A Pattern Matching Algorithm for Codon Optimization and CpG Motif-Engineering in DNA Expression Vectors

Ravi Vijaya Satya and Amar Mukherjee
School of Engineering and Computer Science
University of Central Florida
Orlando, FL 32816
rvijaya, amar@cs.ucf.edu

Udaykumar Ranga
Jawaharlal Nehru Center for Advanced Scientific Research
Jakkur, Bangalore, India
udaykumar@jncasr.ac.in

Abstract

Codon optimization enhances the efficiency of DNA expression vectors used in DNA vaccination and gene therapy by increasing protein expression. Additionally, certain nucleotide motifs have experimentally been shown to be immuno-stimulatory while certain others immuno-suppressive. In this paper, we present algorithms to locate a given set of immuno-modulatory motifs in the DNA expression vectors corresponding to a given amino acid sequence and maximize or minimize the number and the context of the immuno-modulatory motifs in the DNA expression vectors. The main contribution is to use multiple pattern matching algorithms to synthesize a DNA sequence for a given amino acid sequence and a graph theoretic approach for finding the longest weighted path in a directed graph that will maximize or minimize certain motifs. This is achieved using $O(n^2)$ time, where n is the length of the amino acid sequence. Based on this, we develop a software tool.

Key Words: Codon optimization, immuno-modulatory motifs, multiple pattern matching, longest weighted path.

1. Introduction

DNA vaccines have revolutionized the field of vaccine technology by demonstrating the ability to induce humoral and cellular immune responses in experimental animals and humans [9]. Immunization of animals with plasmid DNA encoding a protein antigen was an accidental observation that eventually led the way to a novel strategy of immunization. DNA vaccines, also known as "naked DNA" or "nucleic acid" vaccines, encode the antigens of pathogenic organisms including viruses, bacteria, fungi and parasites [40]. The protein antigen is processed within the cell and presented by the MHC-I and –II pathways thereby eliciting specific immune responses essential for controlling pathogenic infections [2]. Although DNA vaccines have been successful in generating strong immune responses in smaller animal model such as mouse, they have not been as efficient in larger species such as primates and humans [23, 3, 25, 4]. Stimulating both the arms of the immune system is often desirable for efficient control of infectious diseases especially in the larger animals. In the case of recombinant protein vaccines, immune-enhancers technically known as adjuvants, such as Freund's adjuvants and Alum, are in use to enhance antigen specific immune responses. However, no such adjuvants are available for use in the context of DNA vaccines. The lack of suitable adjuvants for DNA vaccines is one important reason for the poor performance of the DNA vaccines in larger animals.

Nucleotide sequence encoding a foreign protein is directly placed under the control of a mammalian promoter to construct a DNA vaccine. Several amino acids are encoded by more than one triplet codon and different organisms have variable requirement for codon preference [13]. Cloning of a wild type gene from a parasite into a DNA vaccine often leads to insufficient levels of protein synthesis in the host cell as a result of codon bias between the species. Successful immunization with DNA vaccines requires high expression of cloned genes to synthesize large quantities of the foreign protein. For instance, the overall genetic content of Human Immunodeficiency Virus-1 is AT-rich, while that of the human beings is CG-rich. The codon frequency of the pathogenic DNA embedded into the mammalian expression vector may not be optimal for adequate protein expression in the host resulting in low level protein expression. A potential solution for the codon bias is to optimize the codon sequences of a gene to suit the requirements of the host without altering the original amino acid sequence of the protein [41, 16]. This approach has been successful in eliciting strong immune responses in several species of experimental animals [8,

28]. While immunization with synthetic genes, codon-optimized for mammalian expression stimulated strong immune response, immunization performed in parallel with wild type genes generated low or moderate levels of immune response [34, 36].

