DNA-based Cryptography

Ashish Gehani, Thomas LaBean, and John Reif

Department of Computer Science, Duke University
Box 90129, Durham, NC 27708-0129, USA
{geha,thl,reif}@cs.duke.edu

Abstract. Recent research has considered DNA as a medium for ultra-scale computation and for ultra-compact information storage. One potential key application is DNA-based, molecular cryptography systems. We present some procedures for DNA-based cryptography based on one-time-pads that are in principle unbreakable. Practical applications of cryptographic systems based on one-time-pads are limited in conventional electronic media by the size of the one-time-pad; however DNA provides a much more compact storage medium, and an extremely small amount of DNA suffices even for huge one-time-pads. We detail procedures for two DNA one-time-pad encryption schemes: (i) a substitution method using libraries of distinct pads, each of which defines a specific, randomly generated, pair-wise mapping; and (ii) an XOR scheme utilizing molecular computation and indexed, random key strings. These methods can be applied either for the encryption of natural DNA or for artificial DNA encoding binary data. In the latter case, we also present a novel use of chip-based DNA micro-array technology for 2D data input and output. Finally, we examine a class of DNA steganography systems, which secretly tag the input DNA and then hide it within collections of other DNA. We consider potential limitations of these steganographic techniques, proving that in theory the message hidden with such a method can be recovered by an adversary. We also discuss various modified DNA steganography methods which appear to have improved security.

1 Introduction

1.1 Biomolecular Computation

Recombinant DNA techniques have been developed for a wide class of operations on DNA and RNA strands. There has recently arisen a new area of research known as DNA computing, which makes use of recombinant DNA techniques for doing computation, surveyed in [37]. Recombinant DNA operations were shown to be theoretically sufficient for universal computation [19]. Biomolecular computing (BMC) methods have been proposed to solve difficult combinatorial search problems such as the Hamiltonian path problem [1], using the vast parallelism available to do the combinatorial search among a large number of possible solutions represented by DNA strands. For example, [5] and [41] propose BMC

N. Jonoska et al. (Eds.): Molecular Computing (Head Festschrift), LNCS 2950, pp. 167–188, 2004.

methods for breaking the Data Encryption Standard (DES). While these methods for solving hard combinatorial search problems may succeed for fixed sized problems, they are ultimately limited by their volume requirements, which may grow exponentially with input size. However, BMC has many exciting further applications beyond pure combinatorial search. For example, DNA and RNA are appealing media for data storage due to the very large amounts of data that can be stored in compact volume. They vastly exceed the storage capacities of conventional electronic, magnetic, optical media. A gram of DNA contains about 10^{21} DNA bases, or about 10^8 tera-bytes. Hence, a few grams of DNA may have the potential of storing all the data stored in the world. Engineered DNA might be useful as a database medium for storing at least two broad classes of data: (i) processed, biological sequences, and (ii) conventional data from binary, electronic sources. Baum [3] has discussed methods for fast associative searches within DNA databases using hybridization. Other BMC techniques [38] might perform more sophisticated database operations on DNA data such as database join operations and various massively parallel operations on the DNA data.

1.2 Cryptography

Data security and cryptography are critical aspects of conventional computing and may also be important to possible DNA database applications. Here we provide basic terminology used in cryptography [42]. The goal is to transmit a message between a sender and receiver such that an eavesdropper is unable to understand it. Plaintext refers to a sequence of characters drawn from a finite alphabet, such as that of a natural language. Encryption is the process of scrambling the plaintext using a known algorithm and a secret key. The output is a sequence of characters known as the ciphertext. Decryption is the reverse process, which transforms the encrypted message back to the original form using a key. The goal of encryption is to prevent decryption by an adversary who does not know the secret key. An unbreakable cryptosystem is one for which successful cryptanalysis is not possible. Such a system is the one-time-pad cipher. It gets its name from the fact that the sender and receiver each possess identical notepads filled with random data. Each piece of data is used once to encrypt a message by the sender and to decrypt it by the receiver, after which it is destroyed.

1.3 Our Results

This paper investigates a variety of biomolecular methods for encrypting and decrypting data that is stored as DNA. In Section 2, we present a class of DNA cryptography techniques that are in principle unbreakable. We propose the secret assembly of a library of one-time-pads in the form of DNA strands, followed by a number of methods to use such one-time-pads to encrypt large numbers of short message sequences. The use of such encryption with conventional electronic media is limited by the large amount of one-time-pad data which must be created and transmitted securely. Since DNA can store a significant amount of information in a limited physical volume, the use of DNA could mitigate this