Poster Abstract from O. Zimmermann and D. Schomburg, *German Conference on Bioinformatics*, Cologne 1998.

Stochastic Language Models for Content Based DNA Sequence Classification

Uwe Ohler and Heinrich Niemann
Lehrstuhl für Mustererkennung (Informatik 5)
Universität Erlangen-Nürnberg
Martensstraße 3, D-91058 Erlangen
Phone: +49-9131/857775, Fax: +49-9131/303811
eMail: ohler@informatik.uni-erlangen.de

1 Introduction

Methods for computer based analysis of DNA primary sequences can be classified into two groups: those which rely on the detection of characteristic short sequences (search by signal), and methods which judge a region as a whole (search by content). Here we introduce a new content based classification method, an approach using stochastic language models (SLMs) which consistently outperforms conventional measures. SLMs have been used successfully by our group for speech recognition applications [4] and turn out to be a straightforward generalization of oligomer measures. We will show the basic concepts of this method, compare different possible techniques and demonstrate the power of this approach by means of the discriminination between coding and noncoding regions. Using SLMs, we also obtained excellent results on the classification of regulatory regions which are described in detail elsewhere [2, 3].

2 Methods

A stochastic language model consists essentially of interpolated Markov chains of different order. Using one part of the training set, maximum likelihood estimates of all conditional probabilities for oligomers $w_1 \dots w_n$ of length n up to certain context length N are obtained by

$$\tilde{P}(w_n|w_1 \dots w_{n-1}) = \frac{\#(w_1 \dots w_n)}{\#(w_1 \dots w_{n-1})},$$

where # denotes the frequency of its argument in the training sample. Then, using a second part of the training sample, the probabilities for the oligos of length N are re-estimated. This can be done for example by *linear interpolation* between all the shorter subsequences:

$$\hat{P}(w_N|w_1 \dots w_{N-1}) = \rho_0 \frac{1}{L} + \rho_1 \tilde{P}(w_N) + \rho_2 \tilde{P}(w_N|w_{N-1}) + \dots + \rho_N \tilde{P}(w_N|w_1 \dots w_{N-1}),$$

where the fraction $(1/L)$ accounts for unseen events, or by *rational interpolation* which in addition to linear interpolation takes into account how often each particular oligomer appeared in the training sample.

The interpolation weights ρ_i can be optimized with respect to the Maximum Likelihood criterion by using an Expectation Maximization approach in the linear case or gradient descent methods in the rational case. Setting all the weights $\rho_0 \dots \rho_{N-1}$ to zero and ρ_N to one results in the well-known N-mer approach. Using interpolation though, the model can take into account statistics of a higher order without running into the danger to overfit the model to the training data.

To classify a new sequence into several classes, e.g. coding and non-coding sequences, one can construct an SLM for each of the considered classes, compute the likelihood of an unseen sequence $w = w_1 \dots w_T$ with each model and classify the sequence with the Bayes rule. Furthermore, the classification can be biased with respect to sensitivity and specificity, e.g. of the exon recognition, by using a normalized threshold value on the difference between the model outputs.

Method	hum54	hum108	hum162
Hexamer Measure	70.5	73.1	74.2
SLM (length six)	72.8	77.3	78.5
Dicodon Measure	80.7	84.3	85.4
Dicodon SLM	81.8	85.5	86.8
Diamino Acid Measure	77.2	84.9	87.7
Diamino Acid SLM	78.7	86.0	89.0

Table 1: Comparison of the results obtained by the hexamer and SLM method on human coding and non-coding sequences of different length. The first two rows give the results for frame independent, the remaining for frame specific models.

3 Data and Results

For training and evaluation of the methods, we used the benchmark data sets of the coding measure survey by Fickett and Tung [1]. These sets contain coding and non-coding sequences of human genes cut into homogeneous (i.e. fully coding or non coding) pieces of different length. Two parts for frame specific as well as frame independent measures are available.

Table 1 shows the results for frame independent and frame specific exon/intron classification using the SLM approach with rational interpolation. We compared the hexamer measure with SLMs of context length six and the dicodon/diamino acid measure with SLMs trained on dicodons resp. diamino acids, strictly following the benchmark guidelines described in [1]. As one can see, the SLMs consistently outperform the conventional hexamer and dicodon/diamino acid measure without interpolation, and except for frame independent sequences of length 162, our methods give better results than the measures described in [1]. This clearly shows that SLMs are able to estimate the statistics more reliably than conventional measures, although in this case a large amount of data is at hand. The effect of interpolation is more obvious when only sparse data is available, as it is the case with eukaryotic promoter sequences: Here, the hexamers overfit to the sparse training data and perform rather poorly on unseen test sequences, whereas the SLM approach still yields good results [3].

References

[1] J. W. Fickett and C.-S. Tung. Assessment of protein coding measures. *Nuc. Ac. Res.*, 20(24):6441–6450, 1992.

[2] U. Ohler. Hidden Markov Modelle und Polygramme zur DNA–Analyse. Studienarbeit, Universität Erlangen–Nürnberg, 1995.

[3] U. Ohler and M. Reese. Detection of eukaryotic promoter regions using stochastic language models. In *Workshop Molekulare Bioinformatik (Informatik 98)*, Magdeburg, 1998. To appear.

[4] E. G. Schukat-Talamazzini, F. Gallwitz, S. Harbeck, and V. Warnke. Rational Interpolation of Maximum Likelihood Predictors in Stochastic Language Modeling. In *Proc. European Conf. on Speech Communication and Technology*, volume 5, pages 2731–2734, Rhodes, Greece, September 1997.