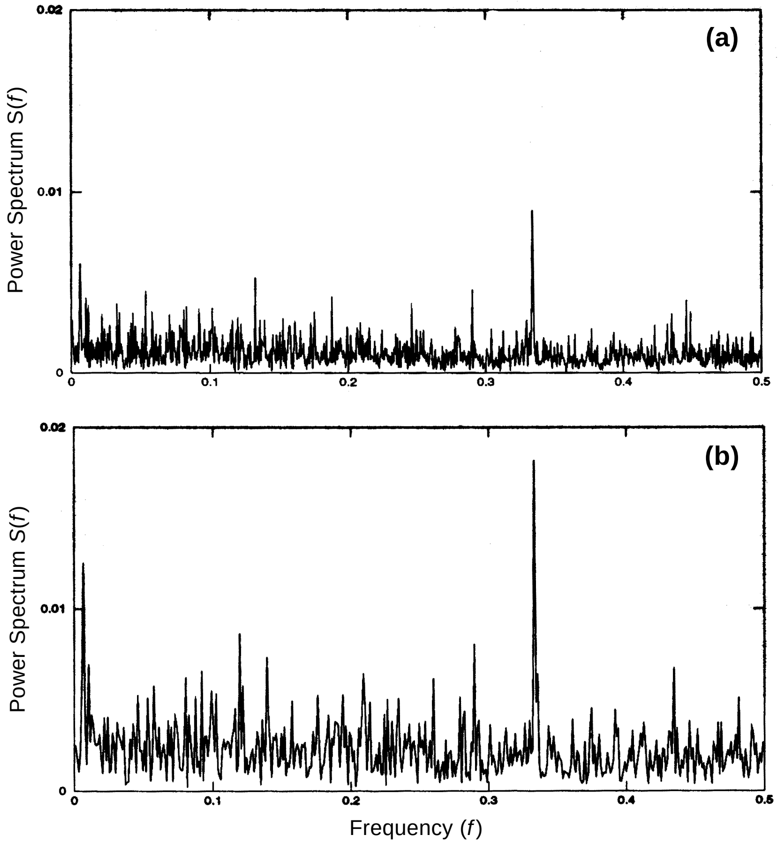

Figure 3. Fourier spectra for FP and FN predictions by both GeneScan and GLIMMER programs in region R2 of *C. jejuni*. Spectrum for (**a**) FN (CJ0672, CDS 627316 . . 629059) with $P_N = 12{\cdot}359$ and (**b**) the False P (628193–629062) with $P_N = 22{\cdot}181$.

protein. Both GeneScan and GLIMMER do not predict this to be coding (a false negative). Figure 3a shows the Fourier spectrum for this sequence, indicating multiple peaks. The FP found by both programs, (628193 . . 629062, Fourier spectrum shown in figure 3b) largely overlaps with this FN, and our suggestion here is that the correct annotation for CJ0672 should be 628193 . . 629059.

To summarize, in this paper we have studied in detail the predictions of two *ab initio* methods against the existing annotations, predicted using GeneMark. The results clearly show that the two programs together predict essentially all the genes correctly. There are, however, a few common False Negative predictions which we believe require re-examination, and the common False Positive predictions indicate new regions in the three genomes which are potential locations of new genes. A full list of these results is available at the website http://202.41.10.146/newgenes.html. Since the potential new coding regions have been specifically located, we hope this will help in the design of suitable experiments in order to verify the predictions made here.

Acknowledgment

This research is supported by a grant from the Department of Biotechnology, New Delhi.

References

Audic S and Claverie J M 1998 Self-identification of protein-coding regions in microbial genomes; *Proc. Natl. Acad. Sci. USA* **95** 10026–10031

Bhattacharya A, Sudha Bhattacharyya and John P Ackers 1999 Nontranslated polyadenylated ribonucleic acids from the protozoan parasite *E. histolytica*; *Curr. Sci.* **77** 564–567

Bhattacharya A, Bhattacharya S, Joshi A, Ramachandran S and Ramaswamy R 2000 Identification of Parasitic Genes by Computational Methods; *Parasitol. Today* **16** 127–130

Borodovsky M and McIninch J 1993 GeneMark: Parallel Gene Recognition for both DNA Strands; *Comput. Chem.* **17** 123–133

Burge C and Karlin S 1997 Prediction of complete gene structures in human genomic DNA; *J. Mol. Biol.* **268** 78–94

Burset M and Guigo R 1996 Evaluation of Gene Structure Prediction Programs; *Genomics* **34** 353–367

Claverie J M 1997 Computational methods for identification of genes in vertebrate genomic sequences; *Hum. Mol. Genet.* **6** 1735–1744

Delcher A L, Hormon D, Kasif S, White O and Salzberg S L 1999 Improved microbial gene identification with GLIMMER; *Nucleic Acids Res.* **27** 4636–4641

Dunham I *et al* 1999 The DNA sequence of human chromosome 22; *Nature* (*London*) **402** 489–495

Fickett J W 1996 The gene identification problem: an overview for developers; *Comput. Chem.* **20** 103–118

Guigo R 1999 DNA composition, codon usage and exon prediction; in *Genetics Databases* (ed.) M Bishop (New York: Academic Press) pp 53–80

Hattori M *et al* 2000 The DNA sequence of human chromosome 21; *Nature* (*London*) **405** 311–319

Lawson D, Bowman S and Bartell B 2000; *Nature* (*London*) **404** 34–35

Ossadnik S M, Buldyrev S V, Goldberger A L, Harvin S, Mantegna R N, Peng C K, Simons M and Stanley H E 1994 Correlation approach to identify coding regions in DNA sequences; *Biophys. J.* **67** 64–70

Parra S, Blanco E and Guigó R 2000 Geneid in Drosophila; *Genome Res.* **10** 511–515

Pertea M, Salzberg S L and Gardner M J 2000 Finding genes in *Plasmodium falciparum* chromosome 3; *Nature* (*London*) **404** 34

Ramachandran S and Ramakrishna R 1999 Gene identification in bacterial and organellar genomes using GeneScan; *Comput. Chem.* **23** 165–174

Tiwari S *et al* 1997 Prediction of probable genes by Fourier analysis of genomic sequences; *CABIOS* **13** 263–270

Uberbacher E C, Xu Y and Mural R J 1996 Discovering and understanding genes in human DNA sequence using GRAIL; *Methods Enzymol.* **266** 259–281

Vukimirovic O G and Tilghman S 2000 Exploring Genome Space; *Nature* (*London*) **405** 820–822

Xu Y and Uberbacher E C 1997 Automated Gene Identification in Large-Scale Genomic Sequences; *J. Comput. Biol.* **4** 325–338