PERGAMON

Computers & Chemistry 23 (1999) 165–174

Computers & Chemistry

Gene identification in bacterial and organellar genomes using GeneScan

Ramaswamy Ramakrishna [a,*], Ramachandran Srinivasan [a]

[a] *The Laboratory of Parasitic Diseases, National Institute of Allergy and Infectious Diseases, Bethesda, MD, 20892, USA*
[b] *School of Physical Sciences, Jawaharlal Nehru University, New Delhi, 110 067, India*

Received 23 June 1998; accepted 13 November 1998

Abstract

The performance of the GeneScan algorithm for gene identification has been improved by incorporation of a directed iterative scanning procedure. Application is made here to the cases of bacterial and organnellar genomes. The sensitivity of gene identification was 100% in *Plasmodium falciparum* plastid-like genome (35 kb) and in 98% in the *Mycoplasma genitalium* genome (~580 kb) and the *Haemophilus influenzae* Rd genome (~1.8 Mb). Sensitivity was found to improve in both the Open Reading Frames (ORFs) which have been identified as genes (by homology or by other methods) and those that are classified as hypothetical. False positive assignments (at the nucleotide level) were 0.25% in *H. influenzae* genome and 0.3% in *M. genitalium*. There were no false positive assignments in the plastid-like genome. The agreement between the GeneScan predictions and GeneMark predictions of putative ORFs was 97% in *M. genitalium* genome and 86% in *H. influenzae* genome. In terms of an exact match between predicted genes/ORFs and the annotation in the databank, GeneScan performance was evaluated to be between 72% and 90% in different genomes. We predict five putative ORFs that were not annotated earlier in the GenBank files for both *M. genitalium* and *H. influenzae* genomes. Our preliminary analysis of the newly sequenced G + C rich genome of *Mycobacterium tuberculosis* $H_{37}R_v$ also shows comparable sensitivity (99%).

Keywords: GeneScan; Mycoplasma; Plasmodium; Haemophilus; Fourier

1. Introduction

Given the enormous amount of information generated by genome sequencing programs, a large number of genes have been identified using computational methods. Analysis of whole genome sequences from various micro-organisms reveals that the functions of between a half and two-thirds of the existing genes can be identified using database searches (Pennisi, 1997). The remaining predictions are putative, and such open reading frames (ORFs) are classified as hypothetical.

The validity of these predictions can, in principle, be ascertained through 'functional genomic analysis' (Rastan and Beeley, 1997), although carrying out these experiments can be time-consuming and expensive (Fickett, 1994). It would be desirable to confirm the coding potential of a given DNA sequence as far as possible through the use of various computational methods, and those sequences which are deemed to be coding by several independent algorithms could then be assigned top priority for functional characterization. The primary aim of several gene identification

* Corresponding author. Tel.: 91-11-618-9701; fax: 91-11-619-8234; e-mail: rama@jnuniv.ernet.in

0097-8485/99/$ - see front matter
PII: S0097-8485(98)00034-5

algorithms that have been developed over the past decade (Guigo, 1998) is to locate exonic regions of a query DNA sequence. Some of the techniques further aim to fully identify genes by locating splice sites and promoter regions and also assign possible functions to the predicted gene.

We have recently described a method for gene identification that exploits the fact of a three-base periodicity which is present in coding DNA but is absent in non-coding DNA (Tiwari et al., 1997). This method, termed GeneScan, identifies this periodic feature by computing the Fourier signal of a given DNA sequence at frequency 1/3. This coding measure is universal in that it is equally applicable to genes in essentially all organisms. Although it is related to the positional asymmetry measure devised earlier by Fickett (1982) and Guigo (1998), it appears to be more discriminating in differentiating between coding and non-coding regions of the genome. The performance of the program was found to be comparable to several other programs (Tiwari et al., 1997). In application to the genomes of organisms with A + T rich base composition, we have observed that this method has very high potential of gene identification. The overall sensitivity was ~86% in a filarial nematode *Brugia malayi* (average 60% A + T), 100% in *Onchocerca volvulus* (average 60% A + T) and 99% in *Plasmodium falciparum* (average 70% A + T) (Bhattacharya et al., 1998). These organisms are targets of ongoing genome projects (Blaxter and Aslett, 1997).

In the present communication we describe the application of GeneScan to a plastid-like genome of *P. falciparum* and to two bacterial genomes *Mycoplasma genitalium* and *Haemophilus influenzae* Rd. Through a directed iterative scanning procedure (DIS), the overall sensitivity of gene identification with GeneScan has been improved for bacterial and organnellar genomes.

