DNA-based Cryptography

Ashish Gehani, Thomas H. LaBean and John H. Reif

5th Annual DIMACS Meeting on DNA Based Computers (DNA 5), MIT, Cambridge, MA, June 1999.

Biotechnological Methods **(e.g., recombinant DNA) have been developed for a wide class of operations on DNA and RNA strands**

Biomolecular Computation (BMC)

makes use of such biotechnological methods for doing computation

- **Uses DNA as a medium for ultra-scale computation**
- **Comprehensive survey of Reif [R98]**
- **splicing operations allow for universal computation [Head92].**
- **BMC solution of combinatorial search problems:**

  Hamiltonian path problem [Adleman94]

  Data Encryption Standard (DES)
  [Boneh, et al 95] [Adleman, et al 96]

ultimately limited **by volume requirements, which may grow exponentially with input size.**

DNA Storage of Data

- **A medium for *ultra-compact information storage:* large amounts of data that can be stored in compact volume.**

- **Vastly exceeds storage capacities of conventional electronic, magnetic, optical media.**

- **A *gram* of DNA contains 10^{21} DNA bases**
  = *10^{8} tera-bytes.*

- **A few grams of DNA may hold *all data stored in world.***

- **Most recombinant DNA techniques are applied at concentrations of 5 grams of DNA per liter of water.**

DNA Data Bases:

- **A "wet" data base of *biological data***

  natural DNA **obtained from biological sources may be recoded using nonstandard bases [Landweber,Lipton97], to allow for subsequent BMC processing.**

- **DNA containing data obtained from more conventional *binary storage media.***

  input and output of the DNA data **can be moved to conventional binary storage media by *DNA chip arrays***

  binary data may be *encoded* in DNA strands by use of an alphabet of short oligonucleotide sequences.

Associative Searches within DNA databases:

- **methods for fast associative searches within DNA databases using hybridization [Baum95]**
- **[Reif95] data base join operations and various massively parallel operations on the DNA data**

Cryptography

Data security and cryptography are *critical* to computing data base applications.

***Plaintext*: non-encrypted form of message**

***Encryption*: process of scrambling plaintext message, transforming it into an encrypted message (*cipher text)*.**

Example:

a fixed codebook provides an initial mapping from characters in the finite plaintext alphabet to a finite alphabet of codewords,

then a sophisticated algorithm depending on a key may be applied to further encrypt the message.

***Decryption*: the reverse process of transforming the encrypted message back to the original plaintext message.**

***Cryptosystem*: a method for both encryption and decryption of data.**

***Unbreakable cryptosystem:* one for which successful cryptanalysis is not possible.**

Our *MAIN RESULT:*

DNA-based, molecular cryptography systems

- **plaintext message data encoded in DNA strands by use of a (publicly known) alphabet of short oligonucleotide sequences.**
- **Based on *one-time-pads* that are in principle *unbreakable*.**

One-time-pads may be practical for DNA:

Practical applications of cryptographic systems based on one-time-pads are *limited in conventional electronic media*, by the size of the one-time-pad.

DNA provides a much more *compact storage media*, and an extremely small amount of DNA suffices even for huge one-time-pads.

Our DNA one-time-pad encryption schemes:

- **a *substitution method* using libraries of distinct pads, each of which defines a specific, randomly generated, pair-wise mapping**

- **an *XOR scheme* utilizing molecular computation and indexed, random key strings**

Applications of DNA-based cryptography systems

- **the encryption of (recoded) *natural DNA***
- **the encryption of DNA encoding *binary data*.**

Methods for *2D data input and output:*

- **use of *chip-based DNA micro-array* technology**
- **transform between conventional binary storage media via (photo-sensitive and/or photo-emitting) DNA chip arrays**

DNA Steganography Systems:

- ***secretly tag*** **the input DNA**
- **then *disguise* it (without further modifications) within collections of other DNA.**
- **original plaintext is *not actually encrypted***
- **very appealing due to *simplicity*.**

Example:

DNA plaintext messages are appended with one or more secret keys

resulting appended DNA strands are hidden by mixing them within many other irrelevant DNA strands (e.g., randomly constructed DNA strands).

[Clelland, Risca, and Bancroft]
genomic steganography:
techniques using amplifiable microdots

Our *RESULTS* for *DNA Steganography Systems:*

- **Potential Limitations of these DNA Steganography methods:**

  Show certain DNA steganography systems can be *broken*, with some assumptions on information theoretic entropy of plaintext messages.

