Eukaryotic Gene Prediction by Spectral Analysis and Pattern Recognition Techniques

T. Eftestøl[1†], *T. Ryen*[1], *S.O. Aase*[1], *C. Strässle*[2], *M. Boos*[2], *G. Schuster*[2], *and P. Ruoff*[1]

[1]University of Stavanger
Faculty of Science and Technology
4036 Stavanger, Norway
[†]trygve.eftestol@uis.no

[2]Hochschule für Technik Rapperswil
Abteilung Electrotechnik
8640 Rapperswil, Switzerland

ABSTRACT

The problem of computational gene prediction in eukaryotic DNA is investigated. The discrete Fourier transform is used to reveal the periodicity of three which is present in the essential subregions of a gene. We introduce a novel method that allows to predict the position of genes in an optimal way (in the sense of minimal error probability) based on the complex Fourier values at the frequency 1/3. Our method is based on training and testing a bayesian classifier. We simulate gene sequences for training, apply the Fourier transform to the sequences, extract feature vectors from the spectral representation of the binary sequences and train classifiers to discriminate coding from non coding regions in the sequence. The classifier is tested on a real gene sequence where the coding and non coding regions are known.

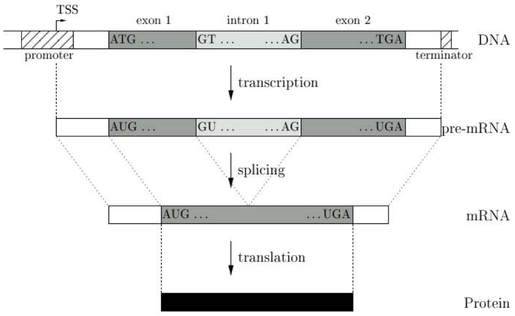

Fig. 1. Central dogma of molecular biology.

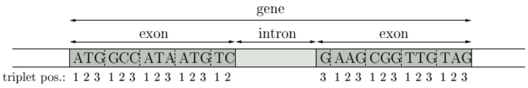

Fig. 2. Illustration of triplet position.

1. INTRODUCTION

With the use of effective sequencing methods genomic information of organisms is growing exponentially with time. One of the important and nontrivial tasks is to develop algorithms which effectively can identify genes. Genes are clustered areas within the DNA double helix which code for proteins.

In eukaryotic DNA, genes generally consist of coding regions (exons) and non coding regions (introns). Proteins are translated from a copy of the gene where introns have been removed and exons are joined together, a process called splicing (figure 1). It is therefore of importance to identify reliably the start of a gene, its exons and introns (if present) as well as the end of the gene.

A DNA sequence is a string of the four characters A, C, G, T which represent the four nucleotides. Each of the twenty possible amino acid is coded by three such nucleotides. In the exon, the nucleotides therefore exhibit a periodicity with period 3 (figure 2). Based on this property in coding regions of DNA, different methods to discriminate between coding and non coding regions using statistical methods [1], autocorrelation[2], and finally Fourier analysis[3] have been investigated.

Based on Fourier analysis different measures to discriminate between coding and non coding regions were developed. There exist the measures known as *spectral content measure*[3], *optimised spectral content measure*[4], and *spectral rotation measure*[5]. These methods apply parameters for discrimination derived from the peak at $f = 1/3$ in the Fourier spectrum[3], or optimisation of the parameters weighing the Fourier transformed binary binary indicator sequences of the nucleotides[4, 5]. Thresholds are set to distinguish parameters computed from coding parts of DNA sequences from the non coding parts[3, 4, 5].

In this study we present a novel method for of gene prediction based on the Fourier spectrum of the binary nucleotide sequences. The novelty consists in the fact that we apply both the real- and complex valued parts of the Fourier coefficients. We also determine the thresholds

1-4244-0413-4/06/$20.00 ©2006 IEEE

in an optimal manner, applying bayesian decision theory. We present preliminary results from experiments applying this novel method.

2. MATERIALS AND METHODS

Our method is based on training and testing a bayesian classifier. We simulate gene sequences for training, apply the Fourier transform to the sequences, extract feature vectors from the spectral representation of the binary sequences and train classifiers to discriminate coding from non coding regions in the sequence. The classifier is tested on a real gene sequence where the coding and non coding regions are known.

