Finding secret messages in DNA microdots

Stephen A. Jarvis[1,*], Jason S. Mirsky[1], John F. Peden[2] and Nigel J. Saunders[3†]

Programming Research Group, Oxford University Computing Laboratory[1],
University of Oxford, Oxford OX1 3QD, UK.
MRC Molecular Haematology Group[2], Molecular Infectious Diseases Group[3],
Institute of Molecular Medicine,
University of Oxford, Oxford OX3 9DS, UK.

September 1, 2000

Abstract

Motivation: A DNA microdot offers a novel way in which to communicate secret information. It is an extension of the more traditional microdot, a greatly reduced photograph of a secret document which replaced a full stop somewhere in an innocent-looking letter. The DNA microdot appears to be the secure modern alternative, exploiting the complexity of DNA in which a coded secret message is hidden. An enemy can only unlock the secret information by first knowing that an intercepted letter is impregnated with microdot DNA, and secondly by finding the message amongst a huge amount of background DNA used to mask the secret information.
Results: Using software developed to identify horizontally acquired DNA, we show that this apparently insurmountable task is in fact possible. With the increased speed of DNA sequencing, the information contained in the DNA microdot is far from secure.
Availability: Software available from first author.
Keywords: DNA; microdot; signature analysis; algorithms.

*now at: Department of Computer Science, University of Warwick, Coventry CV4 7AL, UK. Email: *stephen.jarvis@dcs.warwick.ac.uk*. To whom correspondence should be addressed

†current address: Sir William Dunn School of Pathology, University of Oxford, Oxford. OX1 3RE

Introduction

One of the masterpieces of twentieth century espionage is Professor Zapp's microdot. The microdot was used extensively by German spies during the Second World War to transmit secret information. Traditionally a microdot is a greatly reduced photograph of a secret document which is concealed as a full stop in an ordinary postal letter. The letter would be sent to the intended recipient who, on receiving it, would photo-enlarge the microdot and read the hidden message. Were the letter to be intercepted by an enemy, then the microdot-concealed message should remain secret.

Taylor Clelland *et al.* have extended this method by creating a DNA microdot. Their microdot replaces the reduced photograph by a solution of DNA which is pipetted somewhere in the innocuous letter. This letter is similarly sent to an intended recipient who extracts the DNA from the impregnated letter and retrieves the secret message. If the letter falls into enemy hands, the message remains concealed by two levels of security: firstly the DNA itself is invisible and secondly, even if the DNA is detected, the secret message is masked by the enormous complexity of the DNA in which it is contained [Taylor Clelland *et al.*, 1999].

The DNA microdot is an elaborate communication tool which appears to offer a watertight means of transferring secret messages. However, while this method might seem secure, we show that as the speed of DNA sequencing increases, it will prove to be far from safe.

Creating DNA microdots

Taylor Clelland *et al.* create a DNA microdot by first encrypting a text message as DNA triplets using the simple substitution cipher found in Table 1. The encrypted message is then flanked with 20-base PCR primer sequences and inserted into sonicated human DNA which is between 50 and 150 nucleotides in length. The secret message DNA is added to 225 ng of treated human DNA and pipetted onto a full stop in a letter which is then posted to the intended recipient.

A=CGA	K=AAG	U=CTG	U=ACT
B=CCA	L=TCG	V=CCT	1=ACC
C=GTT	M=TCC	W=CCG	2=TAG
D=TTG	N=TCT	X=CTA	3=GCA
E=GGT	0=GGC	Y=AAA	4=GAG
G=TTT	Q=AAC	=ATA	5=AGA
H=CGC	R=TCA	,=TCG	7=ACA
I=ATG	S=ACG	.=GAT	8=AGG
J=AGT	T=TTC	:=GCT	9=GCG

Table 1: Encryption key for Taylor Clelland's DNA microdots

When the letter is received, the DNA is amplified using PCR analysis and gel electrophoresis. This allows the recipient, who will have been notified separately of the primer sequences flanking the message, to extract the relevant strands of DNA. Once the appropriate DNA has been identified, the DNA is excised, sub-cloned and sequenced. Finally, using the encryption key the DNA is decoded to reveal the original text message.

An adversary's job is not easy. Assuming an enemy knows that an intercepted letter contains a DNA microdot, decrypting a message written using Taylor Clelland's substitution cipher is straightforward using standard computational techniques. However, the difficulty lies in finding a short encoded message in a sequence of DNA potentially millions of times longer than the message itself. Without knowing the primer sequences this is very complicated. If 20-base primer sequences are used, the adversary is faced with 10^{20} possible primer pairs. The problem is compounded by the fact that the large amount of background DNA may in itself take a considerable amount of time to sequence.