- **We also discuss various modified DNA steganography systems which appear to have *improved security*.**

Organization of Talk

- ***Introduction*** **of BMC and cryptography terminology, and results.**

- ***Unbreakable DNA crptosystems*** **using randomly assembled** ***one-time pads*****.**

- **Example of a** ***DNA cryptosystem for two dimensional images*****, using a DNA chip for I/O and also using a randomly assembled one-time pad.**

- ***DNA Steganography Techniques:***

show that they can be ***broken*** **with some modest assumptions on the entropy of the plaintext, even if they employ perfectly random one-time pads.**

Provide possible improvements

- **Conclusions**

Cryptosystems Using Random One-Time Pads

Use *secret codebook* to convert short segments of plaintext messages to encrypted text:

Must be *random* codebook
Codebook can be used only *once*

***In secret*, assemble a large one-time-pad in the form of a DNA strand:**

randomly assembled from short oligonucleotide sequences,
isolated, and cloned.

One-time-pad *shared in advance* by both the sender and receiver of the secret message:

requires initial communication of one-time-pad between sender and receiver
facilitated by compact nature of DNA

A DNA Cryptosystem Using Substitution

Substitution **one-time-pad encryption:**

- **a substitution method using libraries of distinct pads, each of which defines a specific, randomly generated, pair-wise mapping.**

- **The decryption is done by similar methods.**

Input:
plaintext binary message of length n,
partitioned into plaintext words of fixed length,

Substitution One-time-pad:
a table randomly mapping all possible strings of plaintext words into cipher words of fixed length, such that there is a unique reverse mapping.

Encryption:
by substituting each ith block of the plaintext with the cipher word given by the table, and is decrypted by reversing these substitutions.

DNA Implementation of Substitution One-time-pad Encryption:

- ***plaintext*** **messages:**
  one test tube of short DNA strands
- ***encrypted*** **messages:**
  another test tube of different short DNA strands

Encryption by *substitution*:
maps these in a random yet reversible way

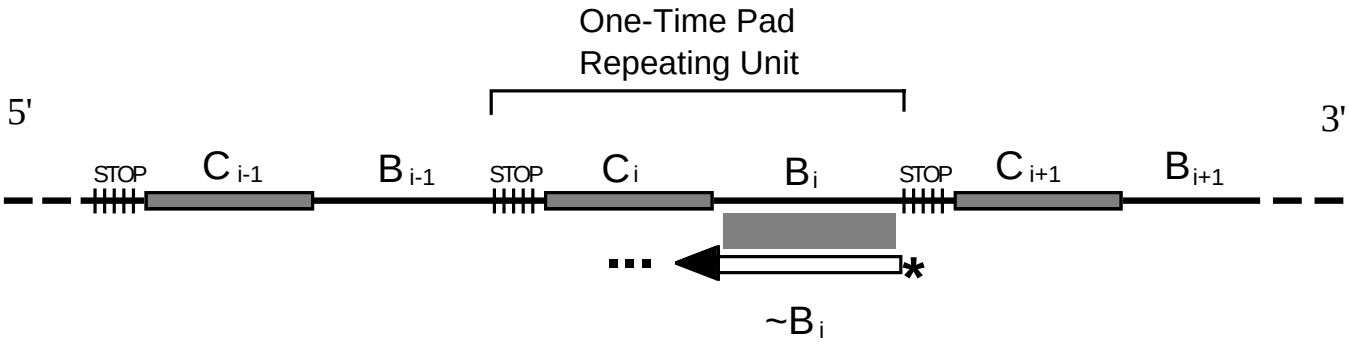
plaintext is converted to cipher strands and plaintext strands are removed

DNA *Substitution one-time pads*:
use long DNA pads containing many segments: each segment contains a cipher word followed by a plaintext word.
cipher word: acts as a hybridization site for binding of a primer
cipher word is appended with a plaintext word to produce word-pairs.

These word-pair DNA strands used as a lookup table in conversion of plaintext into cipher text.

