Chapter 18
Analysis and Prediction of DNA–Recognition by Zinc Finger Proteins:
Applications in Genome Modification

Anita Sarkar
Indian Institute of Technology, India

Ankita Punetha
Indian Institute of Technology, India

Sonu Kumar
Indian Institute of Technology, India

Abhinav Grover
Indian Institute of Technology, India

Durai Sundar
Indian Institute of Technology, India

ABSTRACT

Zinc fingers are the most abundant class of DNA-binding proteins encoded in the eukaryotic genomes. Custom-designed zinc finger proteins attached to various DNA-modifying domains can be used to achieve highly specific genome modification, which has tremendous applications in molecular therapeutics. Analysis of sequence and structure of the zinc finger proteins provides clues for understanding protein-DNA interactions and aid in custom-design of zinc finger proteins with tailor-made specificity. Computational methods for prediction of recognition helices for C2H2 zinc fingers that bind to specific target DNA sites could provide valuable insights for researchers interested in designing specific zinc finger proteins for biological and biomedical applications. In this chapter, we describe the zinc finger protein-DNA interaction patterns, challenges in engineering the recognition-specificity of zinc finger proteins, the computational methods of prediction of proteins that recognize specific target DNA sequence and their applications in molecular therapeutics.

DOI: 10.4018/978-1-4666-3604-0.ch018

INTRODUCTION

The Human Genome Project (HGP) was a landmark achievement for science in more ways than one. The immense data generated by this effort, and thereafter, has given us a holistic view of our genetic make-up. The pool of refined data that we obtain as a result of this endeavor will be the reference source for all future biomedical research. Although the human genome is almost 99.9% similar in everybody, individuality stems from the millions of single-base DNA mutations that vary from person to person. These are also the reason for every individual's varying susceptibility towards cardiovascular disease, diabetes, arthritis, cancers etc. and a determinant of how that person reacts to environmental factors. Deft mining of the biological data generated by the HGP to obtain useful information is essential for researchers who are working in the areas of Genomics and Proteomics. As a result, a new field of science called Computational Genomics and Bioinformatics has emerged, in which the focus is on developing efficient algorithms and computational tools to obtain valuable information which is of interest to the user. The design of such tools and databases is such that the user is readily able to extract useful information from the ever-rising pool of biological sequence and structure data.

The human genome has gene-rich (G≡C rich) and gene-poor (A=T rich) centers. The G≡C rich regions frequently appear adjacent to the gene-rich areas, acting as barriers between genes and junk DNA and aiding gene regulation. This is thus, taken as an indicative in certain computational tools, like those designed for gene identification to detect the presence of genes.

Gene expression in all organisms is controlled by DNA-protein complexes. Every person carries, approximately, half a dozen defective genes, but remains ignorant about their susceptibility till a close relation suffers from a genetic disease. The easy availability of complete genomes has provided a new platform for developing new therapeutic strategies by modulating the transcription of the gene(s) of interest.

Protein-nucleic acid interaction remains the central theme in the area of molecular recognition and transcription factors have been given a priority for many reasons for research in this area. Transcription factors are proteins that regulate gene expression by binding to the promoter elements lying upstream of genes and either facilitate or inhibit transcription. Transcription factors are composed of two essential functional regions: a DNA-binding domain and an activator domain. The DNA-binding domain consists of amino acids that recognize specific DNA bases at the beginning of transcription. Transcription factors are typically classified according to the structure of its DNA-binding domain, which are of one of the following types: zinc fingers, helix-turn-helix, leucine zipper, helix-loop-helix, and high mobility groups (Sauer, 1990; Schwabe & Rhodes, 1991). The class of zinc finger proteins (ZFPs) is especially fascinating due to its ability to recognize DNA, RNA as well as other proteins. Though there are pointers to predict DNA-protein interactions, changing binding specificities of existing transcription factors and probably even devising a rational way to design proteins that would recognize and bind to a particular target DNA sequence, there is no simple formula (or code) for DNA recognition by transcription factors. Advancements in design and engineering of DNA-binding proteins have led to the use of zinc fingers in gene targeting for therapeutic purposes. In this chapter, we discuss the progress towards computational prediction of DNA recognition of zinc finger proteins, the different approaches developed to design them and its potential applications in biomedicine in areas such as cancer, monogenic diseases and detecting infectious disease-causing pathogens. Several important leads have been obtained in such applications for the treatment of Severe Combined Immunodeficiency Disease (SCID) and infectious diseases such as hepatitis B and HIV, Alzheimer's disease, peripheral artery disease, diabetic neuropathy, heart disease etc.