An automated annotation tool for genomic DNA sequences using GeneScan and BLAST

ANDREW M. LYNN, CHAKRESH KUMAR JAIN, K. KOSALAI, PRANJAN BARMAN, NUPUR THAKUR, HARISH BATRA and ALOK BHATTACHARYA*†

*Bioinformatics Centre, and *School of Life Sciences, Jawaharlal Nehru University, New Delhi 110 067, India*

Abstract

Genomic sequence data are often available well before the annotated sequence is published. We present a method for analysis of genomic DNA to identify coding sequences using the GeneScan algorithm and characterize these resultant sequences by BLAST. The routines are used to develop a system for automated annotation of genome DNA sequences.

[Lynn A. M., Jain C. K., Kosalai K., Barman P., Thakur N., Batra H. and Bhattacharya A. 2001 An automated annotation tool for genomic DNA sequences using GeneScan and BLAST. *J. Genet.* **80**, 9–16]

Introduction

DNA sequencing has evolved from a complicated laboratory process to an automated technique using high-throughput sequencers with fluorescent-dye-based chemistry. This technological advance coupled with the replacement of the traditional mapping and sequencing of clones in series to an integrated simultaneous mapping and sequencing approach—'shotgun' genome sequencing (Fleischmann *et al.* 1995)—has significantly reduced the amount of time it takes to sequence a genome (Lee and Lee 2000). Large-scale genome sequencing generates raw sequence data. Annotation is the process of interpreting this data into useful biological information. Preliminary annotation involves detection and description in the sequence of features like the location of protein-coding genes, their structure (the demarcation of control regions, exons, introns and untranslated regions), the location of repetitive sequences and their nature, and the location of genes encoding noncoding RNA. Further annotation involves characterization of these sequence elements in terms of their relationship with other sequence elements within the genome and in other genomes, and the prediction of structural and functional attributes traditionally on the basis of homology. Interpreting the domain fold and attributing function to predicted protein sequences on the basis of sequence comparison with sequences of known proteins are examples of this type of characterization.

Computational gene identification in genomic DNA sequences is the normal start point in creating an inventory of genes. Methods for *ab initio* gene identification can be classified into two types (Haussler 1998): signal sensors are methods that detect local sites such as start and stop codons, branch points, splice sites, promoters and terminators of transcription, polyadenylation sites and ribosomal binding sites, while methods that use nucleotide frequencies and dependencies that help differentiate between coding and noncoding sequences are called content sensors. In the past five years, several systems that combine signal and content sensors have been developed in an attempt to identify complete gene structure.

Coding regions of DNA sequences have a strong three-base periodicity (Fickett 1982). Algorithms based on identifying this periodicity as a signal of protein-coding regions form the basis of programs like TESTCODE (Fickett 1982; Wisconson Package, GCG) and GeneScan (Tiwari *et al.* 1997).

†For correspondence. E-mail: alok@jnuniv.ernet.in.

Keywords. GeneScan; BLAST; genome annotation; DNA sequence; computational genomics; Fourier analysis.

The Fourier transform is a mathematical technique that essentially converts periodicity into a function of its inverse, the frequency. A three-base periodicity is represented as a peak at a frequency of 1/3. The frequency spectrum enables a more detailed assignment of the protein-coding region in two ways. The intensity of the peak at frequency = 1/3 is a measure of the extent of periodicity. The signal/noise ratio of this peak at frequency = 1/3 (P_n) was defined as a quantitative measure of this parameter. By analysis of the cumulative distribution of P_n with sample coding and noncoding DNA sequences, it is found that 95% of coding sequences have P_n value above 4.0, whereas 90% of noncoding regions have this value below 4.0 (Tiwari *et al*. 1997). The presence of other types of periodicity, represented by peaks at regions other than 1/3 on the frequency spectrum, is often a signal of repeat regions in the test sequence. We refer to this method of discriminating between coding and noncoding sequences as the GeneScan algorithm.

Gene finding in bacterial genomes is close to being a solved problem with a number of tools that give excellent results (for details see Bhattacharya *et al*. 2000). The recent adaptation of the GeneScan algorithm to bacterial and organellar genomes has displayed a sensitivity of 98% for the *Mycoplasma genitalium* and *Haemophilus influenzae* Rd genomes (Ramakrishna and Srinivasan 1999). The simplistic gene structure in bacteria—the presence of continuous open reading frames—is one of the reasons for such confidence in prediction. The presence of splice sites and the absence of clear definite rules that define these signals in eukaryotic genomes makes gene identification in these systems more difficult.

