Gene identification *in silico*

Shrish Tiwari, Alok Bhattacharya*, Sudha Bhattacharya† and Ramakrishna Ramaswamy

School of Physical Sciences, *School of Life Sciences, †School of Environmental Sciences, Jawaharlal Nehru University, New Delhi 110 067, India

DNA sequence analysis has emerged as one of the major disciplines of theoretical biology and has become an essential tool for study in molecular biology. In this article we review various methods currently available for the analysis of genomes and large-scale DNA sequences in order to detect potential genes out of sequence information.

THE extreme diversity that exists among biological systems is essentially governed by the information content of DNA molecules that make up the hereditary elements – collectively called the genome – which are transferred from one individual to another[1]. Genetic information is decoded through the 'genetic code', which allows specific proteins to be synthesized according to the information present in DNA.

Many features of this genetic code are by now quite well-known[2]. DNA molecules are composed of four nucleotides, the purines adenine (A) and guanine (G) and the pyrimidines, thymine (T) and cytosine (C). The three-dimensional structure of DNA is double helical, with nucleotides on the two strands paired A–T and G–C. The actual arrangement, i.e. the sequence, and the length of the DNA differs from organism to organism and determines its complete biology. One major function of the DNA is to manufacture specific proteins, and nucleotide sequences that contain this information are called the protein-coding regions or genes. There are also ancillary regions of the DNA which regulate and control expression of the proteins at specific times and under specific conditions.

At the same time, there are vast stretches of DNA sequence, principally in higher organisms, whose function is not yet known[1]. Thus, it is still not possible to predict the complete biological functions of a given organism in spite of knowing the complete sequence of its genome. Gene functions can be experimentally determined using the techniques of molecular genetics and biochemistry. Given the time taken for these experiments, though, analysis of a complete genome is still an intractable task. Moreover, for most organisms, such experimental methods and approaches are not yet feasible. Thus, the development of theoretical and computational methods[3,4] for analysis of DNA sequences could significantly aid and accelerate understanding of functions of different regions of DNA.

Availability of entire genome sequences of organisms can help, at the very least, to establish a complete genetic map of the organisms. Further, a comparative analysis of the genomes of a given family of organisms can help determine the minimal set of genes necessary for development and organization within that family. However, the large amounts of sequence information already available and continuously pouring in (in sequence libraries or repositories such as GenBank) pose a challenge, insofar as interpretation and analysis is concerned. There are several ongoing projects to sequence the entire genome of a number of organisms. Within the next few years the intention is to have a complete genome map of organisms such as *Drosophila melanogaster* (genome length ≈ 165 Mbp, consisting of ≈ 15000 genes), *Escherichia coli* (4.7 Mbp, 3000 genes), *Arabidopsis thaliana* (100 Mbp, 13100 genes), nematode *Caenorhabditis elegans* (100 Mbp, 15000 genes), the puffer-fish *Fugu rubripes* (390 Mbp, 80000 genes) and the human genome (3000 Mbp, 100000 genes). Recently, the entire genomes of *Haemophilus influenzae* (1.83 Mbp, 1727 genes) and *Mycoplasma genitalium* (0.58 Mbp, 482 genes) have been sequenced[5,6], and the entire genome sequence of yeast *Saccharomyces cerevisiae* (12.5 Mbp, 6400 genes) is expected in a few months.

Megabasepair DNA sequencing is, by now, a well-developed technology[5]. In the standard strategy, the genome is organized into a library of overlapping fragments, randomly cloned into a number of available vectors such as cosmids, phage P1, BAC or YAC. These vectors are designed to accept large fragments of DNA so that the entire genome of an organism is represented in the library. The choice of a specific cloning vector, to a large extent, depends upon the size of a specific genome. The large fragments are further subcloned in usually M13-based plasmid vectors, to get short stretches of DNA, which are sequenced using Sanger's dideoxy chain termination method[7]. These segments are then re-assembled to form the genome with the help of overlaps, filling of gaps, etc. To accomplish the task of assembling the sequence of a large genome in a reasonable time-span, there is need for the continual development of innovative methods, especially those leading to automation[8], which can further accelerate the speed of sequencing. Currently, a 'random shotgun' method[5,6] has been used for *H. influenzae* and *M.*

genitalium sequences where the genome was randomly cut into segments and sequenced, with alignment done entirely on computer. This approach is feasible only with short genomes or cloned DNA fragments.

The central issue, stated simply, is to identify the *functional* regions of the sequences – the genes on the genome. As mentioned before, in a complex genome, only a small part is functional, in that it is decoded into a protein with the amino acid sequence determined by the DNA sequence. Another small part performs regulatory role by determining the time and extent of decoding in the life of an organism. Protein-coding DNA, along with associated regulatory sequences, is what 'makes sense'. However, a major part of the genome is composed of highly repetitive sequences of unknown or uncertain function – which were previously termed 'junk' or 'selfish' DNA. Thus the identification of protein-coding regions on the genome becomes a major goal of DNA sequencing and sequence analysis.

DNA can be considered as a linear string of the symbols, A, T, G, and C, and it is necessary to specify the sequence along either one of the strands alone since the sequence along the complementary strand is automatically specified. Proteins are synthesized by reading a code from DNA sequence, with a triplet of nucleotides – a codon – corresponding to a given amino acid. Since 20 amino acids are the constituents of naturally occurring proteins, and there are 64 ($\equiv 4^3$) codons, the genetic code is degenerate. The elementary grammar of the genetic code also includes a rule for initiation of protein synthesis (the start codon) and a rule to signal the end (the three stop or non-sense codons) (Figure 1).

The task of gene recognition and identification poses a challenge for several reasons. One can imagine identifying genes by one of two possible routes, either from the expressed protein back to the DNA or from the DNA directly. The route from the protein back to the DNA is made difficult (and uncertain) by the fact that the amino acid-to-codon correspondence is not unique. Since a complete knowledge of the sequences of all proteins in an organism is a distant goal, this approach is of limited utility. In any case, identifying a gene is not simply a matter of finding an open reading frame (ORF), namely a portion of DNA of length at least 100 bp which starts with a start codon and ends with a stop codon (with no other intermediate stop codons). While prokaryotic genes are often continuous ORFs, the expressed part of a gene in the majority of eukaryotes is split into several discrete segments called 'exons' which are interspersed with noncoding intermediate regions, the 'introns'. Exons may be mixed and matched in various combinations to create new genes, and sometimes one gene's exon may be another gene's intron[1]. The entire gene is transcribed into an RNA molecule, from which introns are spliced out, resulting in a messenger RNA that is a continuous ORF and is translated into the corresponding polypeptide.

There are several theoretical approaches to the problem of identifying coding regions in DNA sequences. A variety of mathematical techniques have been brought into play; these include statistical analysis, stochastic modelling, dynamical systems theory and dynamical programming. Current developments in the physical sciences such as chaos theory or the study of neural networks have also been applied, with equal effectiveness. Most of these methods rely quite heavily on computational tools, and the past decade has seen the development of several computer programs that accomplish the task of predicting the coding properties of unknown sequences with varying degrees of success.

The purpose of this article is to present an exposition of the current state of the art of gene identification through computational (*in silico*) methods. In the following section, we discuss the desirable features of coding sequence finders in general, and in this context describe several of the currently employed techniques for gene identification. These methods include both prokaryotic gene detection methods which need to look for coding ORFs, as well as eukaryotic gene detectors, that must locate exons and also determine how these are to be joined in order to make a functional gene. While we have tried to describe the main methods currently employed, our list is not exhaustive; other techniques and algorithms have been reviewed recently by Fickett[9] and Burset and Guigó[10]. This is followed by a discussion and summary.

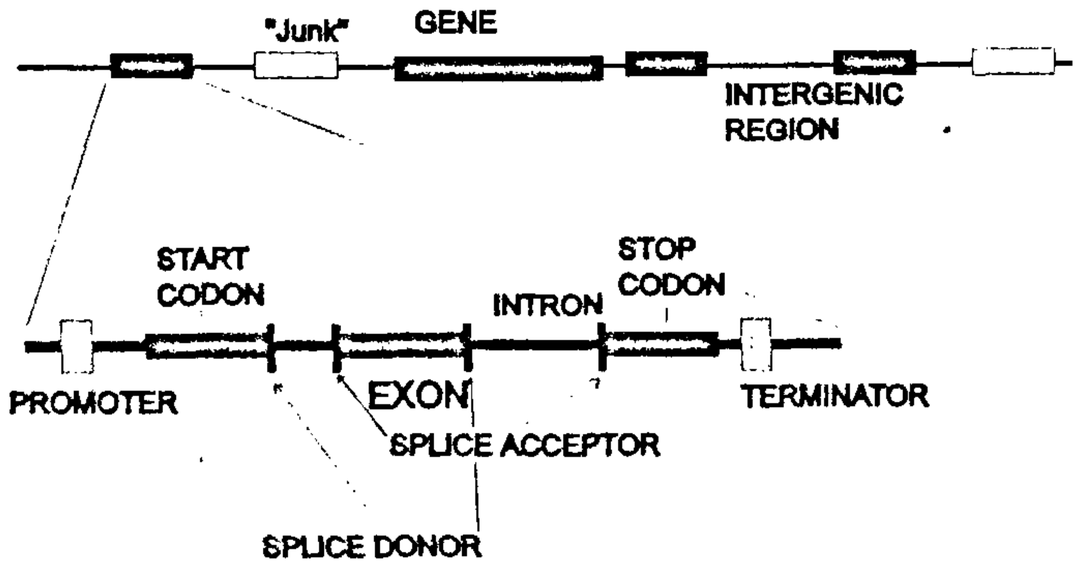

Figure 1. Schematic of the different regions in and around a gene in a genomic sequence, showing the organization of exons, introns, initiation and termination sites, intergenic spacers and promoters.

