SCORING OF CODING REGIONS IN DNA SEQUENCES BASED ON POSITION ASYMMETRY

Thomas A. Jost, Ramon F. Brcich and Abdelhak M. Zoubir

Signal Processing Group, Institute of Telecommunications
Darmstadt University of Technology
Merckstrasse 25, D-64283 Darmstadt, Germany.

ABSTRACT

At present the genomes of many organisms have been sequenced, meaning that the nucleotide structure is known but the location of genes, and most importantly, the coding regions, are unknown. Locating the coding regions is vital as they code for the proteins which control the functioning of the organism, such as its resistance to disease. We propose a new algorithm to score genomic sequences. The algorithm is based on discriminant analysis and can be incorporated into existing programs to analyse DNA sequences.

1. INTRODUCTION

Deoxyribonucleic acid (DNA) is composed of four nucleotides containing the bases Adenine (A), Cytosine (C), Thymine (T) and Guanine (G) connected in sequence. Thus, DNA molecules can be seen as a string created with the alphabet $\{\mathtt{A}, \mathtt{T}, \mathtt{C}, \mathtt{G}\}$. A DNA sequence can be described as genes containing coding regions separated by non-coding regions. Coding regions within genes, called exons, code for the proteins which determine the organism's structure and functioning. Introns are non-coding regions within a gene which separate exons.

Location of genes, specifically exons, within the DNA sequence is an important first step in analysing the genome of an organism. Computational methods for predicting gene locations have been investigated for over 20 years. In that time, many and various techniques have been proposed to detect genes. The biggest challenge in this context is to find small coding regions. Aside from standard molecular methods, algorithms have been developed such as GENSCAN [1], GeneMarkS [2], MZEF [3] and many others [4]. The core units of those algorithms are scoring, i.e., finding the likely start and end of coding regions and scoring the coding region itself. To score a region many approaches [5] have been proposed but most are based on statistical characteristics like hexamer counts or periodicities in the occurrence of a base [6]. Other approaches include entropy measures [7] and, currently, Markov models [8] are being heavily investigated.

A disadvantage of most of these methods is the large training sequence needed to distinguish between coding and non-coding regions. This is especially important for the widely used 5th order Markov models which need to estimate 12288 parameters from training data. A recently published algorithm [9] called the Z-curve claims to give better results than the 5th order Markov model with less than 200 parameters to estimate from training data.

Our proposed method follows the Z-curve in that it is based on the frequency of occurrence of a nucleotide in each of the three possible positions of a codon, but instead of using a linear discrimination function we use a quadratic one.

2. MEASUREMENT VECTORS

2.1. Proposed approach

As 3 bases are required to define an amino acid, coding regions are often represented as a consecutive sequence of 3 bases called codons. The proposed method is based on the frequency of occurrences of base $j \in \{\mathtt{A}, \mathtt{T}, \mathtt{C}, \mathtt{G}\}$ in codon position $c \in \{1, 2, 3\}$. In coding regions it has been observed that there is a strong bias in the probability with which a base occurs in a specific codon position away from what would be expected if bases occurred totally at random, i.e., with equal probability in each codon position [10]. In non-coding regions there appears to be no bias in the occurrence of a base, i.e., bases occur totally at random or with equal probability in each codon position. The bias can be explained by a preference for certain codons brought about by the redundancy of the genetic code, i.e., the 64 possible codons are translated into only 20 amino acids, meaning that several codons represent a specific amino acid. Although these codons usually only differ in the last base or two, out of the set of codons representing a specific amino acid, some are preferred.

To score a genomic region of window length L using mono-nucleotides (a single base), where L is a multiple of 3, a contingency table is constructed where $N_{j,c}$ is the frequency of occurrences of base j in codon position c. Due to the nature of the DNA sequence it is always true that