The most efficient eukaryotic gene identification systems integrate information from different signal sensors such as promoters, splice sites, start and stop codons and the 3′ untranslated regions with statistical properties from content sensors for coding regions. Since signal sensors often vary from gene to gene within organisms and between organisms, the program often has to be 'trained' on sequences that are best representative of gene structure within the organism. The choice of this representative set of genes is critical in determining the efficiency of gene identification.

A recent experiment to evaluate gene identification programs on a well-studied 2.9-Mb region of the *Drosophila* genome (Reese *et al*. 2000) showed disappointing results. In evaluating a program's efficiency, sensitivity (genes identified/genes reported) and specificity (number detected/number detected + false positives) are indices used to gauge the extent of correct prediction and overprediction of genes respectively. At the base level, the best programs reached a sensitivity of 95% and a specificity of 90% in identifying presence of a gene. However, prediction of explicit structure has poorer results. The average sensitivity and specificity of any program was 78% and ~ 50% at the exon level, and at the gene level the values were just over 60% and below 40% respectively.

One of GeneScan's most attractive features is the universal applicability of the algorithm, without the need for representative gene training sets. In the present implementation, we have avoided introducing signal sensors to map gene structure, relying on database comparisons to characterize the elements of coding sequences identified by the GeneScan window analysis. We use databases of both protein and EST (expressed sequence tag) sequences deposited in the public domain to infer information based on homology: amino acid similarity with a protein is used to identify the homologues of the gene, and EST matches imply experimental evidence for the expression of the sequence identified as a coding sequence.

This protocol can be used to generate the gene complement of an organism from genomic sequence data. The relevance of a list of genes can be appreciated by the impact on our understanding of the organism's biology from earlier published results.

Microbial genomes

Genome sequences reveal a gene complement and organization that reflects in detail the specific adaptations and lifestyles of the organism concerned. It is thought that the genome information will be useful to decipher the biology of the organism.

Pathogenic microbes

Two general features of pathogenic microbes are confirmed from the complete genome sequences of pathogenic microbes. First, virulence factors are encoded in clusters (so-called 'pathogenicity islands'), and the first comparisons of whole genomes show that these islands often differ substantially from the rest of the genome in such parameters as G + C content, codon usage and gene density, suggesting that they are relatively recent acquisitions that conferred pathogenicity to a relatively benign symbiont. Second, pathogens evade host immune response through variation of cell-surface antigens, which is due either to polymorphism in the genes encoding the protein or to paralogue expansions of a gene family. The mechanism and mapping of the genes responsible for this antigenic variation are a boost to vaccine development against the pathogen.

These two features are typically exemplified by the genome of *Neisseria meningitidis* (Parkhill *et al*. 2000), the causative agent of bacterial meningitis and septicaemia, where three islands of horizontal DNA transfer are identified, two of which contain genes coding for proteins known to be involved in pathogenicity, such as structural proteins of the pilus, and several coding regions unique to

capsular polysaccharide synthesis. The genome contains an abundance of diverse repeat sequences in each meningococcal strain, with the expression of about 65 genes altered by inaccurate DNA replication of repeat regions. Many of these genes encode antigens or cell-surface molecules involved in pathogenesis. *Mycobacterium tuberculosis* displays a metabolic potential to survive in a variety of environments (anaerobic in tissues). This is reflected to some extent in the gene list deduced from the complete genome sequence (Cole *et al.* 1998), which includes genes of lipid and carbohydrate metabolism to generate the complex structures of the cell wall, and potential virulence factors. The genome contains a high level of sequence conservation during replication though some polymorphic G + C-rich regions encode sets of products with peptide motifs which may be involved in antigenic variation.

The genome of *Borrelia borgdorferi* (Fraser *et al.* 1997), the recently discovered causative agent of Lyme disease, which affects both vertebrate and invertebrate hosts, was published before much of the organism's physiology was known. *Borrelia*'s linear chromosomes challenged the idea of a circular bacterial chromosome. Orthologues of normal virulence factors such as toxin and invasion genes, global regulatory genes, two-component signal transduction pathways and bacteriophages of other pathogenic bacteria were not identified, though duplicated lipoprotein genes, unique to *Borrelia* spp. and of unknown function, were. One member of this family, OspA, has been structurally characterized, and is currently being tested as a vaccine candidate. There are two circular chromosomes of *Vibrio cholerae* (Heidelberg *et al.* 2000). The gene list derived from the complete nucleotide sequence shows genes dedicated to essential cell functions such as DNA replication, cell division, gene transcription, protein translation and cell-wall biosynthesis along both chromosomes. Many of the genetic loci associated with virulence are however located on the larger chromosome. These include the cholera toxin, which is encoded in a phage sequence; the factor essential for colonization of the intestine, which is part of the *Vibrio* pathogenicity island; and genes encoding ToxR and ToxT, which regulate the expression of genes involved in virulence. Several essential proteins, like the ribosomal proteins L20 and L35, are present only on the smaller chromosome. The two chromosomes have 105 genes in common, allowing the opportunity to study interchromosomal recombination in bacteria.