In order to account for any errors in DNA synthesis, amplification and sequencing, 100 copies of the secret message are included per haploid genome. It is noted that even this level of repetition appears to produce some degree of error in Taylor Clelland's results, so perhaps in practice the message might be repeated even more than this.

Systems and methods

Breaking the DNA microdot

When intercepting a DNA impregnated letter, an enemy has to tackle three complex stages in order to break the DNA microdot. Firstly he must extract some of the DNA by amplification and sequencing; secondly he must identify which portion (if any) of the DNA contains the encrypted secret message; and finally he must determine the encryption key to decode the information. These stages are discussed in reverse order.

Encryption keys

The encryption key found in [Taylor Clelland *et al.*, 1999] can be broken by a number of tried and tested methods [Kahn, 1996]. In order to represent all the letters in the Roman alphabet, an encryption key requires 3 nucleotides (4^3 allows 64 representable characters). Assuming that it is known to the nearest few thousand nucleotides where in the DNA sequence the secret message is, it is very easy to construct a computer program to break this key. Each trinucleotide can be systematically replaced with one of the 64 potential alphanumeric symbols. By comparing the results against dictionary files, non-random sequences of characters can quickly be detected.

Of course extending the encryption key so that each character is represented by 4 nucleotides increases the complexity of the problem exponentially. The length of the encryption key can be extended as far as necessary, for example so that there are 20 nucleotides per character. However, it is noted that the larger the encryption key, the easier it is to detect the secret message within the background DNA. In summary, it is harder to detect a secret message with a smaller encryption key, yet once it has been found, the key is easier to break; conversely, it is easier to detect a secret message with a larger encryption key, yet once the message has been found it is harder to break. Either way, the problem proves solvable.

Detecting secret message regions

It is not feasible to try and match each of the 10^{20} possible primer pairs against a segment of DNA and, for each, determine whether all possible encryption keys produce a message which proves meaningful. Instead, a means of detecting secret message 'regions' is needed, which will effectively bypass the complication of the primer pairs.

This problem is analogous to identifying horizontally acquired DNA in bacteria. Using a trinucleotide extension of Karlin's [Karlin and Burge, 1995] dinucleotide signature analysis [Jarvis *et al.*, 2000], we provide a method of detecting regions of DNA which are different enough from the background to warrant further investigation. Even when the chosen encryption key closely resembles the surrounding DNA, this method is sufficiently sensitive to detect the secret message region.

DNA signature analysis is a useful way of detecting when areas of nucleotides display a base distribution which differs in a statistically different way to the normal distribution of bases in a sequence of DNA. Karlin's dinucleotide signature analysis compares the overall proportions of each of the 16 dinucleotides in a sequence of DNA with the proportions found within a smaller subsequence (termed a 'window'). By calculating the dinucleotide difference for windows throughout the sequence, a signature graph is obtained. Unusual concentrations of dinucleotides are highlighted by large peaks in the signature graph. If this method is extended to trinucleotide signature analysis, regions of interest in the DNA microdot can be identified and investigated in more detail, greatly narrowing down the search for the secret message DNA.

Secret message 1

THE SHORELINE OF NORMANDY IS CRAWLING WITH THE ENEMY TROOPS. THE HILLS ARE FILLED WITH SNIPER OUTPOSTS. HOW COULD THEY HAVE KNOWN THAT THE INVASION WAS TODAY THIS WILL RESULT IN A TOTAL LOSS. THE ENEMY SCIENTISTS MUST HAVE DISCOVERED THE DNA MICRODOT. THEY MUST HAVE LOCATED THE MESSAGE USING SHOTGUNCLONES. WE ARE WITHDRAWING ALL TROOPS. THE USS AUGSTA IS HEADING HOME. THE MISSION IS ABORTED. REPEAT. THE MISSION IS ABORTED.

