Motif Recognition in Genomes

Peter Clote clote@informatik.uni-muenchen.de
Sebastian Will wills@informatik.uni-muenchen.de

Institut für Informatik, LFE für Theoretische Informatik,
Ludwig-Maximilians-Universität München, Oettingenstraße 67, D-80538 München, Germany

Keywords: motif recognition, Boltzmann distribution, RNA secondary structure

1 Introduction

A well known and mathematically well established approach to motif detection in DNA is that of *Hidden Markov Models* (HMM), where the model is trained using a set of training sequences. A major property of k-th order Markov models is their locality, i.e. the probability of a state depends only on the last k states. However, especially when 3-dimensional structure plays a role in the recognition of a DNA motif, long range correlations may be as important as local, short range interactions, the latter of which are recognized by HMMs. We present a newly developed and implemented approach to motif recognition, which we call *Window Tuple Potential* (WTP). WTP is motivated by work of Sippl [3] related to protein threading, and is especially suited for the detection of those long range interactions in motifs as well as short range interactions. Further, WTP allows consideration of secondary structure information of RNA for a search for RNA motifs in DNA. Thus, WTP promises to be even more sensitive than HMM approaches.

2 Methods

The WTP approach works essentially by sliding a window $w_i = a_i a_{i+1} \dots a_{i+m}$ over a sequence $a_1 a_2 \dots a_N$ of DNA bases, i.e. $a_i \in \{A, C, G, T\}$, and computing a score E_{w_i} for each window w_i. The best scoring windows will be candidates to contain the motif.

To see how to compute such a score look at the special case of window pair potential. From a set of training sequences, which characterize our motif, we estimate probabilities p_{ij}^{ab} of observing the base a (resp. b) at position i (resp. j). The probability estimation is done by using frequencies such that background frequencies are taken into account. Now, Boltzmann probability is defined by

$$p_{ij}^{ab} = \frac{e^{-E_{ij}^{ab}/T}}{Z^{ab}}, \tag{1}$$

where $Z^{ab} = \sum_{1 \le i < j \le m} e^{-E_{ij}^{ab}/T}$ is the partition function. Pseudo-energy terms E_{ij}^{ab} are computed by taking logarithms of (1), and either setting $Z^{ab} = 1$ or computing relative energy terms, thus yielding a window energy score of

$$E_w = \sum_{1 \le i < j \le m} E_{ij}^{w_i w_j}.$$

This application of Boltzmann probability is similar to the application to amino acid potentials for proteins by Sippl in [3].

RNA secondary structure can be taken into account by incrementally computing the secondary structure S of each window by a dynamic programming approach [2] and computing the score

$$E_w = \sum_{(i,j)\in S} E_{ij}^{w_i w_j},$$

which depends only on those pairs i,j which are base-paired in the RNA secondary structure S.

For secondary structure determination, our implementation uses and modifies parts of the Vienna RNA Package [1]. For the special case of computing the secondary structure iteratively while sliding the window, we developed a modification of Zuker's energy minimization dynamic programming algorithm [2] to reduce the complexity from $O(n^4)$ to $O(n^3)$ by reusing previously computed information.

3 Discussion

We have investigated and compared different variants of the window tuple potential idea and will present empirical evaluation of these variants.

It is planned to further improve WTP for practical applications. We intend to use related methods for the detection of highly secondary structure dependent RNA motifs.

4 Acknowledgments

We would like to thank Prof. Eugene Shakhnovich for helpful discussions on the topic of window tuple potential, especially regarding the incorporation of RNA secondary structure.

References

[1] Hofacker, I.L., Fontana, W., Stadler, P.F., Bonhoeffer, S., Tacker, M., and Schuster, P., Fast folding and comparison of rna secondary structures, *Monatshefte f. Chemie*, 125:167–188, 1994.

[2] Stiegler, P. and Zuker, M., Optimal computer folding of large rna sequences using thermodynamic and auxiliary information, *Nucl. Acid Res.*, 9:133–148, 1981.

[3] Sippl, M., Calculation of conformation ensembles from potentials of mean force, *J. Mol. Biol.*, 213:859–883, 1990.

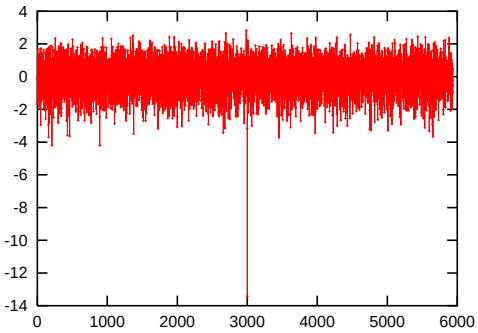

Figure 1: Example plot of window pair potentials E_{w_i}. The positions i are shown on the x-axis and the score is on the y-axis. Here, we have trained using a set of 5S RNA sequences and searched in a artificial DNA sequence, where we have inserted a 5S RNA.