Features of the genome that reflect the organism's adaptation to its environment are typified by *Helicobacter pylori* (Alm *et al.* 1999), which is one of the major causes of stomach ulcers. Proteins encoded by the genome contain double the lysine/arginine content of their normal orthologues from similar organisms, to maintain the high electropositive internal environment of the cell that enables it to survive the acid environment of the stomach. Also essential, the enzyme urease, which converts toxic urea into ammonia and carbon dioxide, seems conserved and is probably the result of horizontal gene transfer. The gene complement of *Xylella fastidiosa* (Simpson *et al.* 2000), a pathogen of citrus plants, reflects its life in the plant xylem in three different ways. First, the bacterium is adapted to use a variety of free sugars found in xylem sap and supplement these with glucose derived from the breakdown of cellulose. Second, a set of genes (67) are devoted to the uptake of iron and other transition metals from the xylem sap; depletion of these micronutrients causes symptoms of disease. Third, the organization produces two distinct types of cell adhesion proteins—one for a matrix of extracellular polysaccharides that embeds the bacterium in the matrix of the xylem, and the other for bacterial adhesion of the gut and mouth parts of the insect vector.

Rickettsia prowazekii (Andersson *et al.* 1998), the causative agent of tick-borne typhus, has 834 genes, and is an example of a highly reduced genome, getting many essential factors complemented by its host. It is of interest also because it belongs to a group of bacteria that are thought to be closest to the eukaryotic mitochondria. Comparisons between the two show that both lack genes for anaerobic glycolysis and amino acid and nucleotide synthesis, though the bacterium has a complete set of genes for the TCA cycle, while in mitochondria only a subset of the genes is found. The importance of finding genes from genome data is clear from these examples.

Microbial evolution

The availability of complete microbial genome sequences has made it possible to examine evolutionary relationships among living organisms in a more comprehensive way. Traditionally molecular evolution has been dominated by the calculation of the 'distance' between sequences from the differences on aligning small-subunit rRNA to yield a bifurcating tree structure. This assumes a linear evolution of organisms from a common ancestor. There is increasing evidence that genomes contain considerable portions that have arisen through genetic recombination from other organisms, mediated by viruses or other genetic elements —a process called lateral or horizontal transfer of DNA. Using whole-genome information, it has been possible to build an average phylogenetic tree on the basis of gene content. Other research has also identified an evolutionary 'core' of genes that code primarily for proteins involved in genome replication and expression. Specific metabolic functions are more sporadically present and may involve lateral transfer of genetic content (Nierman *et al.* 2000).

Eukaryotic genome sequences

The complete annotated genome sequences of *Saccharomyces cerevisiae* (Short *et al.* 1997), *Caenorhabditis elegans* (The *C. elegans* Sequencing Consortium 1998),

Drosophila melanogaster (Adams *et al*. 2000) and *Arabidopsis thaliana* (The *Arabidopsis* Genome Initiative 2000, and references therein) are now available. Yeast genetics is normally used as a model for eukaryotic genetics and major efforts are being made to understand the proteome. There are different approaches and one of them is the use of yeast two-hybrid techniques to determine the interacting patterns of the ~ 6000 hypothetical proteins in a bid to define their function. *C. elegans* and *D. melanogaster* are model systems to understand the genetic basis of multicellular organization, particularly organogenesis, behavioural response, and so on.

The genome annotation tools currently used for eukaryotic genomes are often found wanting in both sensitivity and specificity. A common factor is not to rely on individual programs for *ab initio* prediction but to use them in concert with database comparisons for gene identification (Lewis *et al*. 2000). The rest of this paper describes one of the annotation tools that we have developed and some examples of its application.