One-time-pad DNA Sequence:

- **Length n**
- **Contains d = n/(L_1+ L_2+ L_3) copies of repeating unit:**



Repeating unit **made up of:**

- **B_i = a cipher word of length $L_1 = c_1 \log n$**

- **C_i = a plaintext word length $L_2 = c_2 \log n$**

  Each sequence pair uniquely associates a plaintext word with a cipher word.
- **Polymerase "stopper" sequence of length $L_3 = c_3$,**

To generate a set of oligonucleotides corresponding to the plaintext/cipher word-pair strands:

- **~B_i used as polymerase primer**
- ***extended*** **with polymerase by specific attachment of plaintext word C_i.**
- ***Stopper sequence*** **prohibits extension of growing DNA strand beyond boundary of paired plaintext word.**

Word-pair strands are essentially: a *lookup table for a random codebook.*

Feasibility depends upon:

- **size of the lexicon;**
- **number of possible pads available;**
- **size, complexity, and frequency of message transmissions.**

Parameter	Range
Lexicon size	10,000 – 250,000 words
Word size	8 – 24 bases
Message size	5 – 30% of lexicon size
Pad diversity	10^6 - 10^8

Pad diversity: total number of random pads generated during a single pad construction experiment.

Codebook Libraries:

- **previous gene library construction projects [LK93, LB97]**
- **used in DNA word encoding methods used in DNA computation [DMGFS96, DMGFS98, DMRGF+97, FTCSC97, GDNMF97, GFBCL+96, HGL98, M96].**

Use *two distinct lexicons* of sequence words:

- **for cipher words**

- **for plaintext words.**

Can *generate lexicons* by normal DNA synthesis methods:

- **utilize sequence *randomization at specific positions* in sequence words.**

Example:
For N = A+C+G+T, R = A+G, and Y = C+T,

RNNYRNRRYN

produces 2x4x4x2x2x4x2x2x2x4 = 16, 384 possible sequences.

Methods for Construction of DNA one-time pads.

(1) *Random assembly* of one-time pads in solution (e.g. on a synthesis column).

- ***Difficult to achieve both full coverage* and yet still avoiding possible *conflicts by repetition* of plaintext and/or cipher words.**

- **can set c_1 and c_2 large so probability of repeated words on pad of length n is small, but coverage is be reduced.**

(2) Use of *DNA chip technology* for random assembly of one-time pads

Advantages:

currently commercially available (Affymetrix) chemical methods for construction of custom variants are well developed.

direct control of coverage and repetitions

DNA chip Method for Construction of DNA one-time pads.

- **an array of immobilized DNA strands,**
- **multiple copies of a single sequence are grouped together in a microscopic pixel.**
- **optically addressable**
- **known technology for synthesis of distinct DNA sequences at each (optically addressable) site of the array.**
- **combinatorial synthesis conducted in parallel at thousands of locations:**

  For preparation of oligonucleotides of length L, the 4^L sequences are synthesized in 4n chemical reactions.

Examples:

- **65,000 sequences of length 8 use 32 synthesis cycles**
- **1.67×10^7 sequences of length 10 use 48 cycles**

DNA Chip Method for

Construction of DNA One-time pads

- **plaintext and cipher pairs constructed:**

- **nearly complete coverage of the lexicon on each pad, nearly unique word mapping between plaintext and cipher pairs.**

- **resulting cipher word, plaintext word pairs can be assembled together in random order (with possible repetitions) on a long DNA strand by a number of known methods:**

  blunt end ligation

  hybridization assembly with complemented pairs [Adleman97]

- **Cloning or PCR used to amplify the resulting one-time pad.**

XOR One-time-pad (Vernam Cipher) Cryptosystem

One-time-pad S:
a sequence of independently distributed random bits

M: a plaintext binary message of n bits

- ***Encrypted bits:***

$C_i = M_i$ XOR S_i for = 1,…,n.

XOR: given two Boolean inputs, yields 0 if the inputs are the same, and otherwise is 1.

- ***Decrypted bits:***

Use commutative property of XOR

$$C_i \text{ XOR } S_i = (M_i \text{ XOR } S_i) \text{ XOR } S_i$$
$$= M_i \text{ XOR } (S_i \text{ XOR } S_i)$$
$$= M_i.$$

DNA Implementation of XOR One-time-pad Cryptosystem

- ***plaintext messages:***
  one test tube of short DNA strands

- ***encrypted messages:***
  another test tube of different short DNA strands

Encryption by XOR One-time-pad:

maps these in a random yet reversible way

plaintext is converted to cipher strands and plaintext strands are removed

For *efficient* DNA encoding:

use *modular base 4* (DNA has four nucleotides)

Encryption:

addition of one-time-pad elements modulo 4

Decryption:

subtract one-time-pad elements modulo 4

Details of DNA Implementation of XOR One-time-pad Cryptosystem

- **Each plaintext message has appended unique *prefix index tag* of length L_0 indexing it.**

- **Each of one-time-pad DNA sequence has appended unique *prefix index tag* of same length L_0, forming *complements* of plaintext message tags.**

- **Use Recombinant DNA techniques (annealing and ligation) to *concatenate into a single DNA strand* each corresponding pair of a plaintext message and a one-time-pad sequence**

- **These are *encyphered by bit-wise XOR computation:***
  fragments of the plaintext are converted to cipher strands using the one-time-pad DNA sequences, and
  plaintext strands are removed.

