# Optimization of MISCORE-based Motif Identification Systems

Nung Kion Lee and Dianhui Wang
Department of Computer Science and Computer Engineering
La Trobe University, VIC 3086, Australia
Email:dh.wang@latrobe.edu.au

***Abstract*—Identification of motifs in DNA sequences using classification techniques is one of computational approaches to discovering novel binding sites. In the previous work [16], we proposed a simple and effective method for motif detection using a single crisp rule governed by a mismatch-based matrix similarity score (MISCORE). In this paper, we consider the problem of finding suitable motif cut-off value for MISCORE-based motif identification systems using cost-sensitivity metric. We utilize phylogenetic footprinting data to estimate the parameters in the cost function. We also extend the MISCORE to include entropy to weigh each motif model position to minimize the false positive rate. The performance evaluation is done by using artificial and real DNA sequences. The results demonstrate the feasibility and usefulness of our proposed approach for model based cut-off value estimation.**

## I. Introduction

Motif detection is a computational algorithm that searches for novel protein binding sites giving a motif model built from some known sites. It predicts putative binding sites in DNA sequences for further analysis. In [16], we have developed a mismatch-based similarity score called MISCORE based on position frequency matrix (PFM) [14] for motif detection. MISCORE has been shown to perform well in comparison with three popular methods [16]. This paper aims at further extending our work by proposing a cost optimization method to determine suitable cut-off value of motif. We also include a weighted term in the scoring. We demonstrates that under the circumstance where some true or artificial data is available, we can find MISCORE motif cut-off value using a cost matrix that fulfills user's objective.

The remainder of this paper is organized as follows: Section II defines the notations and presents our proposed MISCORE and its extension, WMISCORE. This section will also discuss the method to optimize the motif cut-off value. Section III reports the evaluation results of MISCORE and WMISCORE, using both artificial and real sequences. Concluding remarks are given in the last section.

## II. Method

### A. Motif Model

Firstly, some notations used in this paper are defined. Let $S$ be a motif that consists of a set of binding sites associated with a certain transcription factor. It is assumed that each binding site in $S$ has a fixed length $k$, which can be achieved through multiple alignment tools. A kmer is a subsequence of length $k$ in DNA sequences, i.e., $T_1T_2\cdots T_k$, where $T_j \in \Sigma = \{A, C, G, T\}, j = 1, 2, ..., k$. In this paper, a binary matrix representation is used, which is compatible to the PFM model. The encoded kmer is given by $e(kmer) = [a_{ij}]_{4\times k}$, $a_{ij} = 1$ if $T_j = V_i$, otherwise $a_{ij} = 0$, where $(V_1, V_2, V_3, V_4) = (A, C, G, T)$. For example, the subsequence AGCGTGT can be encoded as:

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \end{bmatrix}$$

Motifs can be expressed as consensus or probabilistic profiles [14]. Profile representations such as Position Frequency Matrix (PFM) assigns a relative frequency to each possible nucleotide at each position in the motif. It is a $4 \times k$ matrix and each column vector represents the position wise observed nucleotides (i.e. A, C, G, T) frequency in a motif.

### B. MISCORE-based Motif Identification Systems

To build a PFM model based classifier for motifs detection, it is fundamental to define a proper similarity function that reflects the closeness concept in the biological sense. Due to functional associations of binding sites, they are evolutionary constrained as compared to background sequences [5]. Hence, a kmer is likely to be a true site if it has limited mismatches to every binding site in a motif. This understanding forms the basis of this work. In practice, a scaled model mismatch score (see (8)) is employed to predict the kmer class. The following proposition gives a way to compute the number of average mismatch between a kmer and a motif model.

**Proposition 1:**[16] Let $R^{4\times k}_{[0,1]}$ and $R^{4\times k}_{\{0,1\}}$ represent the sets of real matrices with size $4 \times k$ and entries taking values in unitary interval $[0, 1]$ and binary values, respectively; $S = \{K_p \in R^{4\times k}, p = 1, 2, ...\}$ be a motif, which is modelled by its PFM. Define a generalized Hamming distance function over $R^{4\times k}_{[0,1]}$ as follows:

$$d(M_1, M_2) = \frac{1}{2}\sum_{j=1}^{k}\sum_{i=1}^{4} w_j \left| m^{(1)}_{i,j} - m^{(2)}_{i,j} \right|, \qquad (1)$$

where $0 \le w_j \le 1, j = 1, 2, ..., k$.

978-1-4244-2902-8/09/$25.00 ©2009 IEEE