Methods

The GeneScan algorithm and Fourier transform

A DNA sequence can be converted to four binary sequences (U_a) for each base (a = A, T, C, G) by replacing the sequence with 1 for the occurrence and 0 for the absence of the base under consideration. The total Fourier spectrum can be calculated to determine the periodicity of the nucleotide by applying the Fourier transformation $S(f)$ on each binary string.

$$S(f_a) = \frac{1}{N^2}\left|\sum_{j=1}^{N} U_a(x_j)\exp 2\pi i f j\right|^2,$$

where N is the number of bases in the sequence, x_j the value of the jth position in the sequence, f the discrete frequency

$$= k/N, \text{ with } k = (1, 2, 3 \ldots N/2).$$

The total Fourier transform of the sequence is the sum of the individual transforms for each base

$$S(f) = \sum_a S(f_a).$$

The average of the Fourier spectrum can also be calculated from the frequency of occurrence, r_a, of the bases as

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

This value is used in evaluating the relative height of the peak at f = 1/3:

$$P_n = \frac{S(1/3)}{\bar{S}}.$$

A $P_n > 4$ is considered as a discriminator for coding sequences.

Structure and application of the program

I. ***GeneScan-based identification of coding regions*** (figure 1):

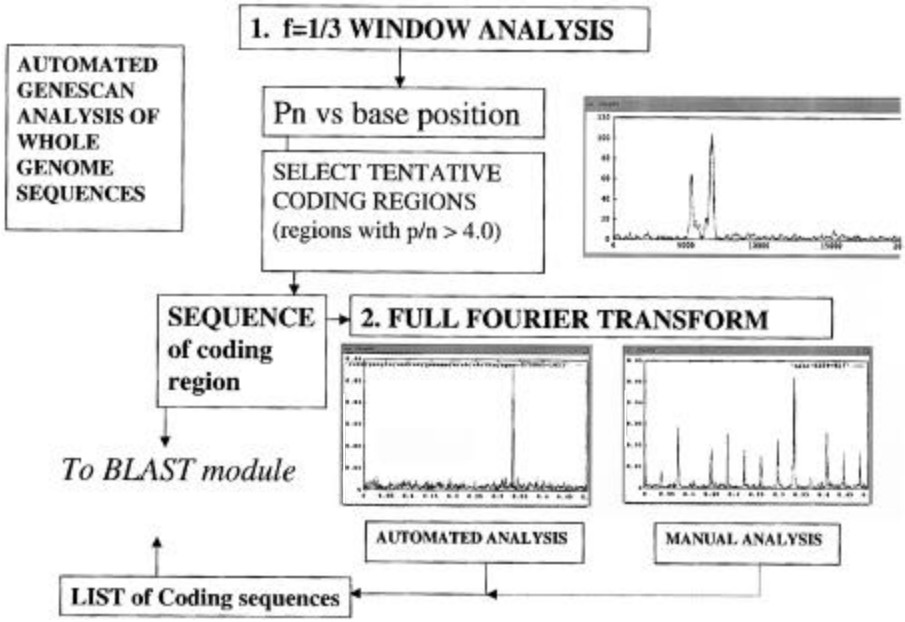

Figure 1. Gene identification using the GeneScan algorithm. The program requires as input genomic DNA sequences, and outputs a list of coding sequences, with the explicit sequence of each in a FASTA file, which is used in the next stage for BLAST analysis.

1. *Sequence input*: The chromosomal sequences were read into a one-dimensional array, the array number corresponding to the base position in the chromosome. Routines were written to read both FASTA and GENBANK formats. The program was compiled on a Silicon Graphics O2 with 64 MB memory running IRIX 6.3, and also on a Compaq Pentium III with 64 MB memory running RedHat Linux 6.2. For computers with less memory, the sequence can be written out as a random-access file and the record number used to trace the base position in the sequence for subsequent steps.

2. *Window analysis*: The primary step in the analysis is to identify regions in the chromosome with a P_n value above 4.0. The algorithm for the Fourier analysis of DNA sequences is applied to the sequence with a sliding window of 300 bases, to find the peak/noise ratio at a frequency = 1/3 (P_n). Since only the P_n is to be calculated, k is set equal to $N/3$ in the first equation above, requiring a single calculation to be made for each x_j. The spectrum is normalized to allow comparison between windows. The result of this routine is a series of data points in two columns: the first column is the start of the window, the second column is the P_n value of that window. The data is written into a file (sequence_name).win.