In addition to codon optimization of the synthetic genes, a range of molecular approaches is being evaluated to up-regulate immune responses generated by DNA vaccines. Co-expression of cytokine genes [20], co-stimulatory receptors [12, 35] or other immunemodulators [33], synthetic assembly of T-helper or CTL epitopes [6] and formulation with a variety of chemical adjuvants [11, 24, 33, 37, 39] have been some of the approaches reported. However, most of these approaches may not be suitable for human application due to toxic manifestations of the adjuvants. An ideal agent used as an adjuvant for DNA vaccine must enhance the immunogenicity without apparent cytotoxicity to the host. Engineering CpG islands into DNA vaccines has been one promising approach that showed enhanced immune responses [19].

The well-established immune-enhancing property of bacterial DNA has been mapped to sequence motifs consisting of un-methylated CpG dinucleotides flanked by base pairs in a specific context [18]. Two important differences between the bacterial and mammalian DNA enable the mammalian innate immune system to recognize the former as a foreign component. CpG motifs in bacteria are found at the expected 1:16 frequency; however, their frequency in mammals is 4 times less than expected. Bacterial CpG are non-methylated while those of mammals are mostly methylated. The mammalian immune system takes advantage of these two chemical differences between the bacterial and mammalian DNA to identify a bacterial infection and wage strong and rapid anti-bacterial immune responses [29].

CpG-mediated activation of the mammalian innate immune system has been extensively exploited in the vaccination technology. Co-injection of CpG containing oligonucleotides or empty vectors with protein antigens elicited potent immune response to the antigen. Methylation of the CpG motifs, on the other hand, abrogated immune response to the antigens suggesting that the CpG motifs possess adjuvant properties only when unmethylated or hypomethylated [[5, 17]. Since the DNA expression vectors used in genetic immunizations are usually grown in bacteria, several CpG motifs on the plasmids are not methylated possibly activating the innate component of the mammalian immune system. Presence of hypo-methylated CpG motifs is essential for induction of immune responses to the antigens encoded by the vectors.

While certain CpG motifs are immuno-stimulatory (CpG-S), enhancing immune responses when the host is vaccinated, others CpG motifs in a different nucleotide context are immuno-inhibitory or -neutralizing (CpG-N) abrogating antigen-specific immune response [18]. Engineering the nature and the frequency of the CpG motifs may be critical for the design of DNA expression vectors. DNA expression vectors are primarily used for two different applications, expression of a foreign gene in a host for vaccination or for correcting a genetic defect of the host [26].

Although the basic design of these two types of vectors is identical in several respects, their requirement for the presence and nature of CpG motifs is diagonally opposite. While genetic vaccines require CpG-S motifs for efficient stimulation of the host immune system, such a strong response must be avoided for long-term survival of the DNA expression vector intended for gene therapy. An ideal strategy for designing DNA vaccines must recruit as many CpG-S motifs as possible, concomitantly eliminating as many CpG-N motifs as possible without altering the original protein sequence of the genes. In contrast, design of a DNA expression vector intended for gene therapy must eliminate as many CpG-S motifs as possible and recruit as many CpG-N motifs as possible for the best result. Thus, depending on the application, some CpG motifs may be **desirable**, while certain others may be **undesirable**.

The software tool we report here is designed to help the researcher to engineer the composition of a gene with respect to codon usage and CpG motifs without modifying the original amino acid sequence of the protein. Codon optimization is advantageous for protein expression in a heterologous expression system such as *E.coli*, *Picchia*, *Saccharomyces*, *Baculo virus*, mammalian cells etc. The software could also engineer the content and nature of the CpG motifs recruited into a DNA expression vector in addition to optimizing the codon usage and resolve potential conflicts arising between these two requirements and find an optimal nucleotide sequence.

2. Prior Work and Summary of Contributions

Objective of this algorithm is to (1) identify the best triplet codon for codon optimization and (2) engineer the nature and content of the CpG motifs of a gene expressed from a genetic vector. Input for the software is the amino acid sequence of a gene of interest or the nucleotide sequence that needs genetic modification. The input amino acid sequence is reverse-translated to generate nucleotide sequence that is optimized for the codon and CpG content. During the process of reverse-translation, there is a one-to-many relationship between the amino acid sequence and the DNA sequences. A given amino acid sequence could correspond to an exponential number of DNA sequences with respect to the length of the given amino acid sequence. Our task here is to choose the particular sequence that has the desirable properties (maximum number of the desirable motifs, and the minimum number of the undesirable motifs) from this large number of possible DNA sequences. A brute force search on all the possible DNA sequences will take exponential time. The motifs (or the patterns) that need optimization are small DNA sequences, generally not exceeding eight nucleotides in length. If different motifs have to be optimized simultaneously, it is possible to have multiple motif occurrences (both desirable and undesirable) sharing the same position in the amino acid sequence. In such cases it may not be possible to have all the motif occurrences in the same DNA sequence. This problem will be explained in detail in Section 4.