Gene identification

Given a genome sequence, the task is to locate all the genes. In an ideal situation, this implies identifying all the exonic, intronic and intergenic regions, and start and stop codons pertaining to each gene.

Gene-finding methods can be classified[11] as being *signal*-based, *content*-based, or, as is increasingly more efficient, a combination of both. Given a DNA sequence, there are six reading frames to be examined. These arise from the three possibilities for the origin of the sequence (owing to the triplet nature of the codon) and the two possibilities for the relevant strand of DNA. Signal-based techniques look for a signature in the sequences – start and stop codons, promoters, consensus splice sites, elements upstream of a gene which determine the transcription start site, polyadenylation signal at the 3′-end of the messenger RNA transcripts and similar motifs. Content methods look for codon usage biases, oligonucleotide frequencies, correlation exponents or related similar indicators.

Three measures which are used to evaluate the performance of a gene-finding technique are the sensitivity S_n, the specificity S_p, and the correlation coefficient C_c. These are defined as[12]

$$S_n = \frac{TP}{TP + FN},$$

$$S_p = \frac{TP}{TP + FP},$$

$$C_c = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(FP + TN)(TP + FN)(TN + FN)}}, \quad (1)$$

where *FP* is the number of false positives, i.e. the number of bases *predicted* as coding, but which are subsequently, through experiments or homology, declared noncoding. Similarly, *FN*, the number of false negatives, is defined as the number of bases, which are experimentally confirmed as coding but are declared noncoding by the computational method. *TP* are the true positives, i.e. the number of bases correctly predicted as exonic, while *TN* are the true negatives, the bases correctly predicted as noncoding. The average of the specificity and sensitivity[12] gives some measure of the accuracy of a given method. Although the above measures are evaluated at the nucleotide level, these can similarly be defined at the exon level[10].

Looking for a signal alone has its drawbacks, chief among which are the degeneracy and the lack of unequivocal definitions of signals, which can reduce the sensitivity or specificity. Looking for content alone involves the calculation of several statistical measures, the analysis of which is organism-specific, and which does not always give unambiguous results for the coding potential of a sequence.

The smallest eukaryotic genes are typically about 300 bp in length. Therefore, the existence of an open reading frame (ORF) of reasonable length suggests the *possibility* of a gene, but this alone is insufficient as evidence of coding potential, since start and stop codons could occur randomly with some (small) probability. In addition, single base errors in sequencing cause *frameshifts*, which may abruptly and erroneously terminate an otherwise long ORF.

'Compositional heterogeneity' is a hallmark of functional proteins. Examination of the entire protein databank (PIR database) shows that the 20 different amino acids are used in very different proportions in functional proteins (Table 1). This automatically implies that there must, similarly, be fairly stringent constraints on nucleotide usage at the level of the DNA sequence. All genomes have a bias in the base composition, possibly decided by evolution. This restricts the choice of which codons are used in designating a given amino acid. Thus coding regions have a very *unequal* usage of codons.

Constraints on the amino acid composition of functional proteins thus can lead to regularities in a coding sequence. This feature can be probed with the help of certain statistical measures, which have been used in the earliest methods to locate genes. These are discussed in the following subsection.

Oligonucleotide distribution-based methods

In these methods, 'unequal codon usage' is the fundamental measure for identifying protein coding stretches. This asymmetry in codon frequency gives rise to compositional variations. However, it has been observed that unequal usage of amino acids, without the codon preference or base composition bias, is enough to produce significant compositional variations in all the three reading frames for codons and bases. It is useful to go

Table 1. Average amino acid usage by proteins

Amino acid	Frequency (%)
Ala	7.60
Arg	5.23
Asn	4.36
Asp	5.21
Cys	1.89
Gln	4.17
Glu	6.32
Gly	7.19
His	2.28
Ile	5.29
Leu	5.81
Lys	9.17
Met	2.29
Phe	3.97
Pro	5.20
Ser	7.15
Thr	5.87
Trp	1.31
Tyr	3.21
Val	6.49

beyond the level of the trinucleotide codon and examine the distribution of other oligomers, which is also quite distinctive.

Testcode. This observation can be used to develop a method to distinguish coding regions from non-coding ones in genomes[14,15]. The first step involves setting up a codon frequency table for coding regions, which can be achieved by using prior information coming from previously determined genes or open reading frames. Then a window is moved along the sequence, three bases at a time. For each window, the codon frequency is evaluated in each of the three reading frames, and compared to the one evaluated for a known set of genes. The frame in which the deviation is comparable to that for the test set of genes is adjudged as a coding frame, whereas if the deviation is large in all the three frames the region is identified as noncoding.

While asymmetry in codon usage has been used by Staden[15] to develop a method for the identification of genes, Fickett[14] exploited compositional variations observed in coding and noncoding regions. In the technique TESTCODE a total of eight parameters – four positional and four content parameters, one for each nucleotide – are used to judge whether a sequence is coding or not. The positional parameter is the ratio

$$P_\alpha = \mathrm{MAX}(f_1^\alpha, f_2^\alpha, f_3^\alpha)/\mathrm{MIN}(f_1^\alpha, f_2^\alpha, f_3^\alpha)\,, \qquad (2)$$

where f_i^α is the frequency of the base α at position i in the codon, and measures the extent to which a base is favoured in one codon position over another. Since it is not relevant which of the codon position favours the base, the positional parameters have fairly similar distributions in all sequences, regardless of the differences in codon usage strategy between organisms. The content parameter for a nucleotide is simply its frequency.

A standardized table, giving the coding probability of a sequence for a range of each of the eight parameters can be devised from existing sequence information available. (In Fickett's implementation of TESTCODE, this table derives from the Los Alamos Sequence Library.) Each of these parameters is used with a different weight, determined as follows: The parameter is used alone to predict coding function, from the standardized table, and the sequence deemed to be coding if the probability exceeds 1/2. The weight of the parameter is the percentage of times this guess is correct, less 50% (the random level). For a test sequence of unknown functionality, one calculates the eight parameters and obtains the probabilities p_i and weights w_i for each of them from the tables. The sum $\Sigma_{i=1}^{8} p_i w_i$ is evaluated to get the TESTCODE indicator, and the prediction is then obtained from the standardized look-up table of the sample set.

By itself TESTCODE cannot find exact boundaries for coding sequences, but it is well adapted for combination with other techniques such as searches for ORFs, ribosome binding sites, intron boundaries. The reliability of the method when checked by taking half of the sequences in the Los Alamos Library as sample set and the other half as test set gives an error rate of prediction of around 5%.

GeneMark. Differences in oligonucleotide frequencies have been exploited in the technique GeneMark[16]. The Markov model is a convenient means for evaluating the probabilities of occurrence of oligonucleotides while taking into account correlations between frequencies in different positions. This model has been widely used for the study of one-dimensional strings generated in dynamical systems or those involved in language theories, to assess the correlation structure of the sequences and frequencies of words of length m in strings of length N.

A Markov process assumes that the state of a system at time t depends on its state at time $t-1$ only. This rule, when interpreted for strings of symbols, states that the probability of the symbol α_i at position i depends only on the probability of the symbol α_{i-1} at position $i-1$. For DNA sequences, this model reproduces the dinucleotide frequencies. For higher oligonucleotide frequencies, higher order Markov processes have to be invoked. For example, in a second order Markov model, the probability of symbol α_i at position i depends on the probability of the doublet $\alpha_{i-1}\alpha_{i-2}$ at positions $i-1$ and $i-2$, and this reproduces the trinucleotide frequencies in DNA.

Since correlation between nucleotides differs in coding and noncoding sequences, the corresponding Markov models are also different. The fact that the reading frame plays an important role in coding regions is accounted for by considering 'phased' or non-homogeneous Markov models. In these the probability of an oligonucleotide depends on which of the three codon positions its first nucleotide occupies. Recent studies have shown that in-phase hexamer statistics are very effective in distinguishing coding regions, since they take into account not only the codon bias but also the correlations between the various positions of neighbouring codons. Thus GeneMark uses phased fifth order Markov chains to make its predictions of coding regions.