	Codon position c		
Base	$c=1$	$c=2$	$c=3$
A	$N_{\mathtt{A},1}$	$N_{\mathtt{A},2}$	$N_{\mathtt{A},3}$
C	$N_{\mathtt{C},1}$	$N_{\mathtt{C},2}$	$N_{\mathtt{C},3}$
T	$N_{\mathtt{T},1}$	$N_{\mathtt{T},2}$	$N_{\mathtt{T},3}$
G	$N_{\mathtt{G},1}$	$N_{\mathtt{G},2}$	$N_{\mathtt{G},3}$

$$N_{j,c} = N - \sum_{\substack{l \in \{\mathtt{A},\mathtt{C},\mathtt{T},\mathtt{G}\} \\ l \neq j}} N_{l,c}$$

where $N = \frac{L}{3}$. Utilising this redundancy, it is possible to discard one of the bases without losing information. Collecting the non-redundant frequencies of occurrence into a feature vector $\mathbf{X}_m$ (where m stands for mono-nucleotide) gives

$$\begin{aligned}\mathbf{X}_m &= [N_{\mathtt{A},1}\ N_{\mathtt{C},1}\ N_{\mathtt{T},1}\ N_{\mathtt{A},2}\ N_{\mathtt{C},2}\ N_{\mathtt{T},2}\ N_{\mathtt{A},3}\ N_{\mathtt{C},3}\ N_{\mathtt{T},3}]^T \\ &= \langle N_{j,c} \rangle, \qquad j \in \{\mathtt{A},\mathtt{C},\mathtt{T}\},\ c \in \{1,2,3\}\end{aligned}$$

where $(\cdot)^T$ is the transpose operator.

To extend this concept it is possible to use di-nucleotides (pairs of bases) instead of mono-nucleotides. Proceeding in the same manner as for mono-nucleotides we build a contingency table of dimension 16x3 with entries $N_{ji,c}$, $j,i \in \{\mathtt{A},\mathtt{C},\mathtt{T},\mathtt{G}\}$. $N_{ji,c}$ is defined as the frequency of occurrences of the following pair: base j in codon position c followed by base i in position $c+1$. If $c=3$, $c+1$ becomes $c=1$, i.e., the first codon position of the next codon. Again, a feature vector $\mathbf{X}_d$ (where d stands for di-nucleotide) of dimension 64x1 can be created from the elements of the contingency table,

$$\mathbf{X}_d = \langle N_{ji,c} \rangle, \qquad j,i \in \{\mathtt{A},\mathtt{C},\mathtt{T},\mathtt{G}\},\ c \in \{1,2,3\}$$

Removing the redundancies in $\mathbf{X}_d$ results in a feature vector with 42 linearly independent parameters. For di-nucleotides we removed the parameters $N_{\mathtt{AA},1}$, $N_{\mathtt{GG},3}$, $N_{\mathtt{C}j,1}$ with $j \in \{\mathtt{C},\mathtt{T},\mathtt{G}\}$ and $N_{\mathtt{CA},2}$.

A further extension can be made by using the frequencies of occurrence of tri-nucleotides (triplets of bases), $N_{jik,c}$, $j,i,k \in \{\mathtt{A},\mathtt{C},\mathtt{T},\mathtt{G}\}$, starting at codon position c. Similarly to the case of mono-nucleotides and di-nucleotides a feature vector $\mathbf{X}_t$ (where t stands for tri-nucleotide) can be defined as

$$\mathbf{X}_t = \langle N_{jik,c} \rangle, \qquad j,i,k \in \{\mathtt{A},\mathtt{C},\mathtt{T},\mathtt{G}\},\ c \in \{1,2,3\}$$

which has 192 elements from the contingency table of dimension 64x3. Again, removing redundancies results in a feature vector with 174 linearly independent parameters. We removed the counts $N_{\mathtt{AAA},1}$, $N_{\mathtt{AAA},2}$ and $N_{ij\mathtt{G},3}$ with $i,j \in \{\mathtt{A},\mathtt{C},\mathtt{T},\mathtt{G}\}$.

It can be shown that $\mathbf{X}_t$ is composed of linear combinations of the elements of $\mathbf{X}_m$ and $\mathbf{X}_d$.

2.2. Z-curve

Instead of directly using the frequencies of occurrence the Z-curve uses a linear combination of them to define a three dimensional point with the coordinates

$$\begin{aligned}x_c &= (N_{\mathtt{A},c} + N_{\mathtt{G},c}) - (N_{\mathtt{C},c} + N_{\mathtt{T},c}) \\ y_c &= (N_{\mathtt{A},c} + N_{\mathtt{C},c}) - (N_{\mathtt{G},c} + N_{\mathtt{T},c}) \\ z_c &= (N_{\mathtt{A},c} + N_{\mathtt{T},c}) - (N_{\mathtt{G},c} + N_{\mathtt{C},c})\end{aligned}$$

for the case of mono-nucleotides with $c \in \{1,2,3\}$. The feature vector $\mathbf{X}_{Z,m}$ of the Z-curve is then defined as

$$\begin{aligned}\mathbf{X}_{Z,m} &= [x_1\ y_1\ z_1\ x_2\ y_2\ z_2\ x_3\ y_3\ z_3]^T \\ &= \langle x_c\, y_c\, z_c \rangle\,.\end{aligned}$$