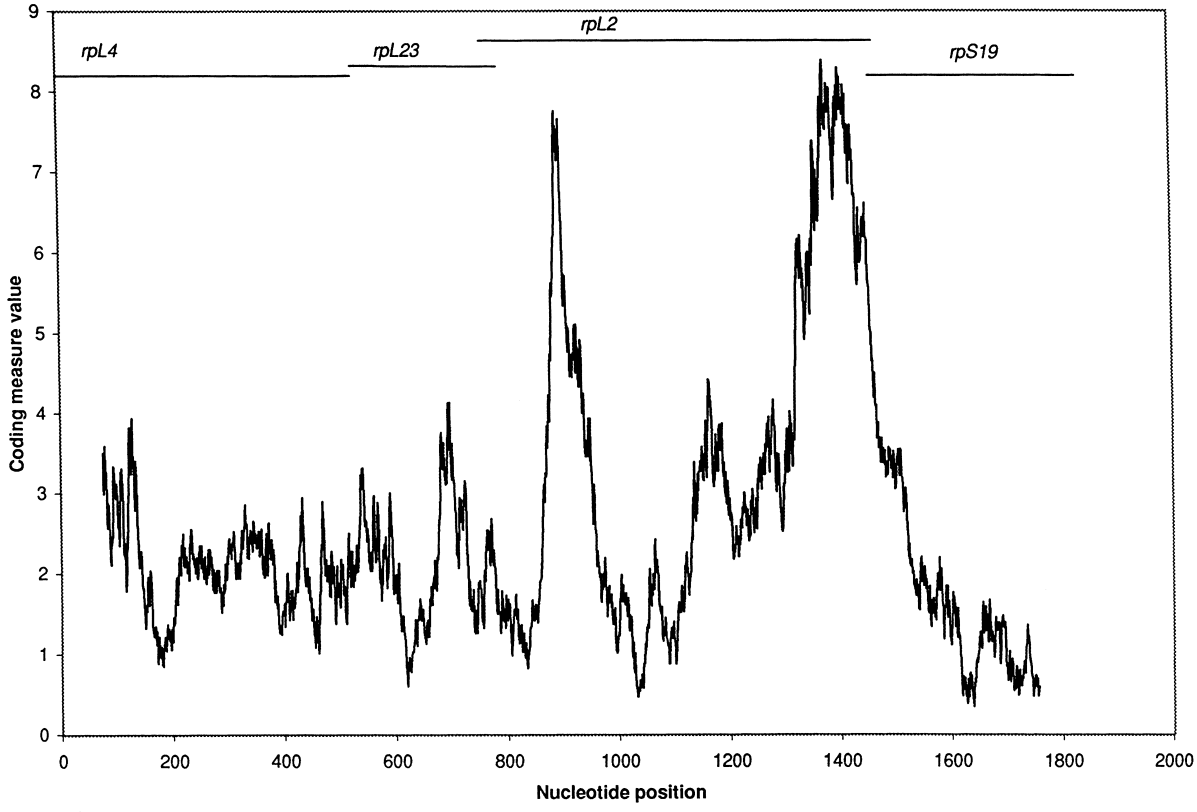

Fig. 1. Graphical representation of GeneScan of a portion of the *P. falciparum* 35 kb plastid-like genomic sequence. The strength of auto-correlation varies along the coding sequence, indicative of a mosaic structure. Scan parameters were window size = 150 bases, step size = 1. This portion of the circular genome contains genes coding for ribosomal proteins. The positions of each gene are indicated in the figure. The gene *rpS19* was identified in the directed iterative scanning at lower threshold.

2. Methods

The basic GeneScan algorithm has been described in detail (Tiwari et al., 1997). The numerical value of the coding measure—or strength of auto-correlation—calculated in a sliding window along the DNA sequence varies sharply along the sequence, indicative of an underlying mosaic structure (Fig. 1). Fragments of the sequence wherein the measure exceeds a set threshold are considered as potentially coding and form the basis of ORF determination.

To determine the start codon, if multiple ATG codons are present at the 5′end of the fragment then the ATG nearest to the fragment is considered as the start. The sequence is then scanned to locate an in-frame stop codon. We have also considered the possibility of alternate start codons (GTG or TTG) in some cases where a coding region was indicated, but no ATG could be located; these are described in the results.

Coding regions may yet be missed. A small proportion of genes 'invisible' to this technique lacks the three-base periodicity either in part or completely. In these cases, the strength of the auto-correlation as measured by GeneScan falls below the threshold used, and simple application of the method is unsatisfactory. To partially alleviate this problem, we introduce an iterative scanning procedure specifically directed to re-examine those sequence portions of the genome that would have been otherwise declared incorrectly as non-coding.