A sliding window is used, which moves along the sequence in steps, which are a multiple of three. For each window, the algorithm calculates if the DNA fragment is modelled by the phased Markov model in one of the six frames or the ordinary Markov model. In the first-order Markov model (which can be generalized to fifth order, below), the probability that a

sequence S of length L is noncoding is given by the ordinary Markov chain formula

$$P(S \mid \mathrm{NON}) = P_N^0(\alpha_1) \times P_N(\alpha_2 \mid \alpha_1) \times \ldots \times P_N(\alpha_L \mid \alpha_{L-1}) . \quad (3)$$

Here, $P_N(\alpha_L \mid \alpha_{L-1})$ is the conditional probability of observing the nucleotide α_L in position L, given that α_{L-1} is observed at position $L-1$, and P_N^0 is the initial probability. P_N values are calculated on the training set of noncoding sequences.

The probability of the frame 1 in S being coding is given by the phased Markov chain

$$P(S \mid \mathrm{COD}_1) = P_1^0(\alpha_1) \times P_1(\alpha_2 \mid \alpha_1) \times P_2(\alpha_3 \mid \alpha_2) \times P_3(\alpha_4 \mid \alpha_3) \times P_1(\alpha_5 \mid \alpha_4) \times \ldots \times P_2(\alpha_L \mid \alpha_{L-1}) . \quad (4)$$

Here, P_1, P_2 and P_3 are respectively the probabilities determined for the three codon positions in frame 1 and P_1^0 is the initial probability. The values P_i and P_i^0 are defined from the training set of coding sequences. For the other frames (2–6) the probabilities $P(S \mid \mathrm{COD}_m)$ are defined by similar formulae. The fifth order case is a generalization of the first order model and the probabilities are determined from Bayes formula[17].

$$P(\mathrm{COD}_m \mid S) = P(S \mid \mathrm{COD}_m) \times P(\mathrm{COD}_m) \sum_{m=1}^{6} P(S \mid \mathrm{COD}_m) \times P(\mathrm{COD}_m) + P(S \mid \mathrm{NON}) \times P(\mathrm{NON}) , \quad (5)$$

where $P(\mathrm{COD}_m \mid S)$ is the *a priori* probability that an unspecified fragment S is coding and its first nucleotide is located in the codon position defined by index m. $P(\mathrm{NON})$ is the probability that S is noncoding and is assumed to be the same as $P(\mathrm{COD}_m)$ and equal to 1/2.

For each sequence to be analysed GeneMark determines all possible ORFs and the average value of $P(\mathrm{COD}_m \mid S)$ is computed for each ORF. If the value is greater than 0.5, the cutoff, the ORF is included in the list of predicted genes. If there is more than one reading frame in which this probability is greater than 0.5, the frame with a higher probability is chosen. When applied to the unannotated sequences of *E. coli*, the technique found many new genes[17].

GeneId. The simple method described above can be made more sophisticated. For example, GeneId[18], a hierarchical rule-based system for identifying probable protein coding genes, starts by identifying all possible signals, such as initiation and stop codons, donor and acceptor sites, promoters, poly-adenylation signals and assigns each a rank, according to the preferred ordering and spacing among the various sites. Using these 'atomic sites', all possible exons are constructed and ranked by computing some of the statistically significant properties equivalent to those described above and comparing with a cutoff value obtained from a sample set. Thus each exon is sequentially filtered through the cutoff for each of the statistical measures. These exons are then classified into equivalence classes. Two exons are said to be equivalent if they occur in exactly the same gene. These classes of equivalent exons are then assembled to form the gene. A function of values assigned to each of the component exon classes is assigned to the potential gene, and this score is used to rank the gene[18].

The sample sets that have been used in GeneId consist of the first, internal and terminal exons from the primate, mammalian, rodent and vertebrate groups of GenBank 64.0, excluding those with alternative splicing sites, mutants, pseudogenes, etc. These have been used to derive profiles for the prediction of the various gene-finding signals and to calculate the cut-off values for the variables used to derive the rules through which the exons are filtered. Weights are assigned to the sample set of exons to correct for the unequal representation of homologous gene families in the database. (Each family of similar exons is assigned a weight of 1.0, the weights of the sequences are assigned according to the topology of the family, and these weights are used in the derivation of the profiles.)

Given a DNA sequence, the first step involves the identification of the various atomic sites. These are confirmed by several context-based rules. For example, the first ATG does not necessarily correspond to the first AUG of the mRNA. To determine the potential start site, sequence context and distance to the cap site are used as criteria. The profile for initiation codon is derived using first expressed exons.

Similar weighted profiles are constructed for donor and acceptor sites from a set of internal exons, and respective cut-offs are established. In the case of start codons, the distance of the codon to the cap site is computed, and from the frequency distribution of this distance, a cut-off can be determined beyond which very few exons have their start. Similarly, for the stop codon, the distance to the end of transcription unit is computed for the set of last exons, and from the frequency distribution, the critical distance can be fixed. These statistics are used to establish the authenticity of a given exon.

Apart from these, some of the standard statistical measures – nucleotide frequency, positional correlations – can also be used to further filter the exons. The average correlation coefficient and sensitivity for this method, for 222 sequences from a variety of organisms – human, chicken, goat, rat, etc. – are 0.79 and 0.88, respectively.

Neural networks

Neural networks[19] are numerical algorithms which allow a system to learn, recognize and classify patterns. The underlying idea for these methods derives from the nervous system of living organisms. A model neuron is a simplified version of the biological neuron, and is a two-state threshold device having outputs +1 or 0 corresponding to the firing or non-firing state of the neuron, respectively. A multiply-connected collection of model neurons forms a neural network[19].

The simplest application of the collective computation of a neural network is associative memory, i.e. storage and recall of information by association with other information. This process is modelled by a neural network as follows. For a set of N two-state neurons the total number of patterns are 2^N and P of these patterns are stored in memory. If a new (or 'test') pattern is now presented to the network, it should be able to recall the stored pattern that resembles it most strongly. This is achieved by defining a dissipative dynamics on a surface wherein the stored patterns are made to correspond to local minima, and the test pattern is allowed to evolve under the dynamics so as to flow into the local minimum with the closest pattern match. The dynamical evolution of the state s_i of neuron i in the presented pattern is defined as

$$s_i(t+1) = \mathrm{Sgn}\left(\sum_{j=1}^{N} w_{ij} s_j(t) - \theta_i \right), \tag{6}$$

where s_i is the state of neuron i, w_{ij} are the synaptic coupling strengths and the threshold θ_i is usually chosen to be 0, for simplicity. The choice of the weights, or coupling strengths, defines the learning rule of the network. The earliest of learning rules used is the Hebb rule, which defines the coupling strength as

$$w_{ij} = \frac{1}{N} \sum_{\mu=1}^{P} \sigma_i^{\mu} \sigma_j^{\mu}, \tag{7}$$

where σ_i^{μ} is the state of neuron i, in the stored pattern μ.

Similar to the task described above, of storing and recalling information, neural nets can be designed to recognize and classify patterns. Such nets have a fundamentally different structure, consisting of an input layer, a black-box consisting of one or several hidden layers, and an output layer. The signals received by the net through the neurons of the input layer are processed in the hidden layers, and the result is sent to the output layer. Since information flows in one direction these nets are called feed-forward layered networks or, more simply, perceptrons. Again, in this case too the output signal depends on the input signal and the coupling strengths of the synapses,

$$S_i = f\left(\sum_k w_{ik} \sigma_k \right), \tag{8}$$

where S_i is the state of the ith neuron of the output layer and σ_k that of the kth neuron in the input layer. As before, the choice of the weights defines the learning rule. The learning is said to be supervised, if the output is known for a sample set and the weights, initially chosen randomly, are monitored by the output error. This learning is not very realistic, since it requires complete and detailed knowledge of the output, which is very uncharacteristic of the actual system they simulate. For this reason other concepts of learning have been studied, which are based on reward and penalty, for example. These come in the general class of unsupervised learning.

In supervised learning, an extensively used learning rule is the gradient learning. In this learning, the total deviation between the actual output and the desired output, and its gradient with respect to the synaptic weights, is computed. In the next iteration, the synapses are modified by a small fraction of the gradient. The disadvantage of this method is that it cannot be applied to perceptrons built with deterministic, binary-value neurons. The major advantage is that it can be generalized to multi-layered perceptrons. Multi-layered perceptrons become necessary because some very simple problems become intractable with simple perceptrons[19].

A method based on the generalization of the gradient rule to multi-layered nets, is error back-propagation, when the gradient rule is applied recursively to the synapses of the output and those of the hidden layers. In applying the rule to the hidden layers, the gradient of the total output deviation with respect to the weights of the hidden layer is computed, and the weights are then modified by a small multiple of this gradient. Thus, the error at the output is propagated back to determine the weights of all the hidden layers. This is a very powerful learning rule, and has been used extensively.

