Chapter 5
Detection and Employment of Biological Sequence Motifs

Marjan Trutschl
Louisiana State University – Shreveport, USA & Louisiana State University Health – Shreveport, USA

Rona S. Scott
Louisiana State University Health – Shreveport, USA

Phillip C. S. R. Kilgore
Louisiana State University – Shreveport, USA

Christine E. Birdwell
Louisiana State University Health – Shreveport, USA

Urška Cvek
Louisiana State University – Shreveport, USA & Louisiana State University Health – Shreveport, USA

ABSTRACT

Biological sequence motifs are short nucleotide or amino acid sequences that are biologically significant and are attractive to scientists because they are usually highly conserved and result in structural and regulatory implications. In this chapter, the authors show practical applications of these data, followed by a review of the algorithms, techniques, and tools. They address the nature of motifs and elucidate on several methods for de novo motif discovery, covering the algorithms based on Gibbs sampling, expectation maximization, Bayesian inference, covariance models, and discriminative learning. The authors present the tools and their requirements to weigh their individual benefits and challenges. Since interpretation of a large set of results can pose significant challenges, they discuss several methods for handling data that span from visualization to integration into pipelines and curated databases. Additionally, the authors show practical applications of these data with examples.

INTRODUCTION

A topic of increasing interest to geneticists and biochemists is the detection and utilization of biologically-significant short sequence sections called *motifs*. Motifs are short nucleotide or amino acid sequences that are intriguing because they are usually highly conserved and have structural or regulatory implications for many biological processes. Motifs are constituents of cellular macromolecules such as nucleic acids, proteins, lipids and carbohydrates. Due to the vast number of motifs, our description of motifs primarily focuses on motifs involved in gene expression and found on deoxyribonucleic acid (DNA), ribonucleic acids (RNA) and proteins.

DOI: 10.4018/978-1-4666-6611-5.ch005

Motifs found in DNA are generally protein binding sites that regulate transcription, replication, and convey spatial organization to the human genome. RNA motifs can serve as regulatory elements for RNA processing and stability of RNA transcripts. Protein motifs are often involved in protein-protein interactions and binding to DNA and RNA for regulation of transcription, translation and DNA replication.

Reliable motif identification can lend itself to streamlining the process of discovering protein function as well as factors involved in gene regulation. To exemplify, a basic research scientist may use motifs found in the promoter region of a gene of interest to predict the context in which the cell turns on the gene. Life scientists may use the identification of RNA motifs to predict alternative splicing or regulatory RNA transcripts. Motif identification is not just useful for basic science researchers; it can also be used by clinicians to pinpoint mutations that are leading to the observed phenotype. Motifs can also be used by industry to predict drug target sites on proteins or nucleic acids. Scientists are seeking automated and reliable methods for discovering motifs that would help them guide their discoveries and generate new hypotheses.

BACKGROUND

The concept of motifs and their relationship to regulation of the cellular environment can be traced back to the late 1950s. Although regulatory elements had been shown to exist in DNA as early as 1951 (McClintock, 1951), it was the work of Jacques Monod and Francois Jacob regarding the regulation of lactose metabolism in *Escherichia coli* that lead to the first generalized theory concerning regulatory elements. Via the *lac* repressor, a protein which moderates the translation of the proteins used in lactose synthesis, Jacob et al. were able to develop a framework accounting for transcriptional regulation (Jacob, Perrin, Sanchez, and Monod, 1960). This seminal work only considered repressor elements and was primitive in comparison to modern views regarding transcriptional regulation; however, it is notable in that it presents the concept of an *operator*, a segment of DNA to which regulatory elements may bind.

At about the same time, a number of motifs were being identified within gene promoter regions. Promoters are DNA sequences which regulate the initiation of transcription of nearby genes. The 1970s saw the discovery of two conserved motifs that recruit the general transcriptional factors and RNA polymerase (Hurwitz, 1960; Stephens, 1960) to promoters: the TATA box "TATAA" (Rifton, Goldberg, Karp, & Hogness, 1978) and the Pribnow box "TATAAT" (Pribnow, 1975). The former is called the Goldberg-Hogness Box in eukaryotes, and the latter is known as the -10 sequence in bacterial promoters. Additional promoter motifs have since been identified and underscore the regulatory complexity between prokaryotes and eukaryotes. Bacterial promoters usually have three unique motifs while eukaryotic promoters can have up to seven (Clancy, 2008).

Over the years, annotation of transcription binding elements has elucidated consensus sequences that regulate binding of these factors to DNA. For instance, the constitutively expressed Sp1, a zinc finger transcription factor was found to bind the GC box motif: 5'(G/T)GGGCGG(G/A)(G/A)(C/T) 3'. The general consensus sequence for NF-kB, an inducible transcription factor involved in activation of many genes under different stimuli, is the kB site: 5' GGGR(C/A/T)TYYCC 3'. These are only 2 examples of the many proteins recruited to promoter motif sequences and involved in activation or repression of transcription (Levine & Tjian, 2003). The transcription factor binding sequences (TFBS) used by many transcription factors have been identified and are available on open sourced sites, such as the JASPAR database and the Encyclopedia of DNA Elements (ENCODE) project's factorbook, that can be readily accessed