In this *directed iterative scanning* (DIS), the sequence portions are re-scanned at a lower threshold if the following criteria are met. The sequence must be at least 200 bases in length. There should be some indication of coding potential, namely if some (even small) fraction of the sequence showed significant correlation structure by having the signal exceed the higher threshold. Finally, the sequence should not contain any gene coding for ribosomal RNA (rRNA), transfer RNA (tRNA) or any other structural RNA molecules. By re-scanning at lower thresholds, the fraction of the sequence that was positive in the previous step is extended into longer fragments. Additionally, a new fragment can occasionally appear. Once again, fragments in which the coding measure scores above the threshold are considered as potentially coding and form the basis of ORF determination. This procedure is carried out iteratively as shown in the flow chart in Fig. 2. It is not profitable to lower the threshold for gene identification below the numerical value 2.0 because the limiting value for random sequences is 1.0.

As microbial genomes are compact, several genes/ORFs can be located in a coding fragment identified by GeneScan. We locate all possible ORFs within each coding fragment by searching for the appropriate start and stop codons. When comparison is made with annotated genomes, a gene/ORF is considered to be correctly predicted if and only if there is exact match between the result of the above procedure and the annotation given in GenBank. The percentage of false positives is calculated as (the number of nucleotides in the false positive fragment/total length of the genome)×100.

2.1. Sequences

The sequences of *M. genitalium*, *P. falciparum*, and *H. influenzae* Rd genomes used in the present work were retrieved from GenBank. Open reading frames in *M. genitalium* sequences were inferred using the *Mycoplasma* codon usage table (Fraser et al., 1995) and in *H. influenzae* sequences were inferred using the Translation Table 11 (Fleischmann et al., 1995).

2.2. Database search

ORFs identified as coding by GeneScan were submitted for database search using the computer program FASTA3 (Pearson and Lipman, 1988) with SWISS-PROT; default parameters were used. Similarities were evaluated as described in the documentation file and taken as significant if a homologous protein was found in other organisms. If no homologous protein was found in SWISS-PROT, then the program TFASTA (Wisconsin Package Version 9.1, Genetics Computer Group, Madison, Wisconsin) was used to verify whether the identified protein was unique. ORFs that showed a definitive homology with existing sequences in the database were classified as very likely genes; otherwise the predicted ORF was termed hypothetical.

3. Results and discussion

3.1. Directed iterative scanning

The results of application of directed iterative scanning using GeneScan to the 35 kb plastid-like genome of *P. falciparum* (Wilson et al., 1996), the ∼580 kb genome of *M. genitalium* (Fraser et al., 1995), and the ∼1.8 Mb genome of *H. influenzae* Rd are summarized in Table 1Table 2. The results show that the overall sensitivity of gene identification using GeneScan is high. Those ORFs that have been definitely identified (either through products or via homology) and those that are hypothetical, conserved hypothetical and putative ORFs identified by GeneMark are discussed separately. The term conserved hypothetical protein that appeared in the revised annotation is preferable to the nomenclature 'hypothetical protein' or 'putative pro-