GRAIL. The widely used method GRAIL employs a neural network algorithm to implement a method similar to that of GeneId. In this case, however, instead of looking for the various atomic sites, several statistical measures are evaluated and the multi-sensor neural network processes them to define a score for the coding region. As depicted in Figure 2, seven sensors are used, including Fickett's measure[14], the positional base frequency to determine the coding frame, and hexamer frequencies. These are evaluated for a simple set, and

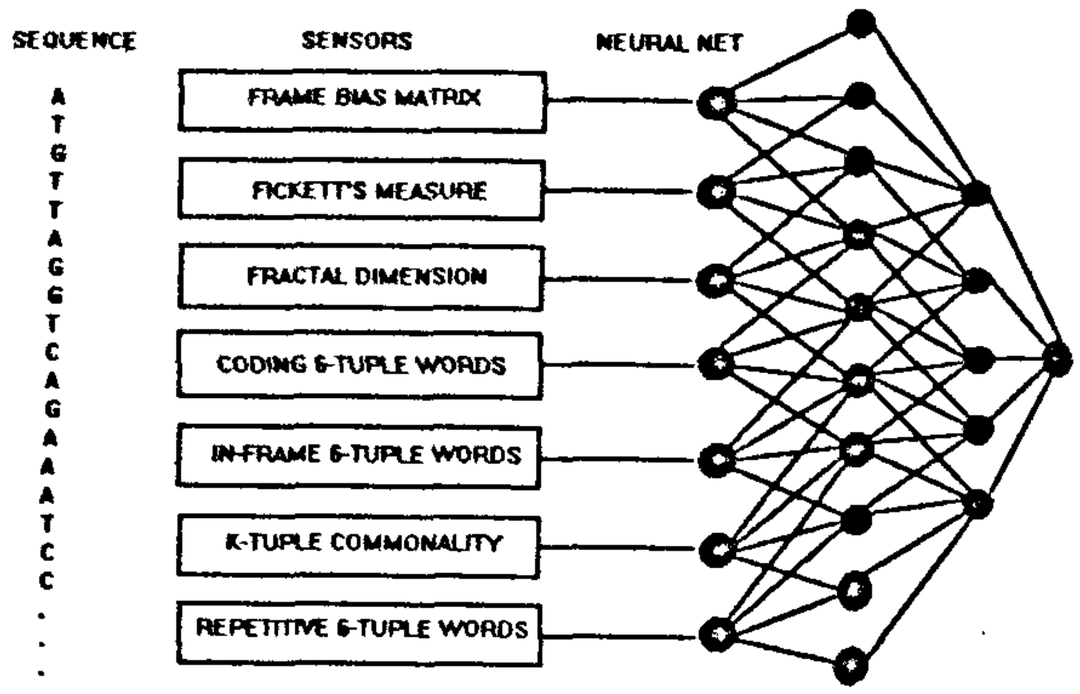

Figure 2. The neural network used by GRAIL with an input layer of seven nodes, two hidden layers of fourteen and 5 nodes respectively and an output layer of single node. (Adapted from ref. 20).

the weights of the measures are extracted. Using these weights the neural net evaluates potential coding regions.

The sample sequences originally used for the evaluation of the weights[20], were 18 genes from the human genome. A 99-base window, centered at each position, was used to evaluate seven sensors, scaled between 0.0 and 1.0. These values were passed through the neural net and at the end of the training, the weights of the net were extracted. The sensors used are:

1. *Frame bias matrix.* This is a 3×4 matrix containing the frequency with which a particular base occupies each of the three positions in a codon. This measure relies on the fact that while in the noncoding region or the incorrect reading frame of a coding region the distribution of positional frequencies is nearly uniform, in the coding frame there is a significant deviation from the random distribution. The standard bias matrix has been obtained from coding exons of human sequences. The correlation coefficient between the standard matrix and that of the three reading frames of each window is evaluated. The difference between the best and the worst coefficient is used as an indicator of the correct coding frame.

2. *Fickett's measures.* These are the same parameters used by Fickett in TESTCODE. The output of TESTCODE is used as the value of this sensor.

3. *Dinucleotide fractal dimension.* It is known that dinucleotide frequencies are far from random. The various dinucleotides can be grouped according to their frequencies of occurrence. It is thus possible to view a DNA sequence as a dynamic function, by examining transitions of sequential dinucleotides, i.e. asking whether the next dinucleotide belongs to the same frequency group or not. These fluctuations are characterized by a fractal dimension[21]. This dimension has a lower value in coding regions as compared to noncoding regions. Thus the sensor value is the difference in the dimension between a reference value derived from introns and that for the test window.

4. *Coding 6-tuple word preferences.* This measure is the sum of the 6-tuple preferences, defined as the logarithmic ratio of the normalized frequency of 6-tuple words in coding and noncoding regions. It has been noticed[22] that, in addition to uneven codon frequencies, there exist correlations between nucleotide positions of adjacent codons in coding exons. Hexamer, or dicodon, frequencies are a measure of both, the codon bias and correlations between nucleotides of neighbouring codons.

5. *Coding 6-tuple in-frame preferences.* This is the same as the previous measure, except that the 6-tuple frequencies are compared with the preference values of in-frame 6-tuples in coding DNA.

6. *Word commonality.* This is the logarithmic ratio of the normalized frequency to the expected random frequency of hexamers. The word commonality measure is summed over the entire window.

7. *Repetitive 6-tuple word preferences.* This measure is the same as the above, except that the comparison here is with several classes of repetitive DNA.

Sensor 1 helps to establish the correct reading frame of a coding sequence. Sensors 2–6 are various statistical measures, which are significantly different for coding and noncoding DNA. Sensor 7 is a negative indicator, since it is a statement that repetitive DNA rarely codes for protein.

These seven sensors form the nodes of the input layer to the neural network constructed in GRAIL. In addition, there are two hidden layers of 14 and 5 nodes, and an output node. For training sets, the correct output value (0 for noncoding and 1 for coding) is also provided. In the learning phase, the net compares the output value with its prediction and adjusts its weights, using the back-propagation algorithm. The net thus optimizes its performance by continuous evaluation of the output error.

When tested[20] on a sample set of human genes, the overall sensitivity and correlation coefficient was found to be 0.54 and 0.69 respectively. However, the performance of a later and improved version, GRAIL2 (ref. 23), on a similar data set, showed a significant improvement, with the overall correlation coefficients and sensitivity as 0.80 and 0.86 respectively.

GeneParser. GeneParser[24] combines the connectionist approach, adopted by GRAIL[20], with a recursive optimization procedure dynamic programming[25], to predict intron–exon structures of genes. The dynamic programming algorithm is used first to parse a sequence into basically four classes, namely introns, first, internal and last exons, subject to certain 'grammatical' constraints. In the second step, a neural network is trained to weigh

several content and site statistics, and score the intervals of interest in the sequence. Finally, the suboptimal solutions of the parsing problem are obtained and presented in a graphical format, enabling the user to identify which of the predicted splice junctions are most likely to be correct.

Like the previous technique, here also multiple lines of evidence are used to identify exons. The content statistics used in this case are in-frame hexamer frequencies, local compositional complexity, intron–exon length distributions, bulk hexamer frequencies and BLAST similarity scores[26]. In the site statistics, splice sites and translation initiation sites were discriminated with the help of occurrence of specific nucleotides in the vicinity of the sites.

In-frame hexamers are used to determine the correct reading frame. The logarithmic ratio of the frequency of a hexamer to that of its frequency in a random sequence of the same base composition is summed over to define the hexamer score. Preferred hexamers have a positive score, while those avoided have a negative score. While in-frame hexamers are evaluated in a particular reading frame, bulk hexamer frequencies are calculated disregarding reading frames. These frequencies differ significantly between sequences of different functionality.

Local compositional complexity is quantified through the Shannon entropy[27] for oligonucleotides of length $L = 8$, which provides a measure of redundancy. This quantity distinguishes between coding and noncoding regions by virtue of repetitive sequences which occur typically in noncoding regions. Error tolerance can be incorporated into the method by introducing random frameshift and substitution errors with predetermined error rates into the sequences of the training set.

The method was tested on 28 human sequences used by GeneId and GRAIL. The sensitivity and the correlation coefficient for this set turned out to be 0.87 and 0.78 for the latest version of GeneParser. This is a significant improvement as far as the sensitivity of the method is concerned.

Linguistic analysis

The spoken languages of the world have two remarkable features. The first is Zipf's law[28], which is essentially the observation that the frequency of word usage has a power-law dependence. A histogram of the total number of occurrences of each word in a text versus the rank of the word follows a power law, with an exponent $\xi \sim -1$. This is an empirical observation and is seen to hold for all languages[29].

