Reversible Data Hiding for DNA Sequences and Its Applications

Qi Tang, School of Information Science, University of Science and Technology of China, Hefei, China
Guoli Ma, School of Information Science, University of Science and Technology of China, Hefei, China
Weiming Zhang, School of Information Science, University of Science and Technology of China, Hefei, China
Nenghai Yu, School of Information Science, University of Science and Technology of China, Hefei, China

ABSTRACT

As the blueprint of vital activities of most living things on earth, DNA has important status and must be protected perfectly. And in current DNA databases, each sequence is stored with several notes that help to describe that sequence. However, these notes have no contribution to the protection of sequences. In this paper, the authors propose a reversible data hiding method for DNA sequences, which could be used either to embed sequence-related annotations, or to detect and restore tampers. When embedding sequence annotations, the methods works in low embedding rate mode. Only several bits of annotations are embedded. When used for tamper detection and tamper restoration, all possible embedding positions are utilized to assure the maximum restoration capacity.

Keywords: Degenerate Bases, Embedding Positions, DNA Sequences, Reversible Data Hiding, Sequence Annotations

1. INTRODUCTION

DNA, as we all know, is the genetic codes for human beings and other plants and animals. A DNA sequence is a string of 4 nitrogenous bases: adenine (A), thymine (T), guanine (G), and cytosine (C). Most of current discovered DNA sequences are stored in GenBank of NCBI and other DNA databases. Normally, a sequence file in GenBack consists of two parts: the original sequence and the annotation. The annotation includes several labels such as LOCUS, DEFINITION, ACCESSION, SOURCE and FEATURES, which help to describe the original sequence. These annotations usually occupy a lot of the storage space of the database. Besides the storage problem, once the sequence is tampered by evil attackers, researches based on those tampered sequences will result in disastrous consequences. Thus, a method that

DOI: 10.4018/ijdcf.2014100101
could either solve the storage problem or detect and restore tampers on DNA sequence is of much help.

Reversible data hiding is a widely used technique in digital media field for annotation, watermarking and authentication. As its name indicates, reversible data hiding losslessly hides data into different media. When this data is extracted, the original media are perfectly retrieved. With different types of data embedded, reversible data hiding can be applied to different applications. For example, embedding media-related annotations could make such media self-describable. This is extensively used in medical image processing. In addition, if the content reference information is embedded, we say that the medium is watermarked. In general, there are three types of watermark: fragile watermark, semi-fragile watermark, and robust watermark. Fragile watermark fails to be detectable so it is widely used for tamper detection. Semi-fragile watermark is designed to resist benign transformations in order to detect malignant transformation. A robust watermark can tolerate a designated transformation and therefore, copy protection applications always choose this kind of watermark.

Introducing the traditionally used reversible data hiding techniques to DNA sequences will help to solve the storage problem and to authenticate the sequences. Some scholars have proposed their data hiding and watermarking methods. Shimanovsky et al. (Shimanovsky, 2003) firstly proposed a method in which utilizing the redundancy of codon to amino acid mapping. There are $4^3 = 64$ different codon combinations with three Nucleotides. But all possible amino acids and signals are 22. So a codon in a sequence can be substituted by its peer codons that will be translated into the same amino acid as the original one does. Hence, several bits can be embedded into each codon. In (Chang, 2007), Chang et al. provide two methods. One of them is based on lossless compressing techniques. They losslessly compressed the DNA sequence and then append the secret data to the end of the compressing result. Another one adopted the difference expansion way to hide data, which is widely used in digital image data hiding. Another three methods: the insertion method, the complementary pair method and the substitution method were proposed by Shiu in (Shiu, 2010). Unfortunately, however, in order to restore the original content, all three methods have to make sure that a reference DNA sequence is transmitted to the client. With this premise, all three methods cannot be denoted as reversible data hiding methods. After that, Guo et al. (Guo, 2012) improved the substitution method. But the reference DNA sequence is still needed.

In this paper, we propose a new reversible data hiding method on DNA sequences. In this method, we utilized degenerate base symbols as labels to mark DNA sequences. Hence no reference DNA sequences are needed. Based on this method, two applications are proposed. One annotates DNA sequences by hiding GenBank sequence annotations to make sequences self-describable. The other is a fragile watermark scheme, in which the watermark to be embedded can be used to detect tampered area(s) in DNA sequences. Even further, if tampered area(s) takes up only little portion of the entire sequence, the original content can be exactly retrieved.

The rest of the paper is structured as follows. Section II describes the proposed reversible data hiding method. The fragile watermarking scheme is explained in Section III. Our experimental results are given in Section IV.

2. PROPOSED REVERSIBLE DATA HIDING METHOD

In this section, we will discuss our reversible data hiding method in detail. The method is called Repeated Segment Substitution (RSS) method. Before we dive into it, we will first explain the degenerate base symbols that we used to mark the DNA sequence. “Degenerate base symbols in biochemistry are an IUPAC (IUPAC) representation for a position on a DNA sequence that can have multiple possible alternatives”(Degenerate). When doing gene sequencing, not all bases in the sequence can
Related Content

**Synthesis Over Analysis: Towards an Ontology for Volume Crime Simulation**
[www.igi-global.com/chapter/synthesis-over-analysis/5263?camid=4v1a](www.igi-global.com/chapter/synthesis-over-analysis/5263?camid=4v1a)

**Human Factors in Information Security and Privacy**
[www.igi-global.com/chapter/human-factors-information-security-privacy/60995?camid=4v1a](www.igi-global.com/chapter/human-factors-information-security-privacy/60995?camid=4v1a)

**Ruler Detection for Autoscaling Forensic Images**
[www.igi-global.com/article/ruler-detection-for-autoscaling-forensic-images/110394?camid=4v1a](www.igi-global.com/article/ruler-detection-for-autoscaling-forensic-images/110394?camid=4v1a)
Dynamic Provable Data Possession of Multiple Copies in Cloud Storage Based on Full-Node of AVL Tree