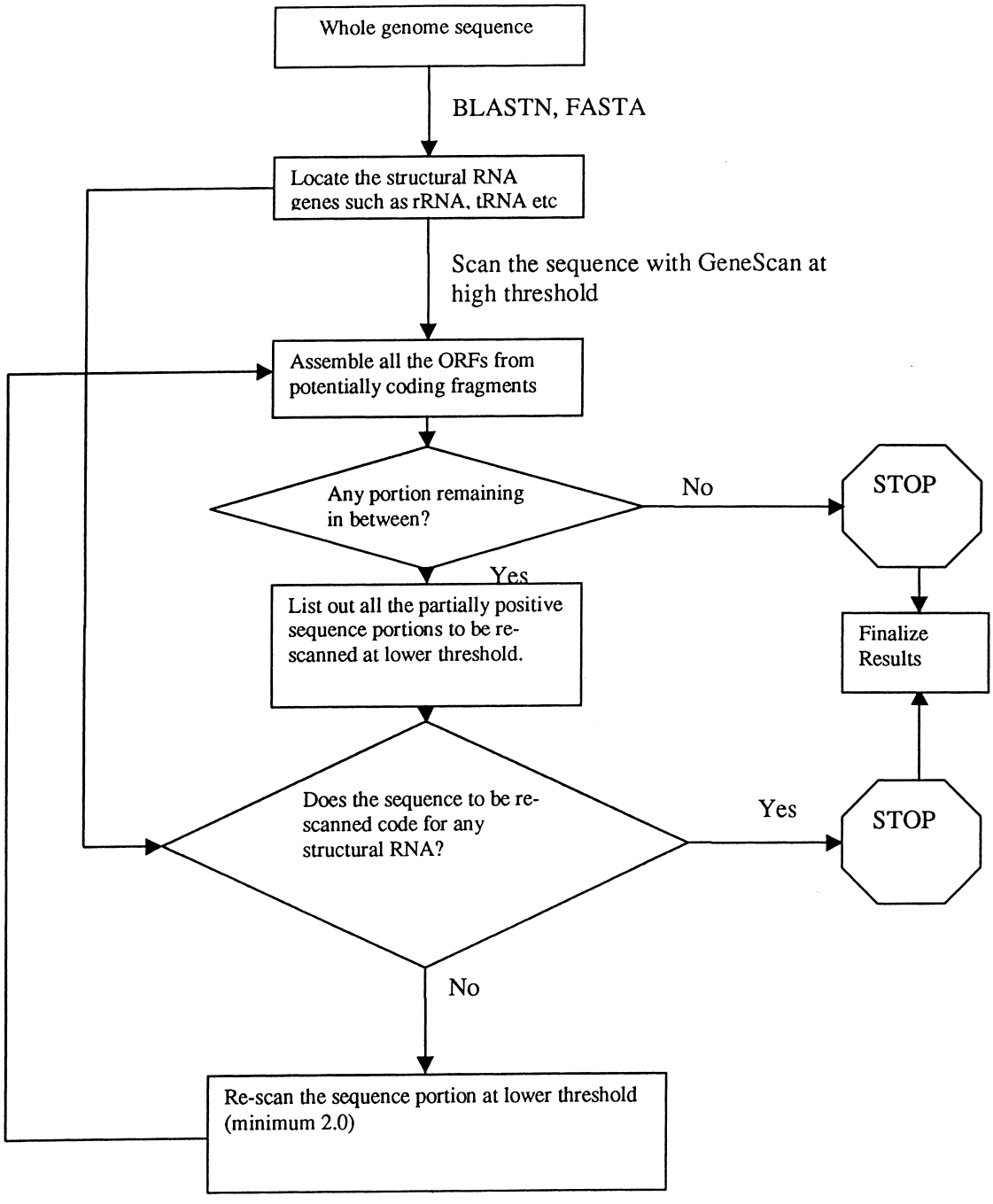

Fig. 2. Flow diagram of GeneScan incorporating the directed iterative scanning procedure.

tein' because these proteins are the most interesting in light of functional genomics.

Our main result here is that the sensitivity of GeneScan, namely the identification of coding ORFs can be very high (sensitivity of 1.0 is achieved for the case of the *P. falciparum* organnellar genome).

Many genes that are invisible to the basic GeneScan algorithm, namely the small proportion of genes that do not have a striking three-base periodicity, are now detected with directed iterative scanning. The sensitivity of gene identification was 100% in the plastid-like genome of *P. falciparum*, 98% in the genome of

Table 1
Results of gene identification through GeneScan. The threshold in the iterative scan was varied down to 2.0

ORF type	*P. falciparum* 35 kb circular plastid-like genome	*M. genitalium* whole genome
Number of known genes	22	330
Number of these genes identified	22	322
Number of known hypothetical ORFs	8	56
Number of these hypothetical ORFs identified	4	54
Number of putative ORFs identified by GeneMark	–	99
Number of these ORFs identified	–	97

Table 2
GeneScan results from *H. influenzae* genome

ORF type	*H. influenzae* whole genome
Number of known genes	1070
Number of these genes identified	1052
Number of known conserved hypothetical protein coding genes	311
Number of these conserved hypothetical protein coding genes identified	298
Number of known hypothetical ORFs	106
Number of these hypothetical ORFs identified	98
Number of putative ORFs identified by GeneMark	226
Number of these ORFs identified	192