Using di-nucleotides for each base j and codon position c, a three dimensional point with coordinates

$$\begin{aligned}x_{j,c} &= (N_{j\mathtt{A},c} + N_{j\mathtt{G},c}) - (N_{j\mathtt{C},c} + N_{j\mathtt{T},c}) \\ y_{j,c} &= (N_{j\mathtt{A},c} + N_{j\mathtt{C},c}) - (N_{j\mathtt{G},c} + N_{j\mathtt{T},c}) \\ z_{j,c} &= (N_{j\mathtt{A},c} + N_{j\mathtt{T},c}) - (N_{j\mathtt{G},c} + N_{j\mathtt{C},c}) \\ &\qquad j \in \{\mathtt{A},\mathtt{C},\mathtt{T},\mathtt{G}\},\ c \in \{1,2,3\}\end{aligned}$$

is constructed. The 36 element feature vector for *phase-specific* di-nucleotides [9] is

$$\mathbf{X}_{Z,d} = \langle x_{j,c}\, y_{j,c}\, z_{j,c} \rangle, \qquad j \in \{\mathtt{A},\mathtt{C},\mathtt{T},\mathtt{G}\},\ c \in \{1,2,3\}\,.$$

Using *phase-specific* tri-nucleotides for the Z-curve the definition is quite similar with coordinates

$$\begin{aligned}x_{ji,c} &= (N_{ji\mathtt{A},c} + N_{ji\mathtt{G},c}) - (N_{ji\mathtt{C},c} + N_{ji\mathtt{T},c}) \\ y_{ji,c} &= (N_{ji\mathtt{A},c} + N_{ji\mathtt{C},c}) - (N_{ji\mathtt{G},c} + N_{ji\mathtt{T},c}) \\ z_{ji,c} &= (N_{ji\mathtt{A},c} + N_{ji\mathtt{T},c}) - (N_{ji\mathtt{G},c} + N_{ji\mathtt{C},c})\end{aligned}$$

where $i,j \in \{\mathtt{A},\mathtt{C},\mathtt{T},\mathtt{G}\}$ and $c \in \{1,2,3\}$. The feature vector follows directly as

$$\mathbf{X}_{Z,t} = \langle x_{ji,c}\, y_{ji,c}\, z_{ji,c} \rangle\,.$$

In [9] a biological interpretation for the Z-curve was given to motivate its use.

3. SCORING FUNCTIONS

In [9] a linear transformation of the measurement vector $\mathbf{X}_Z$

$$T(\mathbf{X}_Z) = V^T \mathbf{X}_Z \, .$$

was used to score a part of DNA as either coding or non-coding. To construct $\mathbf{X}_Z$ either $\mathbf{X}_{Z,m}$ was used alone, $\mathbf{X}_{Z,m}$ and $\mathbf{X}_{Z,d}$ were merged, or a combination of $\mathbf{X}_{Z,m}$, $\mathbf{X}_{Z,d}$ and $\mathbf{X}_{Z,t}$ were taken. As $\mathbf{X}_{Z,m}$ has 9 elements, $\mathbf{X}_{Z,d}$ has 36 elements and $\mathbf{X}_{Z,t}$ has 144 elements, $\mathbf{X}_Z$ has 189 elements.

In quadratic discriminant analysis (QDA) a second order test statistic $T(\mathbf{X}_l)$

$$T(\mathbf{X}_l) = \mathbf{X}_l^T Q \mathbf{X}_l + V^T \mathbf{X}_l$$

with $l \in \{m, d, t\}$ is used.

Fisher's criterion [11] states that the optimal values of Q and V maximise

$$\arg\max_{Q,V} \frac{(\mathsf{E}[T(\mathbf{X})|K] - \mathsf{E}[T(\mathbf{X})|H])^2}{\mathsf{Var}[T(\mathbf{X})|K] + \mathsf{Var}[T(\mathbf{X})|H]}$$

where H denotes the null hypothesis of non-coding regions and K denotes the alternative hypothesis of coding regions.