McInerney[27] presented a program to perform the codon usage analysis on a sequence or a database of sequences. Other work in this area of bioinformatics is mostly on finding CpG islands within the genes and transcriptional elements that regulate gene expression. Ponger et al [30] developed a software package to locate CpG islands associated with the transcription start sites of the genes. Lin et al [22] presented software for locating non-overlapping maximum average segments in a given sequence. These publications analyzed DNA sequences and searched for nucleotide motifs with high CG content.

Our emphasis in this paper is to synthesize a DNA sequence that codes for a functional protein, with certain characteristics and optimization parameters. The main contribution is to use multiple pattern-matching algorithms to synthesize a DNA sequence for a given amino acid sequence and a graph theoretic approach for finding the longest weighted path in a directed graph that will maximize or minimize certain motifs as well as guarantee certain fitness factors of codon frequency usage for a particular species. This is achieved using $O(n^2)$ time and storage resources compared to the brute force algorithm that might take exponential amount of resources, where n is the length of the amino acid sequence. The software tools developed for the purpose will find applications in the rapid development of vaccines and gene therapy.

In Section 3, we introduce the basic terminology. In Section 4, we give a precise combinatorial formulation of the problem in terms of pattern matching operations and present the main algorithm. Section 5 gives the complexity analysis. Section 6 gives a description of the software and in Section 7 we present the results.

3. Terms and definitions

The alphabet of a **DNA sequence** is denoted as $\Sigma_N = \{A,C,G,T\}$[1] where A,C,G,T are the four DNA molecules adenine, cytosine, guanine, and thymine, respectively. The alphabet for the **amino acid sequence** is $\Sigma_A = \{A,R,N,D,C,Q,E,G,H,I,L,K,M,F,P,S,T,W,Y,V,stop\}$, each letter indicating one of the 20 amino acids, and stop indicating the stop codon. For example, 'A' indicates Alanine, 'R' indicates Arginine, 'N' indicates Asparagine, etc. A **codon** c is a triplet of the DNA symbols in Σ_N. As the size of DNA alphabet is 4, there are 64 possible codons. In the context of amino acid translation from mRNA, each element of Σ_A gets mapped to a maximum of 6 codons[2]. This mapping is also referred to as the **genetic code**. Symbolically, the mapping of an amino acid $\sigma \in \Sigma_A$ can be denoted as a set $C(\sigma)$ of n_σ codons, where $1 \le n_\sigma \le 6$. The inverse-mapping associates an amino acid σ with a codon c, denoted by $C^{-1}(c) = \sigma$. The frequency of occurrence of members of $C(\sigma)$ is different for each species. This relative difference in the frequency of occurrence for each codon is called **codon bias**. The **codon usage table** for a species gives the codon bias information for that species[3]. For an amino acid σ, the codon bias (or the frequency information) is given by the set of fractions $F(\sigma)$, corresponding to $C(\sigma)$. For each $f_i \in F(\sigma)$, $1 \le i \le n_\sigma$, $0 \le f_i \le 1$, and $\sum_{i=1}^{n_\sigma} f_i = 1$. When designing a DNA vaccine for a species, it is often desirable to have those codons that occur more frequently

[1] The protein synthesis takes place from mRNA. The actual RNA alphabet is {A,C,G,U}, in which thymine is replaced by uracil. For simplicity of notation, we use T instead of U in this paper.

[2] See http://molbio.info.nih.gov/molbio/gcode.html

[3] See http://www.kazusa.or.jp/codon/ for codon usage tables for different species.