Identifying Regulatory Sites Using Neighborhood Species

Claudia Angelini[1], Luisa Cutillo[1], Italia De Feis[1], Richard van der Wath[2], and Pietro Lio'[2,⋆]

[1] Istituto per le Applicazioni del Calcolo "Mauro Picone" CNR, Napoly Italy
c.angelini@iac.cnr.it, cutillo@na.iac.cnr.it, i.defeis@iac.cnr.it
[2] Computer Laboratory, University of Cambridge, Cambridge UK
rcv23@cam.ac.uk, pl219@cam.ac.uk

Abstract. The annotation of transcription binding sites in new sequenced genomes is an important and challenging problem. We have previously shown how a regression model that linearly relates gene expression levels to the matching scores of nucleotide patterns allows us to identify DNA-binding sites from a collection of co-regulated genes and their nearby non-coding DNA sequences. Our methodology uses Bayesian models and stochastic search techniques to select transcription factor binding site candidates. Here we show that this methodology allows us to identify binding sites in nearby species. We present examples of annotation crossing from *Schizosaccharomyces pombe* to *Schizosaccharomyces japonicus*. We found that the eng1 motif is also regulating a set of 9 genes in *S. japonicus*. Our framework may have an effective interest in conveying information in the annotation process of a new species. Finally we discuss a number of statistical and biological issues related to the identification of binding sites through covariates of genes expression and sequences.

1 Introduction

The identification of the repertoire of regulatory elements in a genome is one of the major challenges in modern biology. Gene transcription is determined by the interaction between transcription factors and their binding sites, called motifs or cis-regulatory elements. In eukaryotes the regulation of gene expression is highly complex and often occurs through the coordinated action of multiple transcription factors. This combinatorial regulation has several advantages; it controls gene expression in response to a variety of signals from the environment and allows the use of a limited number of transcription factors to create many combinations of regulators. Identification of the regulatory elements is necessary for understanding mechanisms of cellular processes. In eukaryotes these sites comprise short DNA stretches often found within non-coding upstream regions. DNA microarrays provide a simple and natural vehicle for exploring the regulation of thousands of genes and their interactions. Genes with similar expression

⋆ Corresponding author.

E. Marchiori, J.H. Moore, and J.C. Rajapakse (Eds.): EvoBIO 2007, LNCS 4447, pp. 1–10, 2007.

profiles are likely to have similar regulatory mechanisms. A close inspection of their promoter sequences may therefore reveal nucleotide patterns that are relevant to their regulation.

In order to identify regulative sites several authors have used the following strategy: 1) candidate motifs can be obtained from the upstream regions of the most induced or most repressed genes; 2) a score can be assigned to reflect the matching of each motif to a particular upstream sequence; 3) regression analysis and variable selection methods can be used to detect sets of motifs acting together to affect the expression of genes [4,5,8].

Most of the current focus on microarray analysis is on integrating results from repeated experiments using the same species or using different species. This paper is extending this focus to transcription factor binding site identification. Following [8], we propose the use of Bayesian variable selection models to use the gene expression of an organism to find transcription binding sites of a closely related species or of a different strain. Variable selection methods use a latent binary vector to index all possible sets of variables (patterns). Stochastic search techniques are then used to explore the high-dimensional variable space and identify sets that best predict the response variable (expression). The method provides joint posterior probabilities of sets of patterns, as well as marginal posterior probabilities for the inclusion of single nucleotide patterns. We have chosen to exemplify our methodology using *S. japonicus* and *S. pombe* genomes and microarray data from cell cycle-regulated gene experiments [6].

Similar to a better known Schizosaccharomyces *S. pombe*, which has been a major model organism for cell cycle and cell biology research for thirty years, *S. japonicus* is a simple, unicellular yeast. Unlike the cousin, it readily adopts a invasive, hyphal growth form. Such growth is an important virulence trait in pathogenic fungi, making *S. japonicus* a potentially important model for fungal disease. The comparison of the *S. pombe* genome, which was sequenced several years ago, with those of its close relatives will greatly improve our understanding of the genomes and the proteins they encode. In addition, the three fission yeasts form an early-branching clade among the Ascomycete (ascus-forming) fungi, which includes yeast, hyphal fungi, and truffles [2]. Although a great deal of molecular information is available from *S. pombe*, a model eukaryote, very little is known about the *S. japonicus* cell-cycle regulative network.

Here we show that our methodology allows us to identify binding sites in *S. japonicus* using *S. pombe* gene expression data. As an example of annotation crossing from *S. pombe* to *S. japonicus* we focus on the Eng1 cluster, a set of very strongly cell cycle-regulated genes, which in *S. pombe* contains nine genes, involved in cell separation [6]. The genes are adg1 and adg2 (cell surface glycoproteins), adg3 (glucosidase), agn1 and eng1 (glycosyl hydrolases), cfh4 (chitin synthase regulatory factor), mid2 (an anillin needed for cell division and septin organization), ace2 (a cell cycle transcription factor), and SPCC306.11, a sequence orphan of unknown function. Motif searches showed that each gene of the cluster has at least one binding site for the Ace2 transcription factor (consensus CCAGCC). The Eng1 cluster has a recognizably similar functional cluster in