***Reverse decryption* is similar:**
use commutative property of bit-wise XOR operation.

BMC Methods to effect bit-wise XOR on Vectors.

Can adapt BMC methods for *binary addition*:

- **similar to bit-wise XOR computation**
- **can *disable carry-sums* logic to do XOR**

BMC techniques for Integer Addition:

(1) [Guarnieri, Fliss, and Bancroft 96] first BMC addition operations (on single bits).

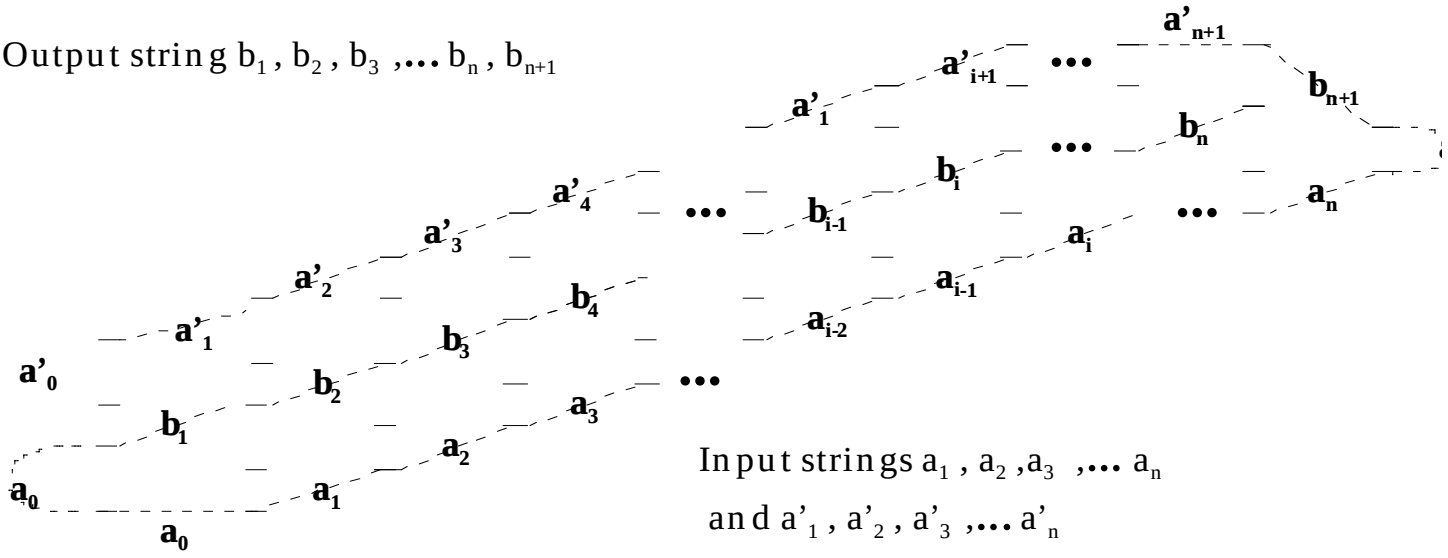
(2) [Rubin el al 98, OGB97,LKSR97,GPZ97] permit chaining on n bits.

(3) Addition by *Self Assembly* of DNA tiles [Reif,97][LaBean, et al,99]

XOR by Self Assembly of DNA tiles

[LaBean, et al,99]

Output string $b_1, b_2, b_3, \ldots b_n, b_{n+1}$



Input strings $a_1, a_2, a_3, \ldots a_n$

and $a'_1, a'_2, a'_3, \ldots a'_n$

XOR by Self Assembly of DNA tiles

[1] For each bit M_i of the message, construct sequence a_i that represents the ith bit.

[2] Scaffold strands for binary inputs to the XOR:

- **Using linkers, assemble message M's n bits into scaffold strand sequence $a_1 a_2 \ldots a_n$,**
- **One-time-pad is further portion scaffold strand $a'_1 a'_2 \ldots a'_n$ is created from random inputs**

[3] add output tiles; annealing give self assembly of the tiling.