3. *Selection of tentative coding regions*: Selecting regions of the chromosome that have a peak/noise ratio above 4.0 is performed by reading the file sequentially and noting the sequence numbers at which the P_n value increases above, and subsequently decreases below, 4.0. Since these numbers correspond to the base position in the chromosome sequence, the potential coding sequence can be read out of the array containing the sequence with an addition of three hundred bases to compensate for the window length. Sequences of length less than 400 bases (corresponding to a continuous stretch of P_n less than 100) are listed but not processed to reduce false positives. The sequence file is saved to disk using a format A5-I_1-I_2, where A5 is a five-character code for the genome under study and I_1 and I_2 are integers that mark the start and length of coding sequence respectively.

4. *Complete Fourier analysis of the sequences*: These selected regions were then analysed by complete Fourier analysis. The Fourier algorithm is used in a loop for all frequencies from $1/N$ to $N/2$ (where N is the number of bases in the sequence). The spectrum is analysed for the presence of multiple peaks, and sorted. Sequences with single peaks or two peaks are selected for automated comparison against sequence databases. Those that have multiple peaks are listed separately for manual interpretation.

II. *BLAST analysis* (figure 2):

EST databases (dbest) are a valuable source for gene identification, a match across two or more exons providing definite clues to the presence of a gene. Sequence homology with a known protein is also useful in associating function with an unknown protein. We use the program BLAST (Altschul *et al.* 1997) to compare potential coding sequences against protein and EST sequence databases. The results are sorted on the basis of BLAST scores into categories as specified for the annotation of human chromosome 23.

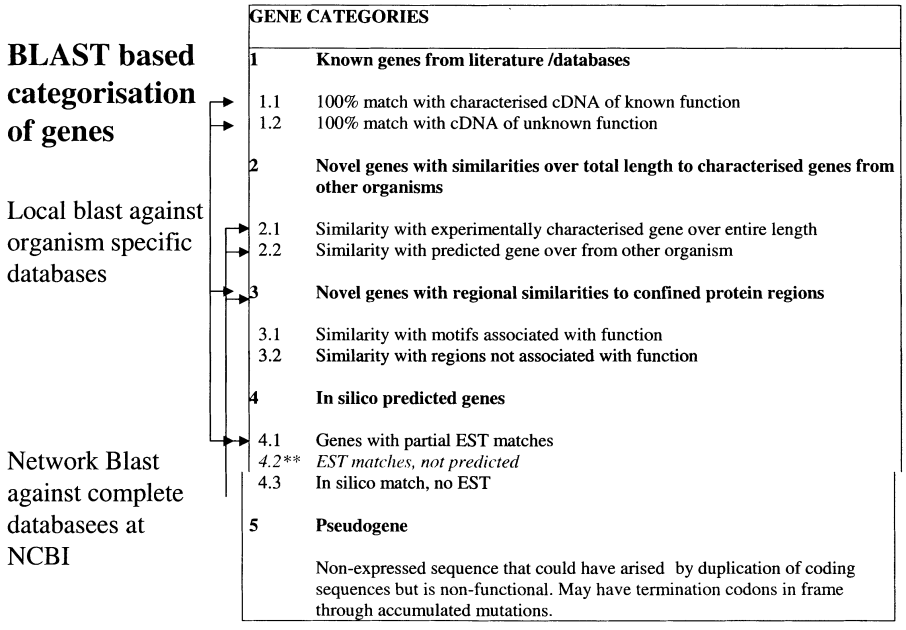

Figure 2. BLAST analysis and categorization.

BLAST is performed at two levels. Databases of all proteins and EST sequences of the organism under study are generated from the NCBI (National Center for Biotechnology Information, Bethesda) databases and the analysis is carried out on local computers. The results are then filtered for sequences that have a BLAST score larger than 100 and an e-value less than 10^{-5}. These proteins are selected tentatively into category 3, and can be upgraded on manual inspection of the BLAST results to category 1 (perfect matches along the entire sequence). The remaining sequences (tentatively categorized 4), along with paralogue sequences are then subjected to BLAST comparisons against the complete databases using the network BLAST client program. This is a program that runs on the local computer to compare sequences to the databases at NCBI using the Internet, in the absence of having the requisite hardware to support the entire GENBANK and associated databases locally. Both the local and network BLAST were downloaded from the NCBI ftp site (ftp.ncbi.nlm.nih.gov). This is the rate-limiting step in the procedure, as it is dependent on the speed of the network. A program, makeblast, reads in the names of sequences listed in the genelist file and generates a series of the appropriate BLAST commands using the sequence name input and appending a tag to identify the output result. The file can be executed to serially run each BLAST command. BLAST results are used to upgrade genes to categories 2 and 3 in the presence of the requisite similarity scores.