Table 3
Genes that were not identified by GeneScan in different genomes

Genome	Functional category	Total number of genes in the respective category	Number of genes missed in that category
M. genitalium whole genome	Protein modification and translation factors	14	1
	Aminoacyl tRNA synthetases and tRNA modification	24	1
	Ribosomal proteins:synthesis and modification	52	6
H. influenzae whole genome	Phage related functions and prophages	11	2
	DNA replication, recombination, repair	75	2
	Cofactors, biotin	7	1
	Cofactors, heme and porphyrin	8	1
	Protein modification	24	1
	Cell envelope, surface polysaccharides, lipopolysaccharides and antigens	32	1
	Cell division	16	1
	Central intermediary metabolism, subcategory 'other'	13	1
	Cationic transport	24	1
	Anionic transport	8	1
	Regulatory	63	2
	Electron transport	9	1
	Protein & peptide secretion	15	1
	Ribosomal proteins:synthesis and modification	54	1

M. genitalium and 98% in the genome of *H. influenzae*. The sensitivity of identification of conserved hypothetical protein coding genes in *H. influenzae* was 96%. The sensitivity of identification of hypothetical ORFs in *M. genitalium* was 96% and 92.5% *H. influenzae* genome. Putative ORFs predicted by GeneMark were detected at the sensitivity level of 95% in the *M. genitalium* genome and 86% in the *H. influenzae* genome. The lowest level of sensitivity of identification was observed for GeneMark predictions even though in this category also sensitivity increases with the application of DIS to the basic algorithm. From Tables 1 and 2 this level of sensitivity appears to vary from one genome to the other. This could be due to the peculiarities of each genome. As these predictions were made purely on a computational basis, we may expect such levels of disagreement in this category of ORFs between GeneScan and GeneMark. Overall, it appears that the GeneScan is able to detect most of the genes whose functions can be inferred from homology search and the genes whose functional role cannot be predicted but are phylogenetically conserved.

The specificity, namely the ability to correctly identify noncoding regions as noncoding is also high: GeneScan identifies very few *false positives*. Indeed, in the genomes studied here, virtually all the tRNA genes and ribosomal RNA genes were uniformly negative and the few false positives (0.3% in *M. genitalium* and 0.25% in *H. influenzae*) were largely due to repeats.

3.2. Missed genes

We analyzed the genes that were not identified even with DIS modification to GeneScan. These have been summarized in Table 3. In *M. genitalium* genome, most invisible genes code for ribosomal proteins; one gene codes for a translation initiation factor IF3 (infC) and another for a peptidyl tRNA hydrolase homolog. Thus, all the genes that were missed belong to the functional category of translation. In *H. influenzae* genome, the genes that were missed belong to diverse functional categories and no systematics is apparent. In our earlier work, we found that GeneScan did not identify the mating type genes in *Saccharomyces cerevisiae* (Tiwari et al., 1997). It has been noted by Sharp et al. (1986) that some ribosomal protein genes fall as outliers to 'high' codon bias grouping in yeast. Because the three-period feature detected by GeneScan is contributed in part by codon bias arising from biased use of certain codons (Tiwari et al., 1997), it is possible that the reduced level of (or lack of) codon bias in some ribosomal protein coding genes, renders them invisible to GeneScan. We are currently investigating these aspects for other missed genes in other genomes.

3.3. Correct predictions and annotations

It is necessary to assess the accuracy of computational predictions of genes when an *ab initio* method such as GeneScan is used to analyze a fresh genomic sequence. For many microbial genomes, the algorithm that has hitherto been employed is GeneMark. Comparison with existing annotation, for example, shows that in the plastid-like genome GeneScan predicted about 90% of the genes and 50% of hypothetical ORFs correctly with the overall accuracy of 80% (see Table 4). In *M. genitalium* about 87% of the genes and putative ORFs and 72% of hypothetical ORFs were predicted correctly with the overall accuracy of

Table 4
Fraction of genes and other ORFs predicted correctly by GeneScan

Genome; overall accuracy of prediction[a]	Genes annotated; accuracy	Conserved hypothetical protein coding genes annotated; accuracy	Hypothetical ORFs annotated; accuracy	Putative ORFs identified by GeneMark; accuracy
P. falciparum[b] plastid-like genome; 80%	22; 91%		8; 50%	
M. genitalium[c]; 85%	330; 87.5%	–	56; 72%	99; 88%
H. influenzae[d]; 72%	1070; 77%	311; 72%	106; 63%	226; 51%

[a] The quoted accuracy is the fraction of genes or ORFs annotated in the respective genomes whose start site and stop site matched the predictions from GeneScan. Values are expressed in percentages. The procedure for start codon prediction is described in the Methods section. Out of the remaining fraction, the genes or ORFs that were detected were not correctly predicted in terms of start codon match.[b]Comparisons with annotated files dated 14 February, 1997.[c]Comparisons with annotation files dated 21 April, 1996.[d]Comparisons with annotation files dated 27 September, 1996. A few genes and ORFs in the *Haemophilus* genome as described in Table 5 were not evaluated for correct predictions as they appeared in the revised annotation