The other common feature of all languages, is redundancy[30], i.e. words (or sentences) do not become unintelligible by the omission of some letters (or words). This notion of redundancy can be quantified through the Shannon entropy. If $p^{(n)}(A_1, \ldots, A_n)$ is the probability to find a word $(A_1, \ldots, A_n)$ in a string of length n then the Shannon entropy is defined as

$$H_n = -\Sigma p^{(n)}(A_1, \ldots, A_n) \ln p^{(n)}(A_1, \ldots, A_n) . \qquad (9)$$

Such analysis can also be adapted to DNA symbol sequences[29], if one can properly define the concept of a word. In coding regions, the 64 codons can be considered as words, but putative words in noncoding regions could have lengths greater than 3. The word of length n can be considered as a parameter that varies from three upward, and a DNA sequence is considered as an overlapping set of word n-tuples. Sequences from different categories of organisms were analysed[29], and the Zipf exponents were larger for noncoding sequences than for coding sequences – indeed the largest exponent obtained was for the noncoding sequences of *C. elegans*, $\xi = 0.537$. This result (which is not entirely uncontroversial[31]) suggests out that noncoding regions were closer to the natural languages than the coding regions, and implies that noncoding regions may have a structured 'language'.

It is conceivable that this difference in language-like properties between coding and non-coding regions can be adapted to a method to distinguish between coding and noncoding regions, similar to the coding sequence finder (CSF) described in section 'Coding sequence finder' below. However, linguistic analysis-based methods for gene identification, on the other hand, exploit the formal structures of languages and grammars.

GenLang. Formal language theory views languages as sets of strings defined over some alphabet with concise sets of rules called grammars. Theoretically, for a given alphabet one can define an infinite number of languages. Grammars have been studied intensively with the help of computers, and have been used to describe the complex structures of strings of symbols. In the process computer programs, called parsers, have been developed, which are capable of determining whether a given string satisfies the rules of the grammar. These programs have been applied to the problem of searching for complex patterns specified by grammars, in a technique known as syntactic pattern recognition.

The information encoded in the DNA sequences uses an alphabet of four letters, with a set of rules determining the protein-coding regions. Thus the techniques of language theory can be applied to identify the syntactic patterns of the DNA language. The problem consists in defining the protein-encoding gene grammars. The core of the grammar[12] can be presented in the form of a binary tree (Figure 3). The root node is the gene, which is hierarchically built up. It is analogous to the sentence of a natural language. It consists of a start (and body)

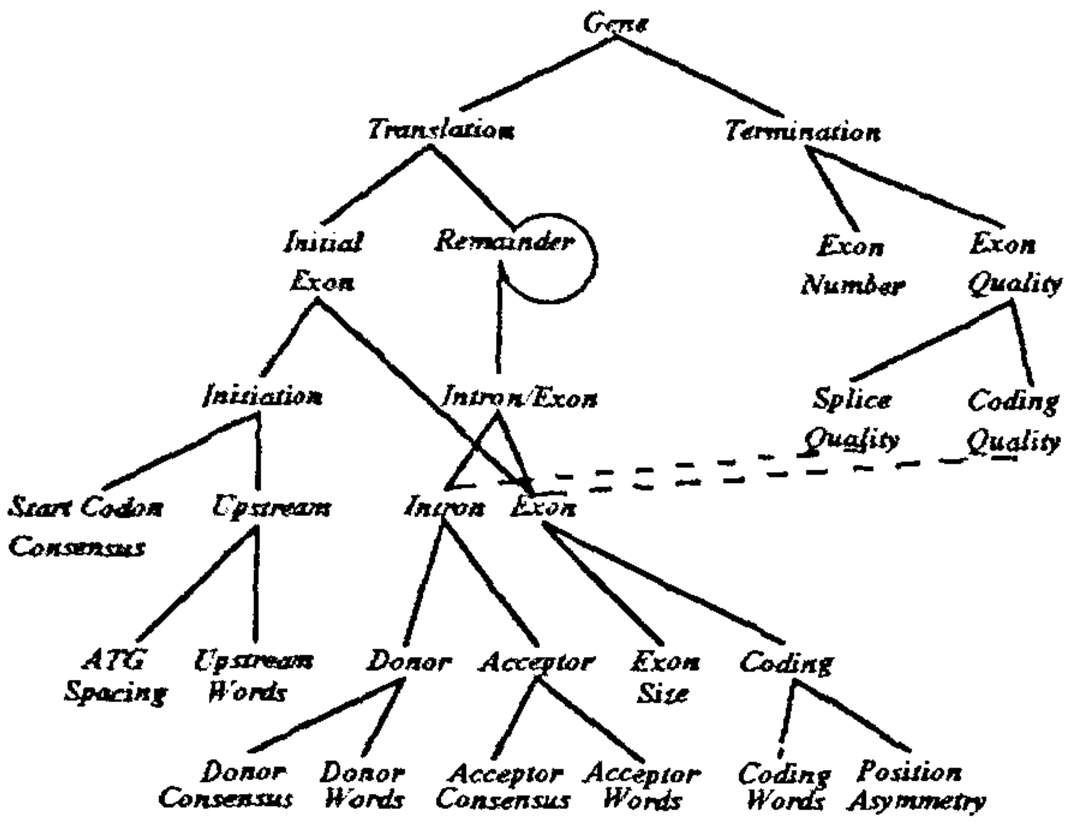

Figure 3. Graphical representation of the core grammar structure used in GenLang. (Adapted from ref. 12).

and a termination. There are various rules which define the start (first exon) and the body (internal exons and introns), and the termination (last exon). At the bottom of the tree are what are termed as 'leaf' rules, variously called sensors (neural network technique) or 'prelexical' in the linguistic context. These are derived directly from the sequence. There are 13 leaf rules used in GenLang, including rules for determining a true initiation of translation, such as consensus for the start codon, its distance from the ATG preceding it upstream, etc.

Each leaf rule is assigned a 'cost', or a threshold error, which it can accept. The cost of consensus sequence (signal), for example, is defined as the sum of the negative logarithm of individual base position frequencies, normalized so that the most likely base in each position contributes zero cost. Costs are propagated up the parse tree and summed at each node, which is connected to two leaves or other nodes, so that each subtree has its own threshold cost. The cost at node N is defined in terms of the costs of the two lower-level rules – left (L) and right (R) child—as follows

$$\text{Cost}_N = (1-\mu)\,\text{Cost}_L + \mu\,\text{Cost}_R\,, \tag{10}$$

where μ is a mixing parameter, with values ranging between 0 and 1. As for the neural networks, values of μ are determined from sample sequences. In addition, each node is associated to a pair of threshold costs, θ_L and θ_R. If the cost accumulated from a subtree at a node exceeds the threshold cost, that path is discarded and the grammar may backtrack or retry the node at the next iteration. Even if a gene is developed successfully, the grammar can be made to backtrack, in an attempt to minimize the overall cost.

The training set currently used[12] consists of sequences from organisms including man, mouse, Drosophila and dicot plants clustered into groups of similar genes, as in GeneId, to take into account the over-representation of some classes of genes (e.g. globins) in the databank. The performance of the technique was assessed, using various measures, including the standard sensitivity and correlation coefficients. For a test set (also from the same organisms) these values were 0.83 and 0.77 respectively. The other measures were more stringent, such as the fraction of genes correctly predicted completely and the fraction of correctly predicted exons. Typical values for these were 0.1 and 0.5 respectively.

Correlation methods

One feature of all methods described above is that they are *context dependent*, i.e. in each case a sample set is required, and parameters obtained from this sample set are then used to determine the potential coding properties of a test sequence. Two approaches that do not require any particular prior information to determine the coding potential of a DNA sequence, which are based on the correlation properties, have recently been developed.

In recent years there has been a flurry of activity, primarily in the physics literature, on the study of long and short-ranged correlations in long genomic sequences. A widely used tool to study the short and long range correlation structure of symbolic strings is the discrete Fourier transform. For a symbolic sequence $\{S_j(\alpha)\}$, $j = 1, \ldots, N$, of symbols $\{\alpha\}$ this can be defined as

$$S(f) = \sum_{\alpha} S_{\alpha}(f) = \sum_{\alpha} \frac{1}{N^2} \left| \sum_{j=1}^{N} S_{j\alpha}\, e^{2\pi i j f} \right|^2 \tag{11}$$

where

$$S_{j\alpha} = 1 \quad \text{if } S_j = \alpha$$
$$= 0 \quad \text{if } S_j \neq \alpha\,.$$

The analysis of a large set of coding and noncoding sequences has revealed[32,33] that the Fourier transform of coding sequences has a distinct peak at frequency $f = 1/3$, as in Figure 4 *a*, while this peak is absent from noncoding sequences as in Figure 4 *b*, *independent of organism and base composition*. On the other hand long-ranged correlations show up as a $1/f$ fall-off in the frequency, and this can be seen either in the simple power spectrum defined above[34,35], or in more involved analyses such as the wavelet transform[36].