For any genome, the total amount of the data generated is voluminous. A program, catalog, was written that reads in the sequence name, and evaluates the BLAST results for the sequence using a discriminator based on the BLAST score and e-value (initially set at 100 and 10^{-5} respectively). For evaluating the program vis-à-vis the annotation, a routine was added that compares the presence of a GeneScan coding sequence to the annotation, allowing the calculation of sensitivity, specificity

Table 1. Genes not identified by automated routine of GeneScan (*Plasmodium falciparum* chromosome 3, 36 proteins).

Location	Length	Product
64604 . . 65358	+ 217	Hypothetical protein
175096 . . 176550	**+ 485**	**Putative ankyrin repeat protein**
297681 . . 298580	– 300	Hypothetical protein
366600 . . 367175	+ 114	Hypothetical protein
460635 . . 460967	+ 111	Hypothetical protein
465980 . . 466811	+ 206	Hypothetical protein
499392 . . 502542	+ 433	Aspartyl protease
530935 . . 531934	**+ 127**	**60S Ribosomal protein L26**
534472 . . 535500	– 343	Hypothetical protein
536439 . . 539506	+ 808	Hypothetical protein
539937 . . 540638	– 234	Hypothetical protein
545414 . . 548149	+ 912	Hypothetical protein
548697 . . 550214	– 420	Hypothetical protein
556030 . . 558650	+ 645	Hypothetical protein
559062 . . 562352	– 1097	Hypothetical protein
569361 . . 575061	– 1828	Hypothetical protein
608028 . . 608708	+ 190	Hypothetical protein
629539 . . 630780	– 283	Hypothetical protein
634976 . . 637717	+ 914	Hypothetical protein
671150 . . 672328	+ 222	Hypothetical protein
674534 . . 675166	**+ 131**	**40S Ribosomal protein S15A**
675926 . . 677494	– 523	Zinc-finger protein (C3HC4-type)
721910 . . 722820	**+ 162**	**40S Ribosomal protein S11**
798950 . . 799594	– 167	Hypothetical protein
803967 . . 804440	+ 158	Ubiquitin-conjugating enzyme E2, 17 kDA
814602 . . 815384	+ 181	Elongation factor 1-beta
837051 . . 838055	+ 174	Hypothetical protein
883651 . . 885353	– 357	*N*-Acetylglucosamine-1-phosphate transferase
891449 . . 892431	+ 298	Hypothetical protein
930612 . . 930923	– 104	Hypothetical protein
962244 . . 963193	**– 263**	**40S Ribosomal protein S3A**
965366 . . 965893	– 55	Homologue of *C. elegans* F49C12.11 protein
982954 . . 984299	+ 316	Hypothetical protein
1002641 . . 1003489	+ 283	Hypothetical protein
1004102 . . 1004891	– 232	Hypothetical protein
1009044 . . 1010153	+ 310	Hypothetical protein

and other statistics. In addition, a simple program was written to convert the list of genes into HTML format, providing links to all associated data, allowing easy navigation and retrieval of data using a Web browser (figure 3).

These modules are shown in the flow chart along with the input and output elements.

Results and discussion

We have tested the system on the published sequence of chromosome 3 of *Plasmodium falciparum* (Bowman *et al*. 1999). The automated routine detected 363 coding regions, corresponding to 184 genes, with a sensitivity of 83.6%. Of the 36 annotated genes not identified by GeneScan (table 1), 25 are labeled 'hypothetical' with no known homologue and may represent significant over-predictions by the annotators. Four genes encode ribosomal proteins, which have different contexts from normal coding sequences and thus have low three-base periodicity. They are also small in size, increasing noise in the Fourier transform. However, because of their conserved nature, ribosomal genes are identified using separate routines. Of the remaining seven genes, one (the ankyrin repeat protein) has repetitive peptide motifs, and was rejected during automation owing to the presence of alternate periodicity peaks arising from the repetitive motif. When we consider only experimentally determined genes or genes with significant homology to known proteins to evaluate the efficiency of the method, the sensitivity of this technique on the test sequence is 92.9%.