Note that since the feature vector for the Z-curve is a linear transformation of the frequencies of occurrence of nucleotides in the codon positions, the approaches of Sections 2.1 and 2.2 are exactly equivalent when used in linear or quadratic discrimination.

Because the QDA is only optimal for Gaussian distributions, another approach is to model the probability density function of coding and non-coding regions separately by a finite Gaussian mixture. By using a mixture of Gaussian curves a non-Gaussian distribution can be better approximated. For estimation of the parameters the Expectation-Maximisation (EM) Algorithm [13] was been used. To score the DNA region the log-likelihood (i.e. as used in a likelihood ratio test)

$$\log \frac{f(\mathbf{X}_l|K)}{f(\mathbf{X}_l|H)}$$

was used.

4. RESULTS

The comparison we present is based on yeast (Saccharomyces cerevisiae) chromosome XVI with Genbank accession number NC_001148 [12]. For non-coding regions we took all bases between genes. For coding regions we took all which are read in the forward direction and were longer than the set window size.

The arbitrarily chosen sequence length was $L = 60$ and the analysis was performed on non-overlapping segments. As training data the yeast chromosome I-VII with Genbank accession numbers NC_001133-NC_001139 was used. To compare the methods we show the obtained operating characteristic (OC) curve, which gives the probability of a true positive (TP), (detection rate) versus the probability of a false positive (FP) (false alarm rate). In the simulations, the threshold was determined to obtain a preset false alarm rate in non-coding DNA regions. The upper left point is the optimum with zero false alarm rate and 100% detection rate.

Figure 1 shows the OC of mono-nucleotides using QDA, Z-curve and Gaussian mixtures where N denotes the number of components. The QDA is superior compared to the Z-curve. Further improvements through using Gaussian mixtures could not be made. This is because $f(\mathbf{X}_l|K)$ and $f(\mathbf{X}_l|H)$ can be well approximated by a Gaussian distribution. For illustration Figure 2 shows the sample distribution for mono-nucleotides as bars together with a Gaussian distribution which has the same mean and variance. As it can be seen the central-limit-theorem (CLT) holds and the feature vector can be approximated by a multivariate Gaussian distribution. This suggests that the second order information contained in the covariance structure has discriminative ability, otherwise the linear and quadratic test statistics would produce the same result. Note that this is consistent with [10] where a kind of variance measure is used to discriminate between coding and non-coding regions. Further improvements using Gaussian mixtures could not be made.

In the case of using di-nucleotides Figure 4 shows that QDA still outperforms the Z-curve measure. The increasing number of parameters results in two drawbacks for the Gaussian mixture such that it basically fails completely for di- and tri-nucleotides and those results are omitted. First the measurement vector space increases resulting in a more difficult search problem for the EM algorithm and second, because the window length stays the same, the number of counts drops such that the CLT is not fulfilled and the distribution of the data is better approximated by a multivariate binomial or Poisson distribution as shown in Figure 3.

This is even more visible in the case of tri-nucleotides where the distribution is shown in Figure 5 and the OC in Figure 6. Because of the computational burden using QDA a subset feature selection algorithm [11] was used to reduce the dimension of the feature vector. In Figure 6 a subset of 70 parameters out of 174 were used in QDA extracted by a forward selection tree algorithm.

5. CONCLUSION

We proposed a new method for scoring DNA sequences to discriminate between coding and non-coding regions based on position asymmetry. As scoring functions we used a second order test statistic and the log-likelihood ratio with distributions under the null and the alternative modeled as

Gaussian mixtures. The QDA is superior compared to the Z-curve. In future work we will try to model the distributions for coding and non-coding sequences by multivariate Poisson distributions.

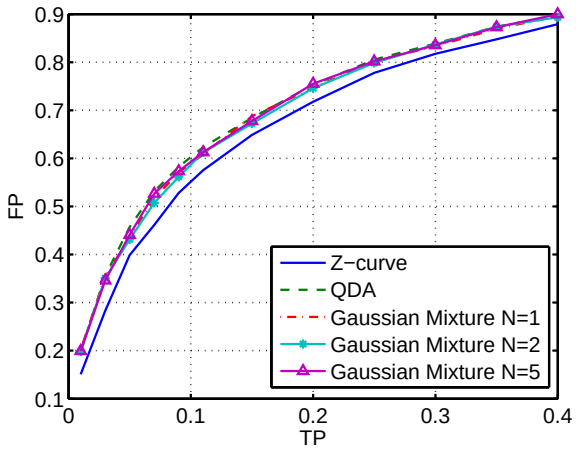

Fig. 1. OC for window length $L = 60$ using mono-nucleotides.

