Characterising DNA/RNA Signals with Crisp Hypermotifs: A Case Study on Core Promoters

Carey Pridgeon[1] and David Corne[2]

[1] Department of Computer Science, University of Exeter, Exeter EX4 4QF, UK
[2] School of MACS, Heriot-Watt University, Edinburgh EH14 8AS, UK
carey.pridgeon@gmail.com, d.w.corne@macs.hw.ac.uk

Abstract. A common way to characterise important and conserved signals in nucleotide sequences, such as transcription factor binding sites, is via the use of so-called *consensus* sequences or consensus patterns. A well-known example is the so-called "TATA-box" commonly found in eukaryotic core promoters. Such patterns are valuable in that they offer an insight into basic molecular biology processes, and can support reasoning regarding the understanding, design and control of these processes. However it is rare for such patterns to be accurate; instead they represent a very approximate characterisation of the signal under study. At the opposite extreme, we may instead characterise such a signal via a neural network, or a high-order Markov model, and so on. These have better sensitivity and specificity, but are unreadable, and consequently unhelpful for conveying an understanding of the underlying molecular biology processes that could support insight or reasoning. We describe a simple pattern language, called crisp hypermotifs (CHMs), that leads to highly readable patterns that can support understanding and reasoning, yet achieve greater sensitivity and specificity than the commonly used approaches to crisply characterise a signal. We use evolutionary computation to discover high-performance CHMs from data, and we argue that CHMs be used in place of classical consensus motifs, and justify that by presenting examples derived from a large dataset of mammalian core promoters. We provide CHM alternatives to the well-known core promoter TATA-box and Initiator patterns that have better sensitivity and specificity than their classical counterparts.

1 Introduction

It is common in molecular biology to seek models that discriminate between different groups of nucleotide sequences. For example, we need to be able to discriminate between intron/exon splice sites and control sequences, between core promoters and control sequences, between different classes of microRNA, and so forth. The model used for discrimination can take many forms, but at a high level we can categorise them as occupying a continuum from 'readable/approximate' to 'opaque/accurate'. At the 'readable/approximate' end, we have straightforward sequences or simple patterns that characterise the positive (i.e. not the control) sequences. For example, such a model might be a simple consensus sequence, from the alphabet {A,C,G,T}, which is commonly observed in the positive sequences and rarely observed in the controls. Or,

E. Marchiori, J.H. Moore, and J.C. Rajapakse (Eds.): EvoBIO 2007, LNCS 4447, pp. 227 – 235, 2007.

we may have a pattern from the IUPAC[1] alphabet, such as YYCARR, which matches several of the positive cases but fewer or none of the controls. Meanwhile, at the 'opaque/accurate' end of this continuum, we may have models such as hidden Markov models (e.g. Henderson et al, 97), higher-order Markov models,(e.g. Salzberg et al, 98) neural networks (e.g. Reese, 01), support vector machines (e.g. Zien et al, 00), and so on. These tend to provide more accurate classification than the simple patterns, and are consequently used for genome annotation and similar tasks, however they are opaque to analysis – i.e. it is very difficult or impossible to glean knowledge and insight concerning the precise sequences and patterns of nucleotides that form the signal of interest.

Conceivably, there are DNA and RNA tasks for which consensus or simple patterns are sufficient, since the signal of interest is highly conserved. But for the majority of such tasks, simple patterns made from the IUPAC alphabet (see Table 1) poorly capture the variability of nucleotide composition in the signal, and score quite low in sensitivity and specificity.

Table 1. The IUPAC alphabet for nucleotide subsets, used in motifs

Nucleotide subset	IUPAC Symbol
A	A
C	C
G	G
T	T
A, C	M
A, G	R
A, T	W
C, G	S
C, T	Y
G, T	K
A, C, G	V
A, C, T	H
A, G, T	D
C, G, T	B
A, C, G, T	N

An interesting challenge is to find ways that can provide accuracy competitive with 'opaque' techniques, yet which have enough readability to support fruitful insight and conjecture concerning underlying molecular biology mechanisms. The idea of using a *hypermotif* (Pridgeon & Corne, 05) was motivated by this challenge. In the latter paper, we focused on the use of *weighted* hypermotifs for use in conjunction with neural networks, to provide discrimination at the 'opaque' end of the model spectrum. In this paper we concentrate only on beginning to examine 'crisp' (i.e. unweighted)

[1] IUPAC: International Union of Pure and Applied Chemistry – a source of standards, include common abbreviations for sets of nucleotides.