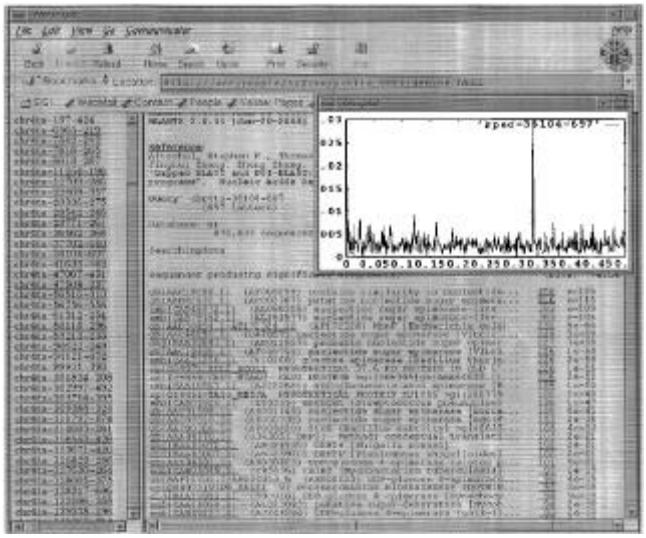

Figure 3. All output from the program is accessible through a Web browser.

As there is no precise definition of the number of genes identified (this routine only identifies tentative coding sequences and not complete gene structure), there is no direct measure of the prediction of false positives. In developing the GeneScan algorithm, evaluation on published sequences resulted in a sensitivity of 86% and a specificity of 100% after manual filtering of repeat sequences. This high specificity is an important reason for using this method to ratify the predictions made by other gene prediction routines, especially when the basis—three-base periodicity—is not explicitly used as a context sensor. Identifying novel coding sequences not previously annotated is of prime importance: GeneScan identified 17 coding sequences not previously annotated (table 2). By BLAST analysis, the *var* gene 3D7-varT3-2 was mapped to four contiguous coding regions at the 3′ telomeric end of the chromosome. Var proteins are expressed by the parasite in the blood stage and transported to the erythrocyte surface where they adhere to host endothelial pro-

Table 2. Novel coding regions identified on *Plasmodium falciparum* chromosome 3 by GeneScan.

8944	12819	4.1	EST match (228, 7e-60)
26606	27683	4.1	EST match (129, 1e-30)
30116	31024	4.1	EST match (129, 1e-30)
46283	47278	1.2	EST match (412, e-113)
256667	256936	4.3	Hypothetical protein
457724	458388	4.3	Hypothetical protein
493592	494094	4.3	Hypothetical protein
565218	565612	4.3	Hypothetical protein
645278	646148	4.3	Hypothetical protein
806597	809078	4.3	Hypothetical protein
841525	841852	4.3	Hypothetical protein
950624	951621	4.3	Hypothetical protein
1026997	1027252	4.3	Hypothetical protein
1027331	1028825	1.1	Identity with *var* gene 3D7-varT3-2 (753, 0.0) and similarity with EST match (325, 2e-86)
1029579	1031220	1.1	Identity with *var* gene 3D7-varT3-2 (879, 0.0)
1031105	1031876	1.1	Identity with *var* gene 3D7-varT3-2 (543, e-154)
1031949	1033024	1.1	Identity with *var* gene 3D7-varT3-2 (593, e-171)

teins. This mechanism prevents the parasite-infected RBC from being transported to the spleen and is also implicated in mortality associated with cerebral malaria where infected erythrocytes bind to the endothelial layers in the brain, blocking blood supply. The protein family is encoded by multiple genes, members of which are selectively expressed to allow antigenic variation. The genetic regulation of *var* gene expression is still unknown, and fragments of *var* genes may serve to mediate recombination events important for their expression.

This routine has application in identifying and characterizing coding sequences from publicly available unpublished genome data. Although the direct result of gene annotation is reserved by sequencing consortia as part of their report on the sequence, advance knowledge of the gene complement is useful in planning experiments. We have applied this routine on data available in November 2000 for *P. falciparum* (www.plasmoDB.org) and have identified 9367 protein-coding sequences. This set of coding sequences may be queried, for example, to identify the genetic basis of observed phenotypes. It was observed that on blocking the erythrocyte-binding antigen known in *P. falciparum*, alternative pathways were available for RBC invasion. We have identified four homologues of the erythrocyte-binding protein (data not shown) which may be tentative candidates to perform this role. Vaccine targets may also be identified: more than 25% of the *P. falciparum* genome is expected to contain members of the *rifin*, *var* and *stevor* genes, all associated with antigenic variation. The gene complement may also be used to identify metabolic pathways unique to the parasite. These can be used as drug targets.