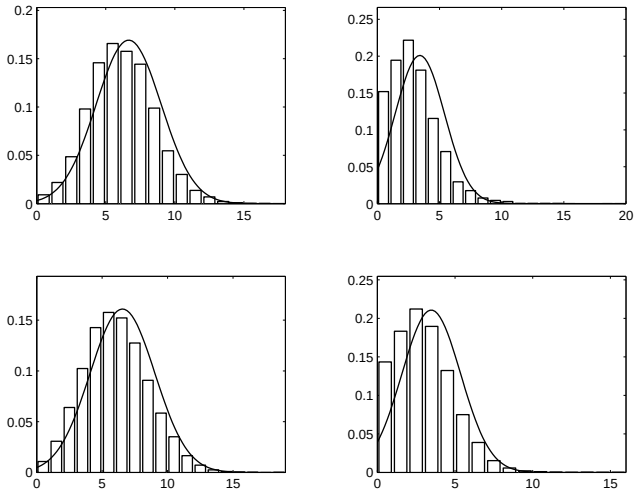

Fig. 2. Histograms for mono-nucleotides with window length $L = 60$. The upper two figures show the histograms of $N_{\mathtt{A},1}$ and $N_{\mathtt{C},1}$ in coding regions together with a Gaussian distribution having the same mean and variance. The lower figures show the same for non-coding regions.

6. REFERENCES

[1] C. Burge and S. Karlin, "Prediction of complete gene structures in human genomic dna," *Journal of Molecular Biology*, vol. 268, pp. 78–94, 1997.

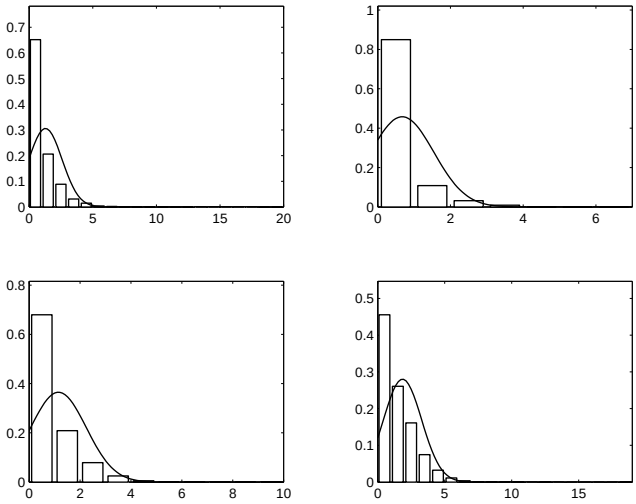

Fig. 3. Histograms for di-nucleotides with window length $L = 60$. The upper two figures show the histograms of $N_{\mathtt{AA},1}$ and $N_{\mathtt{CA},1}$ in coding regions together with a Gaussian distribution having the same mean and variance. The lower figures show the same for non-coding regions.

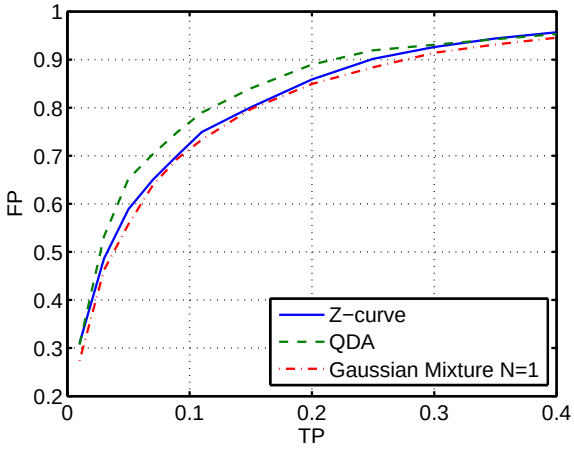

Fig. 4. OC for window length $L = 60$ using di-nucleotides.