[4] adding ligase yeilds reporter strand
$R = a_1 a_2 \ldots a_n.a'_1 a'_2 \ldots a'_n.b_1 b_2 \ldots b_n$
where $b_i = a_i$ XOR a'_i, for i = 1,…,n.

[5] reporter strand is extracted by melting away the tiles' smaller sequences, and purifying.

contains concatenation of:
input message, encryption key, ciphertext

[6] Using a marker sequence:
ciphertext can be excised and separated based on its length being half that of remaining sequence.

[7] Ciphertext can be stored in a compact form

DNA Cryptosystem for 2D Images

using:

- **DNA Chip**
- **Randomly Assembled One-Time Pad**

Encryption and Decryption of 2D images recorded on microscopic arrays of a DNA chip:

Message

Encrypted

Decrypted

Simulated patterns observed by fluorescence microscopy of the DNA I/O chip.

DNA Cryptosystem consists of:

- **Data set to be encrypted: *2-dimensional image***
- ***DNA Chip* bearing immobilized DNA strands:**
  contains an addressable array of nucleotide sequences immobilized s.t. multiple copies of single sequence grouped together in a microscopic pixel.
- ***Library of one-time pads* encoded on long DNA strand**

Initialization and Message Input

- **Fluorescent-labeled, word-pair DNA strands are prepared from a substitution pad codebook**

- **These are annealed specifically to their sequence complements at unique sites (pixels) on the DNA chip.**

- **The message information is transferred to a photo mask with transparent (white) and opaque (black) regions:**

Message Input to DNA Chip

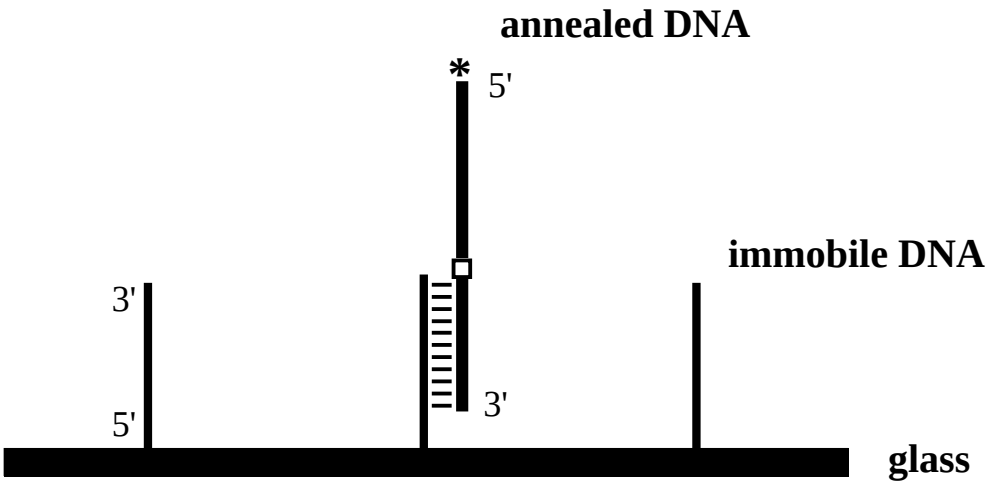
Initialization and Message Input

- ***Immobile DNA strands*** **are located on the glass substrate of the chip in a sequence addressable grid.**
- ***Word-pair strands*** **are prepared from a random substitution pad:**

  the 5' (*unannealed*) end carries a cipher word

  the 3' (*annealed*) end carries a plaintext word.

  contain a *photo-cleavable* base analog between two sequence words (added to 3' end of cipher word during oligo synthesis)