85%. In the case of the *H. influenzae* genome, about 75% of the genes and conserved hypothetical protein coding genes were predicted correctly. The accuracy levels in the categories of hypothetical ORFs and of putative ORFs predicted by GeneMark were 63% and 51% respectively. The overall accuracy in the *Haemophilus* genome was 72%. This compares favourably with other programs such as GeneMark.hmm (Lushakin and Borodovsky, 1998). In *H. influenzae* genome, several coding regions predicted by GeneScan were at variance with the annotation released on September 27, 1996. These are described in Table 5. In some cases it was not clear whether there were genuine ORFs. The revised annotation of 29 May, 1998, which includes putative ORFs identified by GeneMark and those in which possible frameshift errors were verified, is now in closer agreement with our results. One gene (*lic-3*) has not yet been annotated, although it appears that this was independently found using a tetramer DNA repeat search procedure (Hood et al., 1996). For one gene coding for a *HybG* protein, no information was found in the revised annotation. Taken together, it appears that GeneScan application with DIS could be very rewarding in gene annotation of fresh sequences.

3.4. Additional genes/ORFs

GeneScan does, in addition, give independent novel information. The high specificity and sensitivity of GeneScan gives additional support to the analysis of all putatively identified coding regions, and as our results in Tables 6 and 7 indicate, this could reveal new coding ORFs. There was no annotation in GenBank files corresponding to these predictions from GeneScan. In the *M. genitalium* genome, in one case the gene coding for Formamidopyrimidine-DNA gly-

Table 5
GeneScan predictions and their agreement with the revised annotation in GenBank files of *H. influenzae* Rd genome

ORF number	Type of GeneScan prediction[a] and its characteristics, if any	Revised annotation details[b]
HI0686	A; probably *glpT*	Length of *glpT* gene corrected
Between HI0352 and HI0354	B; ORF of 56 amino acids, probably *lic-3* gene	Annotation absent in GenBank file; this ORF was identified independently by Hood et al., 1996 by a 'DNA repeat search' procedure
Between HI1739.1 and HI1740	B; probably gene for glutamate racemase	Annotated in the new file
Between HI1462 and HI1463	B	Annotated as HI1462.1 (GeneMark ORF)and HI1462.2(CHP)[c]
Between HI1436 and HI1437	B	HI1436.2 (GeneMark ORF) and pseudogene (authentic frameshift)
5′ to (reverse strand) HI0220	B	HI0220.2 (GeneMark ORF)
Between HI0221 and HI0222	B	HI0221.1 (brute force ORF identified by GeneMark)
Between HI0602 and HI0603	B	HI0602.1
5′ to (reverse strand) HI0485	B	HI0485.1 (brute force ORF identified by GeneMark)
Between HI0976 and HI0977	B	HI0976.1 (CHP)
Between HI1225 and HI1227	B	HI1225.1 (CHP)
Between HI0213 and HI0214	B	HI0213.1 (brute force ORF identified by GeneMark)
Between HI0559 and HI0561	B	HI0559.1 (brute force ORF identified by GeneMark)
5′ to HI 1467	B	HI1466.1 (brute force ORF identified by GeneMark)
HI1390	A; probably *HybG*	Annotation not revised

[a] Type of prediction, A, B: if the GeneScan predictions indicated that a given gene or ORF was longer than that described in the annotation, it is referred to as type A. In all cases described above the predictions suggested a longer length at the 3′ end. On the other hand, if predictions indicated a new ORF or gene, then it is referred to as type B. These genes and ORFs were not evaluated in terms of correct predictions as they appeared in the revised annotations.[b]The GeneScan analyses were carried out with the files released from GenBank on 27 September, 1996. (The revised annotations appeared on 29 May, 1998, when this work was in progress.)[c]CHP is abbreviation for Conserved Hypothetical Protein. The term conserved hypothetical protein appeared in the revised annotation. We prefer this term to 'hypothetical protein' or 'putative protein' because this will be most interesting in light of functional genomics

Table 6
Additional genes and/or putative ORFs predicted by GeneScan and comparisons with the database annotations in the genome of *M. genitalium*[a]