Another method to detect long-range correlations proceeds through the construction of a so-called DNA walk[37], which is defined as follows. Starting from the origin of a square lattice, one considers a directed

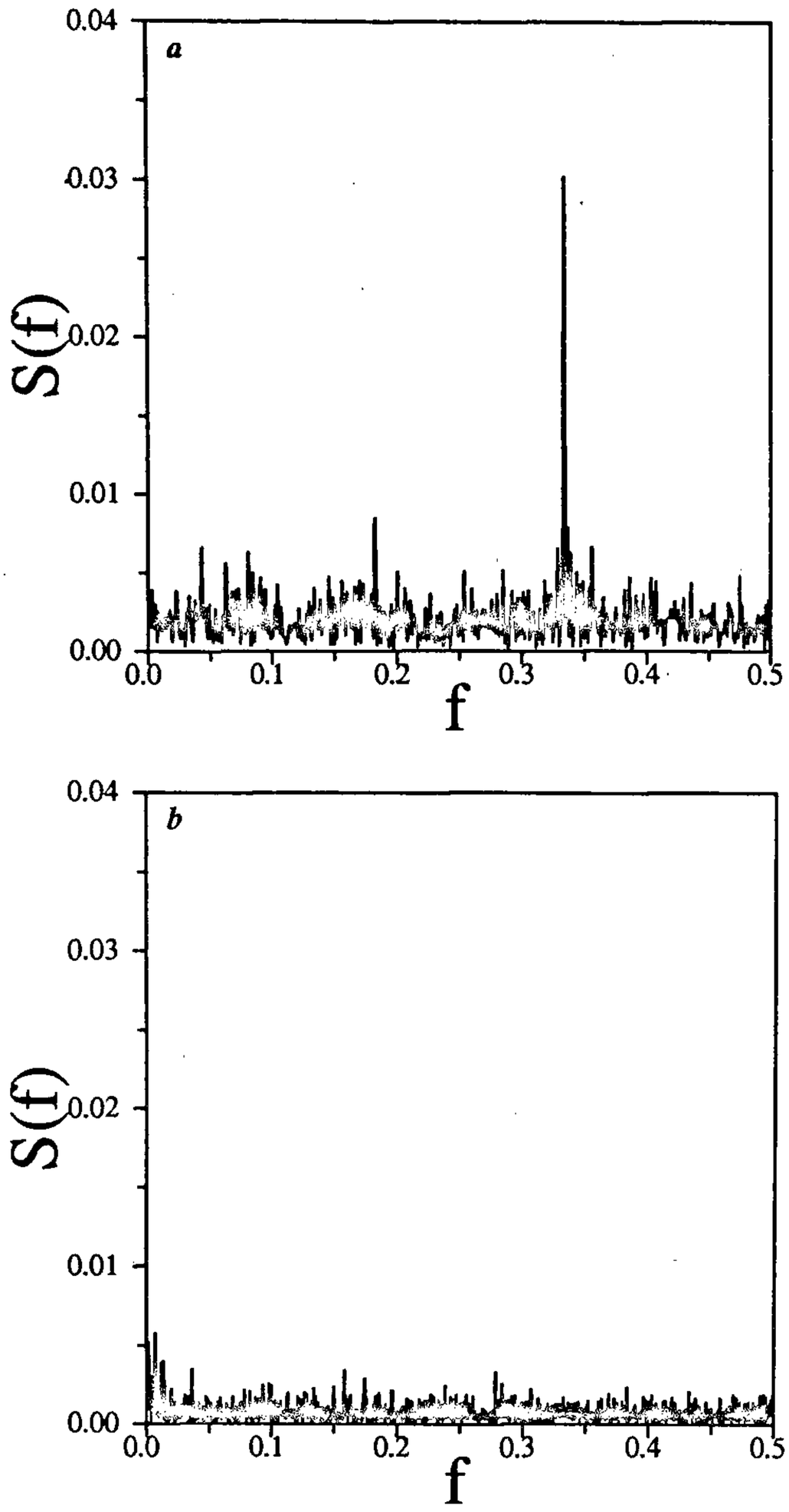

Figure 4. The Fourier transform of a coding (*a*) and a noncoding (*b*) sequence from chromosome III of *S. cerevisiae*.

random walker who takes steps up or right according to the nature of the DNA sequence under analysis. For a pyrimidine at site i, the walker takes a step upward, $u(i) = +1$, and for a purine, a step rightward $u(i) = -1$. The net displacement $y(n)$ after n steps scales as $y(n) \sim n^{\alpha}$. For an uncorrelated or short-range correlated walk, the theory of random walks gives the well-known result that $\alpha = 0.5$, and thus deviations from this scaling behaviour are indicative of long-ranged correlations in the data.

One difficulty of applying this approach directly to DNA sequences is that, DNA has a patchy structure, i.e. the average base composition varies from region to region, and this can give rise to features similar to those observed in long-range correlations. To overcome this difficulty, a technique termed Detrended Fluctuation Analysis (DFA)[38], which involves dividing the region of interest of length N, into N/l non-overlapping windows of length l is adopted. For each window, the ordinate of the least squares fit for the net displacement of the DNA walk is defined as the local trend, and the detrended walk is then defined as the difference between the original walk and the local trend. The variance about the local trend in each window is calculated and averaged over to give a quantity $F_d(l)$. It was shown that $F_d(l) \sim l^{\alpha}$, where $\alpha = 1/2$ for patchy, but otherwise uncorrelated or short-range correlated sequences, while for long-range correlations $\alpha > 1/2$.

Results of studies of several DNA sequences, separately analysed for coding and noncoding regions, have established the result that coding regions have short-range correlations, while noncoding and intronic regions have long-range correlations. Although this result is somewhat controversial[39], there is enough evidence to suggest that the scaling exponents do indeed behave differently in the two cases[40].

The above empirical observations can be adapted as the basis of techniques to then predict coding potential, either by looking for the presence of short-range correlations, in the Fourier method GeneScan[33] outlined below, or by looking for the absence of long-range correlations, as in the coding sequence finder (CSF) algorithm[41].

Another technique developed[42] to study the correlation structure of DNA sequence uses the ideas of the so-called chaos game. The chaos game representation (CGR) of a DNA sequence can be constructed as follows. The four corners of a square are labelled by the four nucleotides in DNA, A, T, G, C respectively. Starting from the origin (the centre of the square), a point is plotted midway to the corner corresponding to the first nucleotide of the sequence. Subsequent points are the successive midpoints between the previous point and the corner corresponding to the current nucleotide, as the sequence is read through. What emerges is a pictorial representation of the DNA sequence. The density of points in the different part of the square indicates the various correlations between nucleotides. For a random sequence, for example, the square fills up uniformly. This technique has so far been used only to classify[43] different groups of genes. The difficulty of using the technique to recognize coding regions is that a CGR pattern becomes distinct only for fairly long sequences (> 1000 bp).

GeneScan. In earlier work[33], we have used the existence of the 1/3 periodicity in coding regions to develop the technique GeneScan which detects the coding potential in genomic regions as follows. A window of length M, is moved along a sequence of length N, and the *local* peak-to-noise ratio at $f = 1/3$, defined as

$P_M(j) = P(1/3)/\bar{P}$, is measured. Here $P(1/3)$ is the peak height at $f = 1/3$, $\bar{P}$ is the average peak height of the spectrum and j is the position of the centre of the window. Study of a very large number of previously identified genes and noncoding sequences has shown that the peak-to-noise ratio of the spectral feature at $f = 1/3$ exceeds 4 in almost every coding sequence, while for a noncoding sequence, this ratio is less than 4 (and usually less than 3). This empirical observation can be used to set a threshold, which, if the local peak to noise, $P_M(j)$ exceeds, then the window is deemed to overlap with a coding region. This simple procedure thereby gives the approximate location of coding exons. Figure 5 shows a representative result, for bases 15000–25000 of *S. cerevisiae* chromosome III. The size of the window to be used depends on whether or not we expect short (< 200 bp) exonic regions. Once the approximate position of the coding region is located, the sequences are further scanned to find, in any of the six reading frames, the exact location of the initiation and the stop codon, or the location of possible consensus splice sites. For prokaryotic genes, our procedure works extremely well, and a typical result, from the analysis of the *H. influenzae* genome is given in Table 2. We have similarly studied a host of organism ranging from *S. cerevisiae* (9 of the 16 chromosomes), *E. histolytica*, *A. vinelandii*, *A. californica*, *C. elegans*, *M. genitalium* and several human genome sequences as well. The overall quality of the results is similar to that in Table 2.

For eukaryotic sequences, where exons can be very short, the scanning windows need to be adjusted to an optimal length. Furthermore, other techniques for locating splice sites to determine the exact location of the introns and exons also need to be used. As a consequence, the technique does not have the same level of sensitivity or specificity at the nucleotide level, for genes composed of several exonic regions[33], although the overall quality of results is very similar to that afforded by other techniques[10].