2.1. Gene data for training and testing

The Fourier analysis gene finder is trained on artificial coding and non coding sequences taken from the codon usage table of *Arabidopsis thaliana*[6]. In addition the non coding and coding regions of *Arabidopsis thaliana* chromosome I is used in the training process to estimate the á priori probabilities of the coding and non coding classes. A region containing two genes from the DNA in *Arabidopsis thaliana* chromosome IV is used for testing.

2.2. DNA Fourier Analysis

The discrete Fourier transform (DFT) $X[k]$ of a given numeric sequence $x[n]$ of length N is defined by

$$X[k] = \sum_{n=0}^{N-1} x[n]\, e^{-j\frac{2\pi}{N}nk} \qquad 0 \le k \le N-1 \tag{1}$$

where n is the sequence index and k corresponds to the discrete frequency of k/N. Because a DNA sequence D is a string of the four characters A, C, G and T which represent the four nucleotides, numerical values have to be assigned to each character. The common way is to assign a binary indicator sequence to each of the four bases[7]. The binary indicator sequence $x_b[n]$ of base b, ($b = \{\mathrm{A, C, G, T}\}$), will take a value of 1 or 0 at position n depending on the existence of base b at position n in the DNA sequence. The total estimated power spectrum $S[k]$ of the DNA sequence is defined as the sum of the four individual estimated power spectra.

Table 1. Example of binary indicator sequences.

D	G	G	A	T	A	T	C	A	C	T	T	T
$x_\mathrm{A}[n]$	0	0	1	0	1	0	0	1	0	0	0	0
$x_\mathrm{C}[n]$	0	0	0	0	0	0	1	0	1	0	0	0
$x_\mathrm{G}[n]$	1	1	0	0	0	0	0	0	0	0	0	0
$x_\mathrm{T}[n]$	0	0	0	1	0	1	0	0	0	1	1	1

$$S[k] = \sum_b |X_b[k]|^2 \qquad f = \frac{k}{N} \tag{2}$$

Figures 3 a and b show the total power density spectrum of a typical coding and a non coding region respectively. The periodicity with period 3 of a coding region can be seen as a distinct peak at frequency $f = 1/3$ in the power density spectrum. Since the only significant difference between the spectra of coding and non coding regions is at $f = 1/3$ or $k = N/3$ it is sufficient to evaluate the DFT $X_b[k]$ at $k = N/3$.

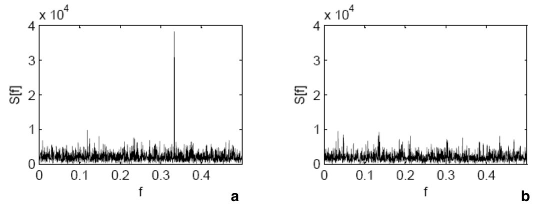

Fig. 3. Power density spectrum of coding (a) and non coding (b) section.

$$X_b\left[\frac{N}{3}\right] = \sum_{n=0}^{N-1} x_b[n]\, e^{-j\frac{2\pi}{3}n} \tag{3}$$

Note that for N being a multiple of 3 the sum over the four values $X_b[N/3]$ will always be zero.

$$X_\mathrm{A}\left[\frac{N}{3}\right] + X_\mathrm{C}\left[\frac{N}{3}\right] + X_\mathrm{G}\left[\frac{N}{3}\right] + X_\mathrm{T}\left[\frac{N}{3}\right] = 0 \tag{4}$$

2.3. Feature extraction

In order to have a measure that tells us if the nucleotides at a certain position in the analysed DNA sequence D belong to a coding region or not, a sliding analysis window of length N (with N being an odd multiple of three) is introduced. The sliding window is moved by the step size Δ and has its centre at position i relative to the beginning of D. The siding window positions are numbered using index variable m={0,1,2,...}.

Taking the random variables $Z_b = X_b[N/3]$ where ($b = \{\mathrm{A, C, G, T}\}$) as a basis for the optimal decision according to the Bayesian decision theory[8], we construct our feature vectors of the real and imaginary parts of these variables according to

$$\mathbf{z} = \begin{bmatrix} Re\{X_{\mathrm{C},i}[N/3]\} \\ Re\{X_{\mathrm{G},i}[N/3]\} \\ Re\{X_{\mathrm{T},i}[N/3]\} \\ Im\{X_{\mathrm{C},i}[N/3]\} \\ Im\{X_{\mathrm{G},i}[N/3]\} \\ Im\{X_{\mathrm{T},i}[N/3]\} \end{bmatrix} \quad \text{with} \quad i = \frac{N-1}{2} + m\Delta \tag{5}$$

For the training vectors to be independent of each other the sliding windows are chosen to be non overlapping ($\Delta = N$).