Secret Message 2

ALICE WAS BEGINNING TO GET VERY TIRED OF SITTING BY HER SISTER ON THE BANK, AND OF HAVING NOTHING TO DO: ONCE OR TWICE SHE HAD PEEPED INTO THE BOOK HER SISTER WAS READING, BUT IT HAD NO PICTURES OR CONVERSATIONS IN IT, AND WHAT IS THE USE OF A BOOK, THOUGHT ALICE, WITHOUT PICTURES OR CONVERSATIONS SHE WAS CONSIDERING IN HER OWN MIND AS WELL AS SHE COULD, FOR THE HOT DAY MADE HER FEEL VERY SLEEPY AND STUPID WHETHER THE PLEASURE OF MAKING A DAISYCHAIN WOULD BE WORTH THE TROUBLE OF GETTING UP AND PICKING THE DAISIES, WHEN SUDDENLY A WHITE RABBIT WITH PINK EYES RAN CLOSE BY HER. DOWN THE RABITHOLE, FROM THE ADVENTURES OF ALICE IN WONDERLAND BY LEWIS CARROLL

Figure 1. The messages as found and decrypted by the location and decryption programs

Sequencing

Taylor Clelland *et al.* plant 100 copies of their secret message DNA in background DNA of around 3 billion nucleotides. To ensure the detection of one secret message region, DNA containing an average of 30 million bases therefore needs to be sequenced. Microdot signature analysis works in such a way that as each sequence is produced, it becomes a candidate for analysis. Consequently it is possible to find the encrypted message as soon as it is sequenced, provided that there is a minimum background with which to work.

The sequencing of such a large amount of DNA may still seem a tall order. However, the process is faster than expected because microdot signature analysis does not require the genome to be sequenced

To obtain the trinucleotide signature the odds ratio for each of the 64 trinucleotides is calculated. This probability is the actual frequency of the trinucleotide in the sequence over the expected probability of finding that trinucleotide in the same sequence.

$$p_{(xyz)}(s) = \frac{f_{(xyz)}(s)}{(f_{(x)}(s) * f_{(y)}(s) * f_{(z)}(s))} \tag{1}$$

$f_{(xyz)}(s)$ is the frequency of the trinucleotide xyz in the sequence s.

$$f_{(xyz)}(s) = \frac{\#xyz}{\#trinucleotides} \tag{2}$$

where $\#trinucleotides = \#s - 2$

To calculate a trinucleotide signature, a window of sufficient size (e.g. 1500 bases) is created - this is termed window α. α's trinucleotide probabilities are then compared with the probabilities for the overall sequence, β. This is done for each of the 64 trinucleotides and the normalised results are added. The result is an overall trinucleotide difference between the window (α) and the sequence average (β).

$$\delta = (\frac{1}{64}) \sum_{xyz}^{64} \mid p_{xyz}(\alpha) - p_{xyz}(\beta) \mid \tag{3}$$

By repeatedly sliding the window one position, and recalculating the difference, a trinucleotide signature can be created for the complete sequence.
Such a signature can be achieved for a sequence of approximately 5 million bases in less than a minute on a modest (250MHz) PC.

Figure 2: Equations for trinucleotide signature analysis

in any particular order. Shotgun clones, for example, do not need to be ordered before analysis takes place. Simply concatenating the DNA segments into a large sequence provides the necessary background against which to detect the secret message DNA.

With advances in sequencing technology the DNA microdot will become less secure as a means of secret message transfer. As sequencing becomes faster we may find that discovering a secret message takes less time than it takes the postal service to deliver the letter on which the DNA microdot is placed.

Implementation

To simulate a DNA microdot we retrieved piecewise segments of relatively homogeneous signature human DNA from NCBI Entrez [Benson *et al.*, 1999]. This simulates the process of sonicating and denaturing DNA which Taylor Clelland *et al.* perform. We concatenated these pieces, in no particular order, to form a sequence of approximately one million bases. Using Taylor Clelland's encryption cipher, we encrypted two randomly chosen secret messages, see Figure 1. Our experiments were based on inserting zero or more copies of one of the secret messages into the NCBI human DNA sequence and recording the detection rate.

All experiments were conducted as blind tests so that the detection of the encrypted message was based solely on experimental results and without any prior knowledge of its location.

Two methods were used for finding encrypted messages concealed in the DNA. The first involved trinucleotide signature analysis (Figure 2 Eq. 3). The

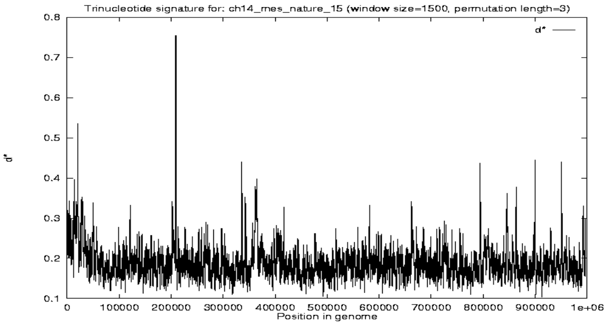

Figure 3. Locating an embedded message in a DNA microdot. The trinucleotide sliding window signature easily finds the encrypted message (at the 200000 mark in these results).

signature is calculated from ratios of the actual frequencies of the trinucleotides in the sequence to their expected frequencies based on a first order probability (Figure 2 Eq. 1 & 2). The second method exploits any repeats of the secret message. If the secret message appears a number of times in a given sequence it is possible to detect the location of the message by discovering the repeats of the primer sequences and/or the cipher text within the DNA.