Position in the genome	Direction of predicted ORF relative to the neighbouring ones	Comment
Section 1 of 56 of complete genome, 5′ to gene *dna_N* (MG001)	Same as MG001	Existing annotation: 1026–1829. Positions 727 onwards is highly similar to the homologous protein from *Mycoplasma pneumoniae*
Section 21 of 56 of complete genome, 3′ to ORF MG218	Same as MG218	Annotation 6662– > 7184. An ORF of a minimum of 174 amino acids[b]. Homologous ORF found in *M. pneumoniaea*[c].
Section 27 of 56 of complete genome, between gene MG262 and ORF MG263	Same as MG262	Annotation 5183–6035. Gene coding for formamidopyrimidine-DNA glycosylase. Data present in SWISS-PROT as MG262.1; data not present in GenBank file. Identified by similarity to the homologous protein from various bacteria
Section 28 of 56 of complete genome, between gene MG269 and ORF MG270	Same as MG269 and MG270	Annotation: complement (1235–1504). An ORF of 89 amino acids. Homologous ORF found in *M. pneumoniaea*[c].
Section 37 of 56 of complete genome, before gene MG324	Same as MG324	Annotation complement (67–831). An ORF of 254 amino acids, with GTG (Valine) as start codon. No homology to any known gene or ORF. Homologous ORF found in *M. pneumoniaea*.
Section 38 of 56 of complete genome. Between ORF MG335 and gene MG336	Same as MG336	Annotation: 3518–4543. An ORF of 341 amino acids. Homologous ORF found in *M. pneumoniaea*[c].
Section 41 of 56 of complete genome, between MG350 and MG351	Opposite to both MG350 and MG351	Annotation: complement (2599–3273). An ORF of 224 amino acids. Homologous ORF found in *M. pneumoniaea*[c].
Section 49 of 56 of complete genome, between ORF MG415 and MG417	Same as MG415	Existing annotation of MG415: complement (3262–4080). Probable annotation of MG415: complement (3262–5256). The inferred protein sequence between 3262 and 4080 is in a continuous frame with the larger protein 3262–5256. The existing annotation is from an internal methionine.

[a] ORFs with start codon other than ATG are indicated in the tables. If no mention is made about the start codon, then the start codon is ATG.[b]These homologues were found using the TFASTA program.[c]Sequence file ends within the ORF

cosylase is absent in GenBank although it is present as MG262.1 in SWISS-PROT. In the second case, GeneScan prediction suggested alternative initiation sites of a hypothetical ORF predicted using GeneMark. In the third case, the revised annotation in the database of The Institute of Genomic Research (TIGR) of *dna_N* gene, agrees with the GeneScan predictions. The remaining putative ORFs predicted in the genome of *M. genitalium*, did not have any homologous protein in the SWISS-PROT. Using TFASTA, we found that they all had homologues in the genome of *Mycoplasma pneumoniae*. Similar observations were made independently while analyzing the *M. pneumoniae* genome (Himmelreich et al., 1997). In the genome of *H. influenzae*, all five putative ORFs were smaller than 100 amino acids length. No homologue was found for these ORFs in the SWISS-PROT database and by TFASTA search. Two of these putative ORFs did not have any motif resembling the purine rich Shine-Dalgarno sequence within 20 bases upstream of

Table 7
Putative ORFs predicted by GeneScan in the genome of *H. influenzae*[a]

Position in the genome	Direction of predicted ORF relative to the neighbouring ones	Comment
Section 141, between ORFs HI1484 and HI1485	Same as HI1484	Annotation: 1060–1329. ORF of 89 amino acids; GTG start codon; Shine-Dalgarno motif at eight bases upstream of start codon; has 20 amino acid hydrophobic region; no homology to any known protein
Section 10, between ORFs HI0096 and HI0097	Opposite to HI0096	Annotation: complement (3826–3987). ORF of 53 amino acids; GTG as start codon; purine rich sequence at seven bases upstream with partial similarity to Shine-Dalgarno motif; no homology to any known protein
Section 153, between ORFs HI1651 and HI1652	Same as HI1652	Annotation: complement (2814–3023). ORF of 69 amino acids; similarity to HI1523 protein; no purine rich sequence motif within 20 bases upstream
Section 6, between ORFs HI0062 and HI0063	Same as HI0062	Annotation: 8745–8987. ORF of 80 amino acids. No purine rich sequence within 20 bases upstream of start codon; no homology to any known protein
Section 110, between ORFs HI1152 and HI1153	Opposite to HI1152	Annotation: complement (1427–1624). ORF of 65 amino acids; purine rich sequence with partial similarity to Shine-Dalgarno motif at seven bases upstream of the start codon; no homology to any known protein

[a] ORFs with start codon other than ATG are indicated in the tables. If no mention is made about the start codon, then the start codon is ATG

the putative start codon. After the ribosome binding site matrices from these organisms become available, these predictions can be re-evaluated. It would be interesting to investigate whether these ORFs are expressed and if so, whether they are expressed constitutively or under conditional stimulation. Based on the results described in the previous sections we would place about 97% confidence on these novel ORFs.