[2] J. Besemer, A. Lomsadze, and M. Borodovsky, "Genemarks: a self-training method for prediction of gene starts in microbial genomes. implications for finding sequence motifs in regulatory regions," *Nucleic Acids Research*, vol. 29, no. 12, pp. 2607–2618, 2001.

[3] M.Q. Zhang, "Identification of protein coding regions in the human genome by quadratic discriminant analysis," *Proc. Natl. Acad. Sci.*, vol. 94, pp. 565–568, 1997.

[4] M.Q. Zhang, "Computational prediction of eukaryotic

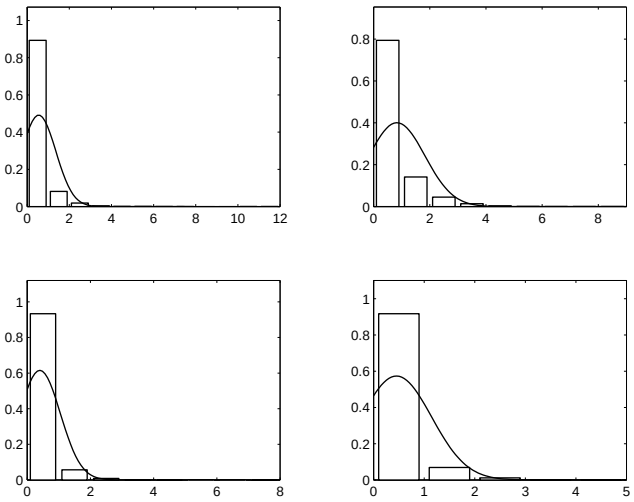

Fig. 5. Histograms for tri-nucleotides with window length $L = 60$. The upper two figures show the histograms of $N_{\mathtt{AAA},1}$ and $N_{\mathtt{TAA},1}$ in coding regions together with a Gaussian distribution having the same mean and variance. The lower figures show the same for non-coding regions.

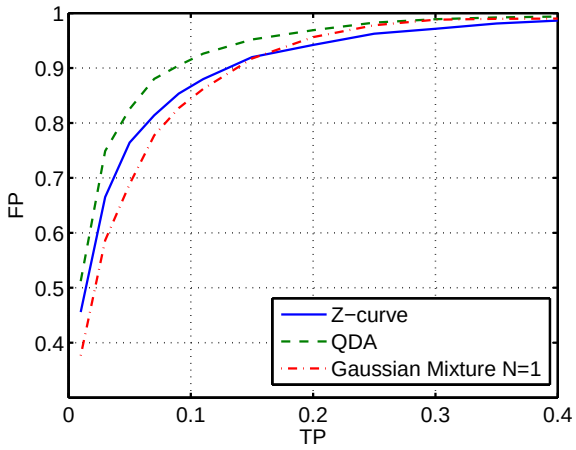

Fig. 6. OC for window length $L = 60$ using tri-nucleotides.

protein-coding genes," *Nature*, vol. 3, pp. 698–710, 2002.

[5] J.W. Fickett and C.-S. Tung, "Assessment of protein coding measures," *Nucleic Acids Research*, vol. 20, no. 24, pp. 6441–6450, 1992.

[6] D. Anastassiou, "Genomic signal processing," *IEEE Signal Processing Magazine*, pp. 8–20, 2000.

[7] G. L. Rosen and J. D. Moore, "Investigation of coding structure in DNA," in *ICASSP 2003*, Hong Kong, China, Apr. 2003, pp. 361–364.

[8] M. Borodovsky and J. McInnich, "Genmark: Parallel gene recognition for both dna stands," *Computers Chem.*, vol. 17, no. 2, pp. 123–133, 1993.

[9] F. Gao and C.T. Zhang, "Comparison of various algorithms for recognizing short coding sequences of human genes," *Bioinformatics*, vol. 20, no. 5, pp. 673–681, 2004.

[10] R. Staden, "Measurements of the effects that coding for a protein has on a dna sequence and their use for finding genes," *Nucleic Acids Research*, vol. 12, no. 1, pp. 551–567, 1984.

[11] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press, 1990.

[12] National Center for Biotechnology Database, "Genbank," *http://www.ncbi.nlm.nih.gov/entrez*.

[13] G.J. McLachlan, *Finite Mixture Models*, Wiley, 1999.