In summary, our automated annotation tools will be very useful in making use of genome data. In all automated predictions there is a certain amount of error, which can be avoided by manual examination of the results.

References

Adams M. D., Celniker S. E., Holt R. A., Evans C. A., Gocayne J. D., Amanatides P. G. *et al.* 2000 The genome sequence of *Drosophila melanogaster*. *Science* **287**, 2185–2195.

Alm R. A., Ling L. S., Moir D. T., King B. L., Brown E. D. Doig P. C. *et al.* 1999 Genomic sequence comparison of two unrelated isolates of human gastric pathogen *Helicobacter pylori*. *Nature* **397**, 176–180.

Altschul S. F., Madden T. L., Schaffer A. A., Zhang J., Zhang Z., Miller W. and Lipman D. J. 1997 Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389–3402.

Anderson S. G., Zomorodipour A., Andersson J. O., Sicheritz-Ponten T., Alsmark U. C., Podowski R. M. *et al.* 1998 The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature* **396**, 132–143.

The *Arabidopsis* Genome Initiative 2000 Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796–815.

Bhattacharya A., Bhattacharya S., Joshi A., Ramachandran S. and Ramaswamy R. 2000 Identification of parasitic genes by computational methods. *Parasitol. Today* **16**, 127–131.

Bowman S., Lawson D., Basham D., Brown D., Chillingworth T., Churcher C. M. *et al.* 1999 The complete nucleotide sequence of chromosome 3 of *Plasmodium falciparum*. *Nature* **400**, 532–538.

The *C. elegans* Sequencing Consortium 1998 Genome sequence of the nematode *C. elegans*: a platform for investigative biology. *Science* **282**, 2012–2018.

Cole S. J., Brosch R., Parkhill J., Garnier T., Churcher C., Harris D. *et al.* 1998 Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* **393**, 537–544.

Fickett J. W. 1982 Recognition of protein coding regions in genomic DNA. *Nucl. Acids Res.* **10**, 5303–5318.

Fleischmann R. D., Adams M. D., White O., Clayton R. A., Kirkness E. F., Kerlavage A. R. *et al.* 1995 Whole genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**, 496–512.

Fraser C. M., Casjens S., Huang W. M., Sutton S. G., Clayton R., Lathigra R. *et al.* 1997 Genomic sequence of a Lyme disease spirochaete, *Borrelia borgdorferi*. *Nature* **390**, 580–586.

Haussler D. 1998 Computational gene finding. *Trends Biochem. Sci.* suppl. 'Trends in Bioinformatics' 12–15.

Heidelberg J F., Eisen J. A., Nelson W. C., Clayton R. A., Gwinn M. L., Dodson R. J. *et al.* 2000 DNA sequence of both the chromosomes of the cholera pathogen *Vibrio cholerae*. *Nature* **406**, 477–484.

Lee P. S. and Lee K. H. 2000 Genome analysis. *Curr. Opin. Biotechnol.* **11**, 171–175.

Lewis S., Ashburner M. and Reese M. G. 2000 Annotating eukaryote genomes. *Curr. Opin Struct. Biol.* **10**, 349–354.

Nierman W. C., Eisen J. A., Fleischmann R. D. and Fraser C. M. 2000 Genome data: what do we learn? *Curr. Opin. Struct. Biol.* **10**, 343–248.

Parkhill J., Achtman M., James K. D., Bentley S. D., Churcher C., Klee S. R. *et al.* 2000 Complete DNA sequence of a serogroup A strain of *Neisseria meningitidis*. *Nature* **404**, 502–506.

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

Reese M. G., Hartzell G., Harris N. L., Ohler U., Abril J. F. and Lewis S. F. 2000 Genome annotation assessment in *Drosophila melanogaster*. *Genome Res.* **10**, 483–501.

Short N., Dennis C. and Cohen B. (ed.) 1997 The yeast genome directory. *Nature* **387**, suppl. 1–105.

Simpson A. J., Reinach F. C., Arruda P., Abren F. A., Acencio M., Alvarenga R. *et al.* 2000 The genomic sequence of a plant pathogen *Xylella fastidiosa*. *Nature* **406**, 151–157.

Tiwari S., Ramachandran S., Bhattacharya A., Bhattacharya S. and Ramaswamy R. 1997 Prediction of probable genes by Fourier analysis of genome sequences. *CABIOS* **13**, 263–270.

Wisconsin Package Version 10, Genetics Computer Group (GCG), Madison, USA.

Received 26 February 2001