4. Summary and conclusions

We have shown that the GeneScan algorithm, with the incorporation of a directed iterative scanning procedure, can achieve a high sensitivity in *ab initio* gene prediction. Obtaining sensitivity levels of 98% is typical and the algorithm also has the potential to reveal new ORFs or genes. Although the genomes analyzed in this paper are biased in A + T, we found that on scanning about 5% of the G + C rich genome of *Mycobacterium tuberculosis* $H_{37}R_v$ a sensitivity level of ~99% could be obtained. Thus, it appears that high level of sensitivity is achievable in any genome. The group of genes invisible to the algorithm varies from one organism to the other. Between 72 and 90% of the genes/ORFs from different genomes can be correctly predicted using GeneScan. The number of false positives is very low. Taken together it appears that the application of GeneScan (in this modified form) to fresh genomic sequences could be rewarding. The methodology is general and is therefore applicable to any bacterial and organellar genome. Scanning of the rest of sequenced genomes is currently being carried out to extend the application. The DIS technique with GeneScan, as applied here, is suitable only for bacterial (and organellar) genomes where several genes are expressed as polycistronic messenger RNAs and some genes could even be overlapping. For eukaryotes, implementation of a DIS will require identification of potential splice sites using existing gene information and splice site prediction programs. This work is also currently in progress.

Acknowledgements

S. R. is grateful to Dr Franklin A. Neva for support, and the NCBI for computational services and The Institute of Genomic Research (TIGR) for access to their databases through the internet. R. R. would like to thank Alok and Sudha Bhattacharya for discussions and the Department of BioTechnology for support.

References

Bhattacharya, A., Bhattacharya, S., Joshi, A., Ramachandran, S., Ramaswamy, R. 1998. Parasitology Today. (submitted) .

Blaxter, M., Aslett, M., 1997. Internet resources for parasite genome projects. Trends in Genetics 13, 40.

Borodovsky, M.Y., McIninch, J.D., 1993. GeneMark: parallel gene recognition for both strands. Computers and Chemistry 17, 123.

Fickett, J.W., 1982. Recognition of protein coding regions in DNA sequences. Nucleic Acids Research 10, 5303.

Fickett, J.W., 1994. Inferring genes from open reading frames. Computers and Chemistry 18, 203.

Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A. et al, 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269, 496.

Fraser, C.M., Gocayne, D.J., White, O., Adams, M.D. et al, 1995. The minimal gene complement of *Mycoplasma genitalium*. Science 270, 397.

Guigo, R. 1998. DNA composition, codon usage and exon prediction. (in press) .

Himmelreich, R., Plagens, H., Hilbert, H., Reiner, B., Herrmann, R., 1997. Comparative analysis of the genomes of the bacteria *Mycoplasma pneumoniae* and *Mycoplasma genitalium*. Nucleic Acids Research 25, 701.

Hood, D.W., Deadman, M.E., Jennings, M.P., Biseric, M., Fleischmann, R.D., Venter, J.C., Moxon, E.R., 1996. DNA repeats identify novel virulence genes in *Haemophilus influenzae*. Proceedings of National Academy of Sciences USA 93, 11121.

Lushakin, A.V., Borodovsky, M., 1998. GeneMark.hmm: new solutions to gene finding. Nucleic Acids Research 26, 1107.

Pearson, W.R., Lipman, D.J., 1988. Improved tools for biological sequence comparison. Proceedings of National Academy of Science U.S.A. 85, 2444.

Pennisi, E., 1997. Microbial genomes come tumbling in. Science 277, 1433.

Rastan, S., Beeley, L.J., 1997. Functional genomics: going forwards from the databases. Current Opinion in Genetics and Development 7, 777.

Sharp, P.M., Tuohy, T.M.F., Mosurski, K.R., 1986. Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. Nucleic Acids Research 14, 5125.

Tiwari, S., Ramachandran, S., Bhattacharya, A., Bhattacharya, S., Ramaswamy, R., 1997. Prediction of probable genes by Fourier analysis of genomic sequences. Computer Application in Biosciences 13, 263.

Wilson, R.J.M., Denny, P.W., Preiser, P.R. et al, 1996. Complete gene map of the plastid-like DNA of the malaria parasite *Plasmodium falciparum*. Journal of Molecular Biology 261, 155.