Z_{A} is omitted from the feature vector because it is implicitly given by equation (4) when Z_{C}, Z_{G}, and Z_{T} are known.

2.4. Classification

We denote class ω_C and ω_{NC} as the coding and non coding classes of patterns respectively.

According to Bayes decision rule[9] the current analysis window would be classified as coding when the following inequality is true:

$$P(\omega_C)p(\mathbf{z}|\omega_C) > P(\omega_{NC})p(\mathbf{z}|\omega_{NC}) \tag{6}$$

$P(\omega_i)$ and $p(\mathbf{z}|\omega_i), i = 1, 2$ denotes the á priori probabilities and class specific probability density functions respectively.

The á priori probabilities reflect the ratio between the total length of coding regions and non coding regions in the genome of a given organism and are calculated from training data.

2.5. Experimental setup

The Fourier analysis gene finder is trained on artificial DNA sequences. The coding sequence which serves for the calculation of the coding distribution $p(\mathbf{z}|\omega_C)$ is a memoryless random sequence of codons (nucleotide triplets). The non coding sequence which serves for the calculation of the coding distribution $p(\mathbf{z}|\omega_{NC})$ is a weighted memoryless random sequence of nucleotides[6]. The probabilities used for weighing the random sequences are taken from the codon usage table of *Arabidopsis thaliana*[6].

The á priori probabilities are estimated as

$$P(\omega_C) = \frac{n_C}{n_C + n_{NC}} \tag{7}$$

and

$$P(\omega_{NC}) = \frac{n_{NC}}{n_C + n_{NC}} \tag{8}$$

where n_C and n_{NC} are the number of coding non coding triplets in *Arabidopsis thaliana* chromosome I respectively.

To estimate the probability density functions, both maximum likelihood (ML) estimation and the Parzen windowing technique with window width varied to achieve good correspondence between training and testing is used[9].

The training vectors for each class $\mathbf{z}$ are calculated from 200 non overlapping sliding windows of length N ($\Delta = N$) for N=351 – which is a common window length in DNA Fourier analysis[3]. 200 testing vectors for each class are calculated in a similar manner. The probability density functions are estimated from the training vectors. The classifier is evaluated by first reclassifying the training vectors and then classifying the test vectors using the estimates.

The best classifier with a reasonable match between reclassification and test performance is chosen for further classification of a subsequence containing two genes of a real DNA sequence, the DNA in *Arabidopsis thaliana* chromosome IV is analysed for this purpose. The feature vectors $\mathbf{z}$ are calculated from non overlapping sliding windows of length N ($\Delta = N$) for N=351 from the DNA sequence. For each feature vector the true class, ω_C of ω_{NC} is known from the DNA information provided. Using the estimated probabilities and density functions, each feature vector is classified according to equation 6. Now, when analysing an unknown sequence the codons inside the exons can be at any position relative to the sliding window . Therefore the density function values has to be calculated three times, once for every possible position of the codons. Thus the densities are extended by the phase parameter $\varphi = \{0, \frac{2\pi}{3}, \frac{4\pi}{3}\}$ and the ratio,

$$M_{\mathrm{bay}}[N, i, \varphi] = \frac{p(\mathbf{z}|\omega_C)}{p(\mathbf{z}|\omega_{NC})} \tag{9}$$

is calculated for each phase.

The error rate and sensitivity and specificity for each classifier is reported.

3. RESULTS

Table 3 show the classifier performances after training reclassification and testing on simulated DNA sequences: The ML classifier shows the best performance. The Parzen window based classifier with the largest window size (hn=1) has considerably lower performance. Both these classifiers have acceptable correspondence between reclassification and testing performance which is not the case for the Parzen classifier with narrow window width (hn=0.1). Therefore the ML classifier is used in the further detection of coding regions in real DNA.

Table 2. Performance characteristics.The error rate, sensitivity and specificity for discriminating coding from non coding regions in artificial DNA sequences. The results are presented as reclassification/testing performances

Classifier	Error	Sensitivity	Specificity
ML	0.05/0.06	0.89/0.88	1.00/0.99
P (hn=0.1)	0.34/0.47	0.32/0.23	1.00/0.84
P (hn=1)	0.45/0.48	0.10/0.08	1.00/0.97

Table 3. Performance characteristics. The error rate, sensitivity and specificity for discriminating coding from non coding regions in a real DNA sequence using a ML classifier trained on simulated data.