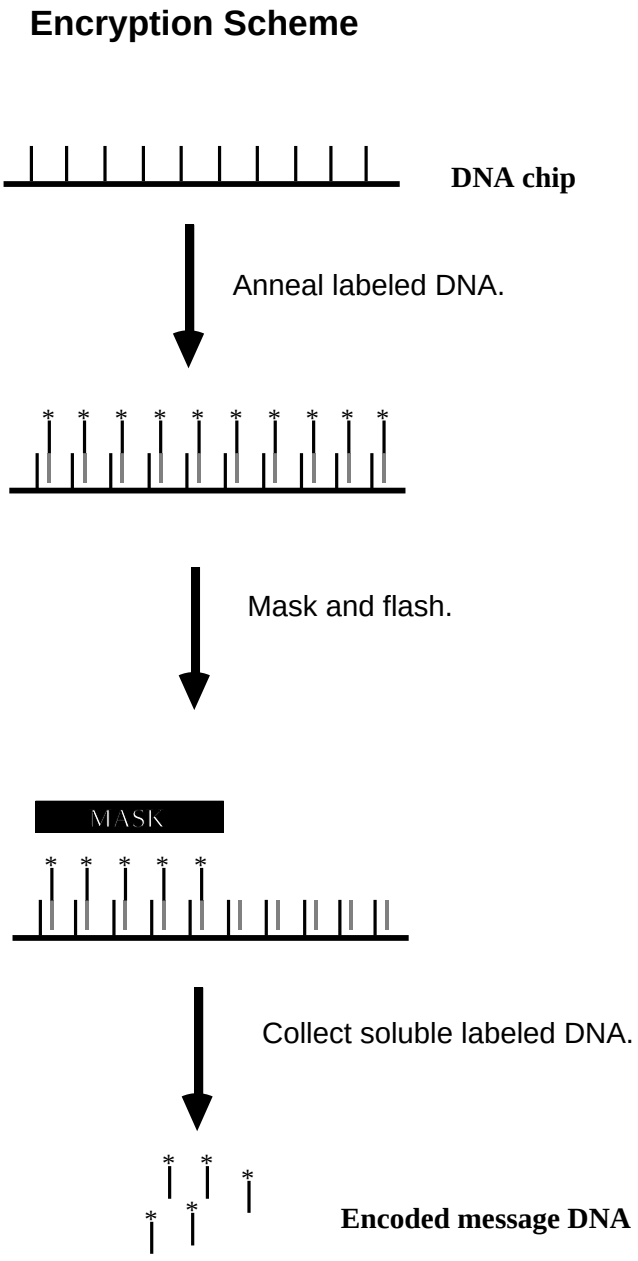
- **The *annealed DNA* contains:**

  a fluorescent label on its 5' end (asterisk);

  a codebook-matching sequence word (not base-paired on the chip);

  a photo-labile base (white square) capable of cleaving the DNA backbone; and

  a chip-matching word (base-paired to immobile strand).



Encryption Procedure:

[1] start with DNA chip displaying sequences *complementary* to plaintext lexicon.

[2] fluorescent-labeled word-pair strands from one-time-pad are *annealed* to chip at pixel bearing complement to plaintext 3' end.

[3] mask protects some pixels from a light-flash. At unprotected regions, DNA is *cleaved* between plaintext and cipher words.

[4] cipher word strands, still labeled with fluorophore at 5' ends, are collected and transmitted as *encrypted message*.

Encryption of the Message

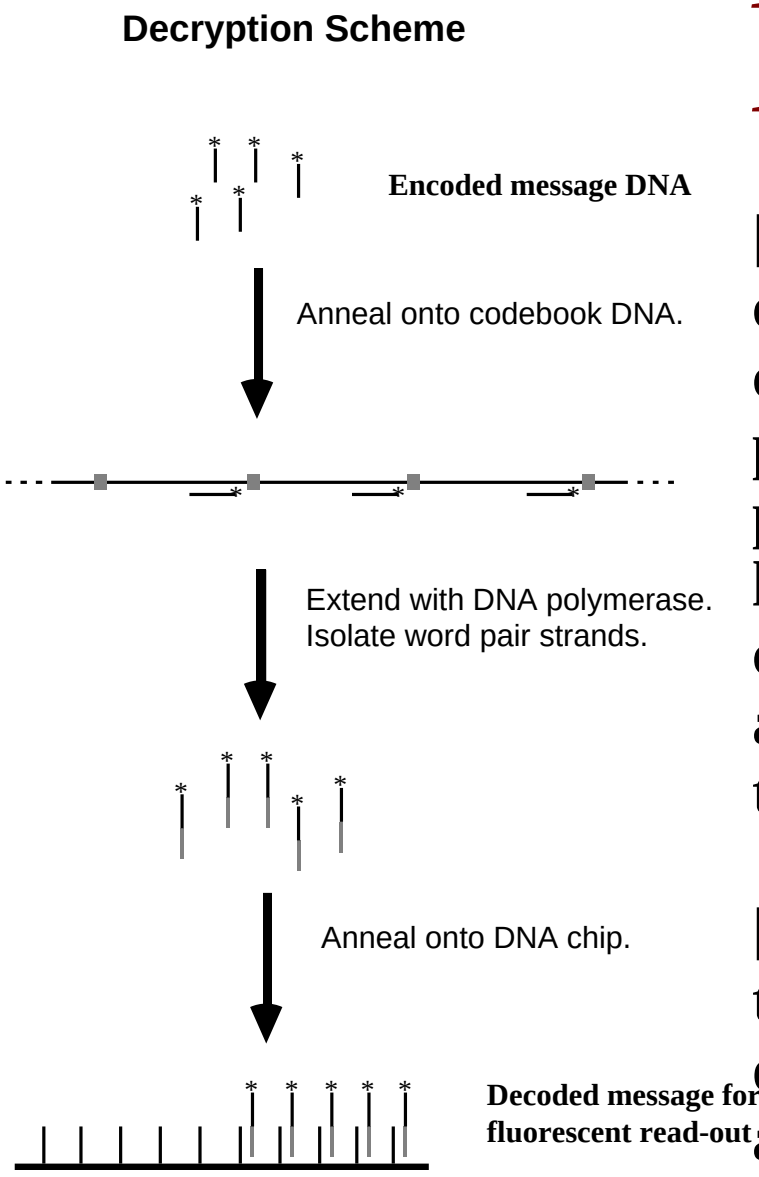
- **Following a light-flash of mask-protected chip, *annealed* oligonucleotides beneath transparent mask pixels are *cleaved* at a photo-labile position: their 5' sections are dissociated from annealed 3' section and collected in solution.**