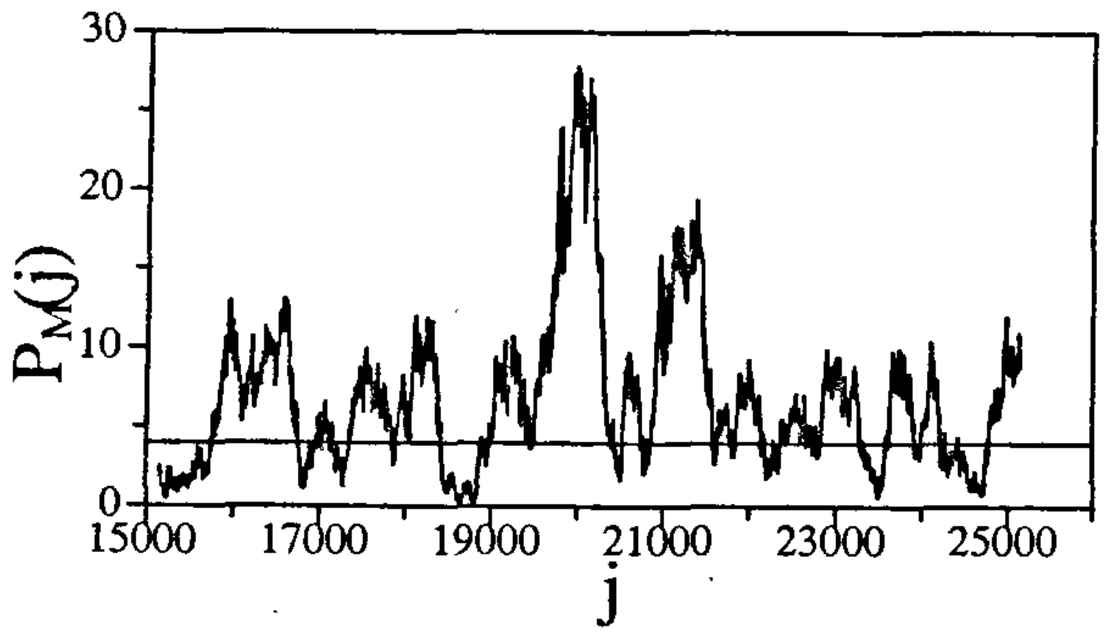

Figure 5. The results of window analysis using GeneScan on a stretch of the chromosome III of yeast. The regions where the graph exceeds the base (here taken as 4) indicate probable coding regions[33].

Table 2. Summary of results from GeneScan for the complete genome of *H. influenzae* (the quoted specificity and sensitivity are at the gene level)

ORFs reported[45]	1727
ORFs detected	1499
False positives	0
Specificity	1.0
Sensitivity	0.87
Genes reported[46]	933
Genes detected	867
Sensitivity	0.93

Coding sequence finder. The observation of the difference in correlation exponents between coding and noncoding regions can also be implemented to scan a genome to determine the potential coding regions[41]. Here also, a window of length W is moved along the genomic sequence. For each window, a double logarithmic plot of $F_d(l)$ vs l is constructed. The exponent α is obtained as the (least squares) slope of this graph, and this value of α is plotted against the position of the window (defined as the centre of the window). For noncoding regions $\alpha > 0.5$, while for coding regions $\alpha \sim 0.5$. Thus a dip in the plot indicates a probable coding region and one can essentially read the coding regions off the graph; see Figure 6.

The major drawback of this method is the large size of the window required to observe long-range correlations, which sets a limit on the smallest size of coding region detectable (usually about 1000 bp). Furthermore, the boundaries of the coding stretches are only very

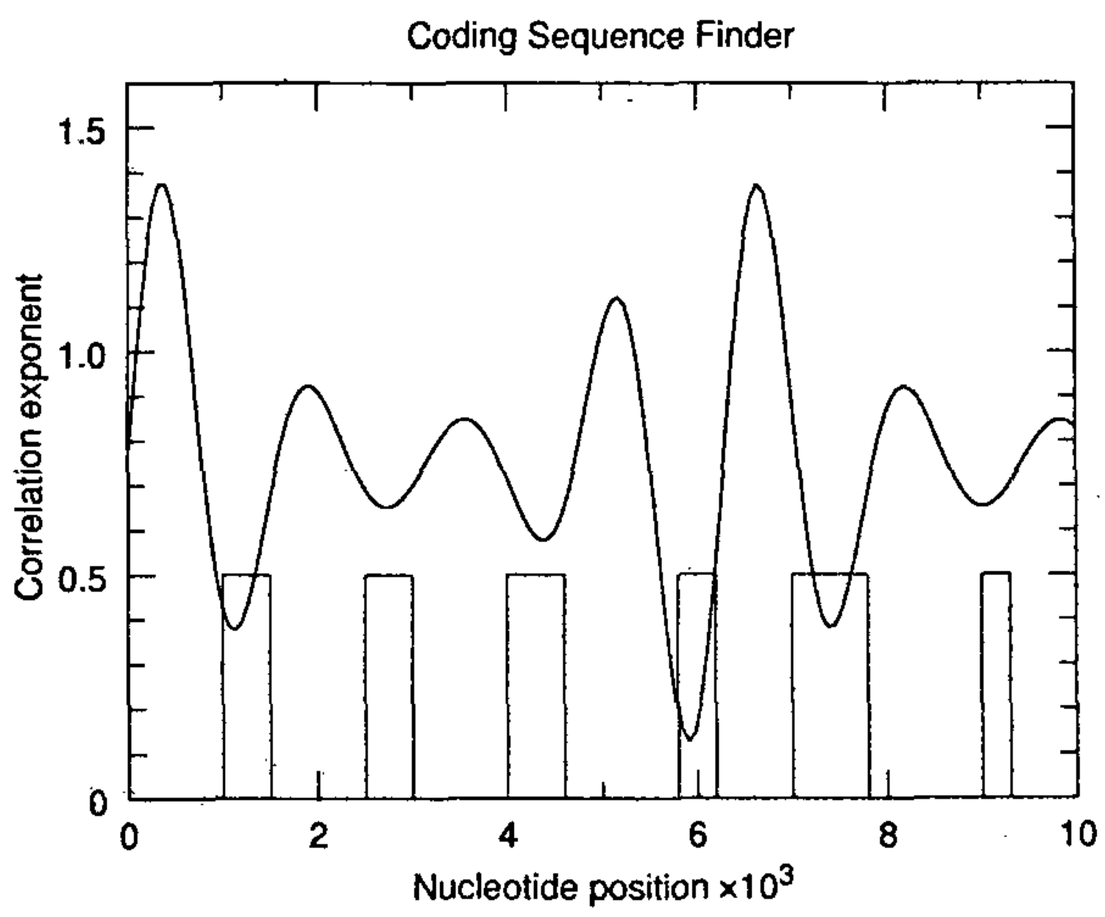

Figure 6. Typical results from the coding sequence-finder algorithm. A detrended DNA walk is constructed from a DNA sequence and the correlation exponent is determined for a stretch of about 1000 bp. The dip in the value of the scaling exponent indicates potential coding regions. (Adapted from ref. 41).

approximately determined by this method, although this can be improved by combining this with other techniques.

Discussion and summary

Rapid assessment of sequence information is possible only through computational techniques which are now being developed and refined. The different methods available have been briefly described and reviewed in the previous section, and it is clear that each algorithm or technique has something to offer.

Experience-based methods – for example those that use neural networks[20,23,24] or prior information from Markov chain analysis[16,18], are currently the most widely used. These offer the advantages of pattern recognition and pattern selection. Methods that do not rely on prior information, such as those based on correlation properties[33,41], offer, in principle, the possibility of wider applicability. It is also likely that no single technique is complete in itself. A recent review by Burset and Guigó[10] has benchmarked a variety of gene structure prediction programs against a large database, and finds that on average, the predictive accuracy of most methods ranges between 60 and 70%, for eukaryotic genes. Some techniques that are presently being developed further look for homology between the derived protein sequence and existing protein databases to completely identify genes; these are currently the most accurate methods available. However, it is perhaps unreasonable to expect that any one computer algorithm – or even a combination of several of them – will predict the location of protein coding regions with perfect specificity and sensitivity. Practical wisdom would dictate that one uses some or all of the methods available to decipher a given genome[44].

As the technology for DNA sequencing becomes increasingly sophisticated, there is bound to be a virtual flood of nucleotide sequences pouring in from a wide variety of organisms. The challenge is to make sense out of this sequence information, for the ultimate answers to the mystery of life may well lie in nature's deliberate choice of certain nucleotide sequences over others.

Data coming out from sequencing projects has brought home the realization that our level of genetic ignorance is much higher than we imagined. The example of yeast, *S. cerevisiae*, is telling. An organism which one thought was genetically beaten to pulp by decades of intensive research, has actually revealed only about one-third of its secrets to molecular geneticists. From new sequencing data – and as of today all 16 chromosomes of *S. cerevisiae* have been completely sequenced – we now know that this organism houses a much greater number of potential genes than hitherto suspected. Since it is difficult to experimentally identify genes on a mass scale given the level of technology today, it is absolutely necessary to weed out noncoding sequences (especially noncoding ORFs) so that one has fewer potential genes for experimental verification. Computational methods do provide a quick assessment of possible coding regions, and this information can help workers know where to look for probable genes and make an experimental verification of the predictions.