Phase	Error	Sensitivity	Specificity
1	0.32	0.44	0.93
2	0.44	0.17	0.94
3	0.46	0.11	0.96

Table 3 shows the classifier performances after applying the ML classifier to the real DNA sequence: Figure 4 shows how the function $\frac{p(\mathbf{z}|\omega_C)}{p(\mathbf{z}|\omega_{NC})}$ changes throughout the analysed sequence. A coding region is detected when the function exceeds the threshold value $\frac{P(\omega_{NC})}{P(\omega_C)}$ corresponding to applying the decision rule in equation 6. The results are not convincing. The results from testing the classifier based on artificial data do not correspond satisfactorily with the performance achieved when challenging this classifier with real data.

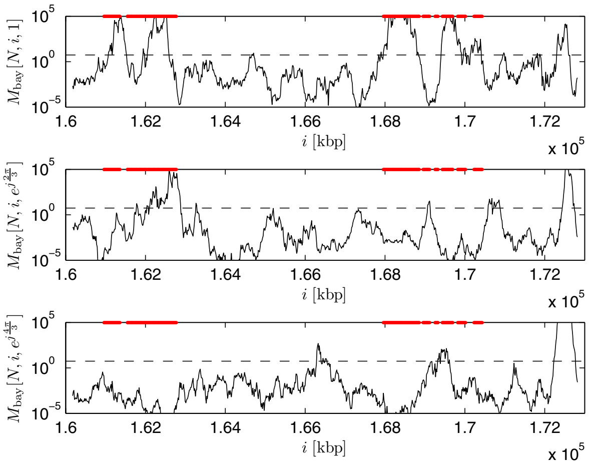

Fig. 4. Detection of coding regions. Function values exceeding the threshold corresponds to a detection of a coding region. True coding triplets are marked in red.

One problem is obviously that the model used for generating the artificial data is unreliable. This assumption is supported by the large deviation of error rates when testing the classifier on artificial data as compared to the increase in error of 0.27 (best case) when predicting true coding regions. One possible solution to this problem would be to improve the model. Alternatively the classifier can be trained on real data. In this case, aspects of constructing of a representative data set for training the classifier has to be considered.

Another problem that has to be handled in future work is the different phase signals. A possible way might be to calculate an expanded feature vector covering all phases and use a principal component decomposition to a lower dimensionality to capture the dominant phase.

4. CONCLUSION

We have proposed a method to detect coding regions in DNA sequences based on Fourier analysis in the framework of Bayes classification.

The preliminary results presented in this paper indicates that the model used to generate artificial data for training of the classifier is inadequate. Future work should focus on improving the model or generating a training set of real DNA sequences.

REFERENCES

[1] R. Staden and A. D. McLachlan. Codon preference and its use in identifying protein coding regions in long DNA sequences. *Nucleic Acids Res*, 10:141–156, 1982.

[2] J. W. Fickett. Recognition of protein coding regions in DNA sequences. *Nucleic Acids Research*, 10:5303–5318, 1982.

[3] S. Tiwari, S. Ramachandran, A. Bhattacharya, S. Bhattacharya, and R. Ramaswamy. Prediction of probable genes by Fourier analysis of genomic sequences. *Comput Appl Biosci*, 13:263–270, 1997.

[4] D. Anastassiou. Frequency-domain analysis of biomolecular sequences. *Bioinformatics*, 16:1073–1081, 2000.

[5] Daniel Kotlar and Yizhar Lavner. Gene prediction by spectral rotation measure: a new method for identifying protein-coding regions. *Genome Research*, 13:1930–1937, 2003.

[6] R. Grantham, C. Gautier, M. Gouy, R. Mercier, and A. Pavé. Codon catalog usage and the genome hypothesis. *Nucleic Acids Research*, 8:r49–r62, 1980.

[7] R. Voss. Evolution of long-range fractal correlations and 1/f noise in DNA base sequences. *Physical Review Letters*, 68:3805–3808, 1992.

[8] Jürgen Schürmann. *Pattern Classification: A Unified View of Statistical and Neural Approaches*. John Wiley & Sons, Inc., 1996.

[9] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. John Wiley and Sons, Inc., New York, second edition, 2001.