- **This test tube of strands is *encrypted message*.**

- **Annealed oligos *beneath opaque mask* are unaffected by light-flash and can be *washed off chip*.**

- **If encrypted message oligos are reannealed onto a (washed) DNA chip, message information would be *unreadable:***

Simulated Read-Out of Encrypted Message from DNA Chip



Decryption Procedure:

[1] word-pair strands constructed, *appending* cipher word with proper plaintext word, by polymerase extension or lop-sided PCR using cipher words as primer and one-time-pad as template.

[2] cipher strands *bind* to their specific locations on the pad and are appended with their plaintext partner.

[3] binding reformed *word-pair strands* to DNA chip and reading message by fluorescent microscopy.

Decryption of the Message

- **use the fluorescent labeled oligos as primers in one-way (lopsided) PCR with the same one-time codebook which was used to prepare the initial word-pair oligos.**

- **When word-pair PCR product is bound to the same DNA chip, the decrypted message is revealed:**

Decrypted Message

Simulated Read-Out of Decrypted Message from DNA Chips

Steganography

a class of techniques that hide secret messages within other messages:

plaintext is not actually encrypted but is instead disguised or hidden within other data.

Historical examples:

- **use of grills that mask out all of an image except the secret message,**
- **micro-photographs placed within larger images**
- **invisible inks, etc.**

Disadvantages:

- **Cryptography literature generally consider conventional steganography methods to have *low security*:**

steganography methods have been often broken in practice [Kahn67] and [Schneier96]

Advantages:

- **it is very appealing due to it's *simplicity*.**

DNA Steganography Techniques:

- **take one or more input DNA strands (considered to be the plaintext message)**

- **append to them one or more randomly constructed *"secret key"* strands.**

- **Resulting *"tagged plaintext"* DNA strands are *hidden* by mixing them within many other additional "distracter" DNA strands which might also be constructed by random assembly.**

Decryption:

- **Given knowledge of the "secret key" strands,**
- **Resolution of DNA strands can be decrypted by a number of possible known recombinant DNA separation methods:**

plaintext message strands may be separated out by hybridization with the complements of the "secret key" strands might be placed in solid support on magnetic beads or on a prepared surface.

These separation steps may combined with amplification steps and/or PCR

Cryptanalysis of DNA Steganography Systems:

DNA steganography system's security is entirely dependent on degree that message DNA strands are *indistinguishable* from "distracter" DNA strands.