At the heart of it is the problem that we still do not know *all* that makes a stretch of DNA evolve into a protein-coding sequence. Only as more and more DNAs are sequenced and analysed, will the statistics will get better and we may then develop a better understanding of why some parts in DNA are coding. Mathematical methods may also provide insights into the vast amounts of non-protein-coding DNA found in complex genomes. Is this DNA truly 'junk' or is there a pattern in it which we do not comprehend? The architectural details of genomes will provide clues to the inevitable question of how life originated and evolved on this planet.

1. Lewin, B., *Genes V*, Oxford University Press, Oxford, 1994.
2. Volkenstein, M. V., *Biophysics*, MIR Publishers, Moscow, 1983.
3. Staden, R., *Nucleic Acids Res.*, 1984, **12**, 505–512.
4. Karlin, S. and Cardon, L. R., *Annu. Rev. Microbiol.*, 1994, **48**, 619–654.
5. Fleischmann, R. D. *et al.*, *Science*, 1995, **269**, 496–512.
6. Fraser, C. M. *et al.*, *Science*, 1995, **270**, 397–403.
7. Sanger, F., Nicklen, S. and Coulson, A. R., *Proc. Natl. Acad. Sci.*, 1977, **74**, 5463–5467.
8. Nowak, R., *Science*, 1995, **268**, 1134–1135.
9. Fickett, J. W., *Comput. Chem.*, 1996, **20**, 103–118.
10. Burset, M. and Guigó, R., *Genomics*, 1996, in press.
11. Staden, R., in *Patterns in Nucleic Acid Sequences*, Academic Press, New York, 1990.
12. Dong, S. and Searls, D. B., *Genomics*, 1991, **23**, 540–551.
13. Fickett, J. W. and Tung, C.-S., *Nucleic Acids Res.*, 1992, **20**, 6441–6450.
14. Fickett, J. W., *Nucleic Acids Res.*, 1982, **10**, 5303–5318.
15. Staden, R. and McLachlan, A. D., *Nucleic Acids Res.*, 1982, **10**, 141–157.
16. Borodovsky, M. *et al.*, *Molek. Biol.*, 1986, **20**, 833–840.
17. Borodovsky, M., Koonin, E. V. and Rudd, K. E., *TIBS*, 1994, **19**, 309–313.
18. Guigó, R., Knudsen, S., Drake, N. and Smith, T., *J. Mol. Biol.*, 1992, **226**, 141–157.
19. Müller, B. and Reinhardt, J., *Neural Networks*, Springer, Berlin, 1990.
20. Uberbacher, E. C. and Mural, R. J., *Proc. Natl. Acad. Sci. USA*, 1991, **88**, 11261–11265.
21. Hsu, K. and Hsu, A., *Proc. Natl. Acad. Sci.*, 1990, **87**, 938–941.
22. Claverie, J.-M., Sauvaget, I. and Bougueleret, L., *Methods Enzymol.*, 1988, **183**, 237–252.
23. Xu, Y., Mural, R. J. and Uberbacher, E. C., *Comput. Appl. Biol. Sci.*, 1994, **10**, 613–623.
24. Snyder, E. E. and Stormo, G. D., *Nucleic Acids Res.*, 1993, **21**, 607–613; *J. Mol. Biol.*, 1995, **248**, 1–18.
25. Bridle, J. S. and Sedgwick, N. C., *Proc. IEEE Int. Conf. Acoustic Speech, Signal Process.*, 1977, **27**, 656–659; Cohen, J. R., *J. Acoust. Soc. Am.*, 1981, **69**, 1430–1438.
26. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J., *J. Mol. Biol.*, 1990, **215**, 304–410.
27. Shannon, C. E. and Weaver, W., *The Mathematical Theory of Communication*, The University of Illinois Press, Urbana IL, 1964.

28. Zipf, G. K., *Human Behaviour and the Principle of Least Effort*, Addison-Wesley, Cambridge, MA, 1949.
29. Mantegna, R. N. *et al.*, *Phys. Rev. Lett.*, 1994, **73**, 3169–3172.
30. Shannon, C. E., *Bell Syst. Tech. J.*, 1948, **27**, 379.
31. See the discussion in Israeloff, N. E., Kagalenko, M. and Chan, K., *Phys. Rev. Lett.*, 1996, **76**, 1976; Bonhoeffer, S. *et al.*, *Phys. Rev. Lett.*, 1996, **76**, 1977; Voss, R. F., *Phys. Rev. Lett.*, 1996, **76**, 1978; Mantegna *et al.*, *Phys. Rev. Lett.*, 1996, **76**, 1979–1981.
32. Tsonis, A. A., Elsner, J. B. and Tsonis, P. A., *J. Theor. Biol.*, 1991, **151**, 323–331.
33. Tiwari, S., Ramachandran, R., Bhattacharya, A., Bhattacharya, S. and Ramaswamy, R., *Prediction of Probable Genes by Fourier Analysis of Genomic Sequences*, communicated.
34. Voss, R. F., *Phys. Rev. Lett.*, 1992, **68**, 3805–3808.
35. Li, W., Marr, T. G. and Kaneko, K., *Physica*, 1994, **D75**, 392–416.
36. Arneodo, A., Bacry, E., Graves, P. V. and Muzy, J. F., *Phys. Rev. Lett.*, 1995, **74**, 3293–3296.
37. Peng, C.-K. *et al.*, *Nature*, 1992, **356**, 168.
38. Peng, C.-K. *et al.*, *Phys. Rev.*, 1994, **E49**, 1685–1689.
39. Azbel, M. Y., *Phys. Rev. Lett.*, 1995, **75**, 168–171.
40. Buldyrev, S. V. *et al.*, *Phys. Rev.*, 1995, **E51**, 5084–5091.
41. Ossadnik, S. M. *et al.*, *Biophys. J.*, 1994, **67**, 64–70.
42. Jeffrey, H. J., *Nucleic Acids Res.*, 1990, **18**, 2163–2170.
43. Solovyev, V. V., Korolev, S. V. and Li, H. A., *Int. J. Genome Res.*, 1993, **1**, 108–128.
44. Most of the methods discussed in this article can be accessed via email servers. Details of how to submit a sequence for analysis can be found in the cited articles, refs. 10, 12, 14, 16, 18, 20, 24. GeneScan analysis can be obtained by sending sequences to genescan@jnuniv.ernet.in.
45. The ORFs that have been reported as coding in ref. 5, using the method GeneMark.
46. These are the genes that have been positively identified as coding through protein homology.

ACKNOWLEDGEMENTS. This research has been supported by grants from the Department of Biotechnology to A. Bhattacharya and from the Department of Science and Technology and the CSIR to R. Ramaswamy. We thank Dr S. Ramachandran for collaboration on GeneScan.

Received 26 February 1996; revised accepted 18 May 1996

Record of prolific and indubitable acritarchs from the Lower Paleozoic strata of the Tethyan Garhwal Himalaya and age implication

H. N. Sinha, S. S. Srivastava and B. Prasad*

Department of Earth Sciences, University of Roorkee, Roorkee 247 667, India
*K.D.M. Institute of Petroleum Exploration, Oil and Natural Gas Corporation Ltd, Dehradun 248 001, India

Exceptionally well-preserved and prolific fossil acritarchs (a group of acid insoluble microfossils with uncertain affinities) have been recorded from the Shiala and Yong formations of the Tethyan Garhwal Himalaya, India. The presence of age marker forms of acritarch reveals that the Ordovician/Silurian boundary lies within the Shiala Formation and not at the contact of Shiala and the Yong formations, as was proposed by earlier workers[1,2].

THE Lower Paleozoic acritarchs are known from throughout the world, especially from UK, USA, Canada, Norway, Spain, Belgium, Southern Africa, Russia, China and Arabian Sahara. However, the record from India is almost negligible due to rare occurrence of marine Paleozoic sediments in the Peninsular India. Only the simple *Leiosphaerids*, sphaeromorphs, acanthomorphs and netromorphs, are so far recorded from the marine Precambrian/Cambrian rocks of the Peninsular India (Vindhyan, Kaladgi, etc.)[3–6] and the Extrapeninsular India (Krol belt, Lesser Himalaya, Tethys Himalaya)[2,7–13]. In the Extrapeninsular India, the Lower Paleozoic marine sediments are well recognized in the Tethyan zone of Kashmir, Spiti and Kumaon–Garhwal Himalaya and contain a variety of invertebrate Paleozoic fossils.

The term 'Tethys' was conceived by Suess[14] for a long expanse of Mesozoic seaway separating the old continental masses of the Gondwanaland in the south and Angaraland in the north. The 'Tethys Himalaya' refers to the widespread sedimentary basin to the north of the central crystalline rocks[15,16]. The Tethyan sediments range in age from Precambrian/Cambrian to Early Tertiary[17] and are rich in fossil contents. The Tethyan sediments of the Garhwal Himalaya have received