Discussion

Results

In 40 blind tests our trinucleotide extension to Karlin's dinucleotide signature analysis easily located a single copy of the encrypted message placed in the sequence, see Figure 3.

The trinucleotide signature analysis is effective on non-repeated messages as small as 250 characters (cipher-text size approximately 750 nucleotides). It is also possible to detect similar sized segments of smaller repeated messages.

The DNA context in which the message is embedded is important when trying to conceal the message. DNA signature analysis relies on comparing a smaller context with a larger context and assessing the difference. Messages composed and inserted without regard to context are easily found.

The second method of message detection has no minimum message size as it works solely on the basis of the repeated primers. In each of the 40 blind tests the two test messages were easily detected. It was even possible to detect an 'empty' message from the repeats of the primers alone. The primer sequences are designed to alert the intended recipient to the location of the secret message. The fact that these primers are supposed to be unique means that if they are repeated they are likely to be detected. Of course there often are repeats of 20-base words in DNA sequences, but their occurrence is rare enough to warrant an investigation which quickly produces a result in the case of a hidden secret message.

Conclusions

Taylor Clelland *et al.* have shown that it is possible to apply modern DNA technology to the traditional microdot to create a doubly secure method of passing secret information.

This paper describes how these DNA microdots can be broken.

Using trinucleotide signature analysis it is possible to detect small or repeated messages consisting of 750 nucleotides (250 characters) or more. The secret message region is highlighted clearly on the signature graph allowing the location of the secret message to be pinpointed.

Messages smaller than 250 characters can also be detected by signature analysis if the user fails to take into account the contextual DNA in which the secret message is placed. The DNA background in which copies of the secret message DNA are inserted is important. If an encryption key is chosen which closely matches the background DNA the detection of small messages is made more difficult.

Using our second method - based on detecting the repeats of the primer sequences - there is no limit to the size of the message which can be detected. This is a reliable method because successful communication of the message requires it, and its primer sequences, to be repeated within the source DNA.

The computing power which we employ for trinucleotide signature and repeats analysis is modest. Our 1000-line C program is able to compute the required results in less than a minute and cracking the encryption key can be done just as quickly. With this in mind, the DNA microdot is clearly not as secure as it might first seem.

Acknowledgements

Nigel Saunders is supported by a Welcome Trust fellowship in medical microbiology.

References

Benson,D., Boguski,M., Lipman,D., Ostell,J., Ouellette,B., Rapp,B. and Wheeler,D. (1999) Genbank. *Nucl. Acids Res.* **27** 12-7.

D. Kahn,D. (1996) *The Codebreakers*. Scribner, New York.

Karlin,S. and Burge,C. (1995) Dinucleotide relative abundance extremes: a genomic signature. *Tren. Genet.* **11** 283-290.

Jarvis,S.A., Mirsky,J.S., Pedan,J.F. and Saunders,N.J. (2000) Identification of horizontally acquired DNA using genome signature analysis. Submitted for publication, 2000

Taylor Clelland,C., Risca,V. and Bancroft,C. (1999) Hiding Messages in DNA Microdots. Nature **399** 533-534.

References

[Benson *et al.*, 1999] D. Benson, M. Boguski, D. Lipman, J. Ostell, B. Ouellette, B. Rapp and D. Wheeler, *Genbank*, Nucl. Acids Res. **27(1)**, 12-7, 1999

[Kahn, 1996] D. Kahn, *The Codebreakers*, Scribner, New York, 1996

[Karlin and Burge, 1995] S. Karlin and C. Burge, *Dinucleotide relative abundance extremes: a genomic signature*, Trends Genet. **11(7)**, pp 283-290, 1995

[Jarvis *et al.*, 2000] S. Jarvis, J. Mirsky, J. Pedan and ,N, Saunders *Identification of horizontally acquired DNA using genome signature analysis*, Submitted for publication, 2000

[Taylor Clelland *et al.*, 1999] C. Taylor Clelland, V. Risca and C. Bancroft, *Hiding Messages in DNA Microdots*, Nature **399(6)**, pp 533-534, 1999