Cryptanalysis Assumptions:

- **no knowledge of the "secret key" strands**
- **secret tags are indistinguishable from "distracter" DNA strands.**
- **plaintext is not initially compressed, and comes from a source (e.g., English or natural DNA) with Shannon information theoretic entropy $E_S > 1$**
- **the "distracter" DNA strands are constructed by random assembly**

Then:

the original plaintext portion of "tagged plaintext" DNA strands are *distinguishable* from "distracter" DNA strands, and

the DNA Steganography System can be *broken*

Shannon (information theoretic) Entropy E_S

- **provides a measure of the factor that a source can be *compressed* without loss of information.**

Examples:

many images have entropy nearly 4
English text has entropy about 3
computer programs have entropy about 5
most DNA have entropy range 1.2 to 2

Lossless Data Compression [Lempel-Ziv 77]

Input: **text string of length n with entropy E_S**

[1] Form a *dictionary* D of the d = n/L most frequently occurring subsequences of length at least $L = E_S \log_2 n$ in the known source distribution.

[2] In place of subsequences of the input text matching with elements of the dictionary D, *substitute their indices* in the dictionary D.

Cryptanalysis of DNA Steganography Systems:

Input: **test tube T containing:**
a mixture of "tagged plaintext" DNA strands mixed with a high concentration of "distracter" DNA strands, of length n.

- **form a *dictionary* D of the d = n/L most frequently occurring subsequences of length at least L = $E_S \log_2 n$ in the known plaintext source distribution.**

- **Give procedure for *separating* out plaintext message strands by repeated rounds of hybridization with complements of elements of D.**

r(T) = *ratio of concentration* of "distracter" DNA strands to "tagged plaintext" DNA strands.

On each *round of separation:*
form a new test tube F(T) with expected r(F(T)) considerably *reduced* from the previous ratio r(T).

Separation Procedure:

[1] Pour a fraction s = 1/2 of volume of current test tube T into a test tube T_1 and pour remaining fraction 1-s of T into test tube T_2.

[2] Choose a *random text phrase* x in D (not previously considered in a prior trial), and using Watson-Crick complement of x, do a *separation* on test tube T_2 , yielding a new test tube T_3 whose contents are only DNA strands containing phrase x.

[3] Pour contents of test tubes T_1 and T_3 into a new test tube F(T).

- **Ratio r(F(T)) of "distracter" DNA strands to plaintext DNA will expect to *decrease* from original ratio r(T) by a constant factor c < 1**

- **After O(log(r/r')) repeated rounds of this process, ratio of concentration in test tube T will expect to *decrease* from initially r = r(T) to any given smaller ratio r'.**

Another cryptanalysis technique for breaking steganographic systems:

Cryptanalysis using *"hints"* that disambiguate plaintext.

Example:

- **wish to make secret the DNA of an individual (e.g., the President)**

- **use an improved steganography system where "distracter" DNA strands (that are mixed with DNA of an individual) are DNA from a similar but not identical genetic pool.**

steganography system may often be *broken* by use of distinguishing *"hints"* concerning DNA of the individual

e.g., the individual might have a particular set of observable expressed gene sequences (e.g., for baldness, etc.).

These hints may allow for subsequent identification of the full secret DNA:

use of a series of separation steps with complement of portions of known gene sequences.

Improved DNA Steganography Systems with Enhanced security:

Idea: make it more difficult to *distinguish* probability distribution of plaintext source from that of "distracter" DNA strands.

(1) Mimicking Distribution of "Distracter" DNA:

- **use improved construction of the set of "distracter" DNA strands, so distribution better mimics the plaintext source distribution**
- **construct the "distracter" DNA strands by random assembly from elements of Lempel-Ziv dictionary.**
- ***Drawback:*** **Cryptanalysis using "hints" that disambiguate plaintext.**

(2) Compression of Plaintext.

- **recode the plaintext using a universal lossless compression algorithm (e.g., Lempel-Ziv 77].**
- **resulting distribution of the recoded plaintext approximates a universal distribution, so uniformly random assembled distracter sequences may suffice to provide improved security.**
- ***Drawback:*** **unlike conventional steganography methods, plaintext messages need to be preprocessed.**

Conclusion and Open Problems

Presented an initial investigation of DNA-based methods for Cryptosystems.

- ***Main Results*** **for DNA one-time-pads cryptosystems:**

  Gave DNA substitution and XOR methods based on one-time-pads that are in principle *unbreakable*.

  Gave an implementation of our DNA cyptography methods including *2D input/output.*

- **Further Results for *DNA Steganography:***

  a certain class of DNA steganography methods offer only limited security; can be *broken* with some reasonable assumptions on entropy of plaintext messages.

  modified DNA steganography systems may have *improved security.*

Open Problem:

Show whether DNA steganography systems with natural DNA plaintext input can or cannot be made to be *unbreakable*.