Special course in Computer Science: Molecular Computing

Lecture 12: DNA Memory

Vladimir Rogojin

Department of CS, Abo Akademi

http://combio.abo.fi/teaching/special-course-in-computer-science-molecular-computing/

DNA-based RAM Memory

- It is possible in theory to build memory vastly larger than the brain based on DNA (Baum 1995)
  - Conjectured vessels storing 10^{20} words
    - Compare to 10^{14} synapses in the brain
- can be used in various application fields including biotechnology and nanotechnology
- Scheme for molecular memory:
  - DNA molecule contains an address portion and data portion
- Positive design problem:
  - Each address portion should be accessible only by the sequence that is complementary to the sequence of the address portion.

Sequence Design

- DNA sequences should hybridize with their complements at the same efficiency to prevent hybridization bias.
  - It is necessary to make melting temperature uniform to obtain uniform hybridization efficiency
- GC content – ration of the sum of occurences of G and C:
  - Double strands with a higher GC content tend to have a higher melting temperature
  - Because Double strands with a higher GC content tend to hahydrogen bond between G and C is more stable
  - sequences are often designed such that their GC content is 50%

Specific and non-specific Hybridization

- Specific - intended/expected/planned hybridization/
  - hybridization between an intended pair of sequences
- Non-specific hybridization:
  - hybridization between an unintended pair of sequences
  - Causes various problems in DNA computing:
    - Errors in the computation,
    - Failure the construction of intended DNA nanostructures,
    - Misdiagnosis in DNA microarrays.
- To design specific sequences,
  - We must design sequences among which specific hybridizations are stable and nonspecific ones are unstable.
  - Specific-sequence design requires a metric for stability between two sequences.

H-measure

- The most often used criterion for stability is h-measure (Garzon et al. 1997,1998)
- For two sequences xi and xj h-measure is
- $$|x_i, x_j| := \min_{-n<k<n} H(x_i, \sigma^k(\overline{x_j})),$$
  - H(*, *) denotes the Hamming distance,
  - σ^k denotes the right (left) shift for k>0 (k<0),
  - k denotes the number of the shift,
  - $\overline{x_j}$ is the sequence which is Watson–Crick complementary to x_j.

Hybridization stability

- a double strand with a certain number of base pairs is more stable than one with fewer base pairs

- The large h-measure
  - Less table the double strand

Preventing Secondary Structures

- Secondary structures often prevent the sequence from hybridizing to the correct complementary sequence.
- avoid sequences forming stable secondary structures
  - Unless they are specifically designed to form them.
- mfold program (Zuker 2003) is commonly used to calculate the stability of secondary structures.

Preventing Secondary Structures – constraints

- Repeated Bases
  - harmful effect on specific hybridization.
  - Example: guanine-rich motifs can form four-stranded complexes.
- Forbidden Subsequences
  - Restriction enzyme recognition sequences, unless intendet cut-site
  - When using restriction enzyme, which cuts off the sequence at specific sites
    - The recognition sequence must appear only at an intended site
    - Otherwise, sequences will be cut off at some unintended site.

Preventing Secondary Structures – constraints

- Three-Base Constraint
  - Use only three kinds of bases: A, T, and C
  - Reason:
    - G can form G-C and G-T pairings, and
    - Whiplash polymerase chain reaction (PCR) uses only A, T, C; the fourth base is used as stopper sequences to stop the polymerase extension at the designated position

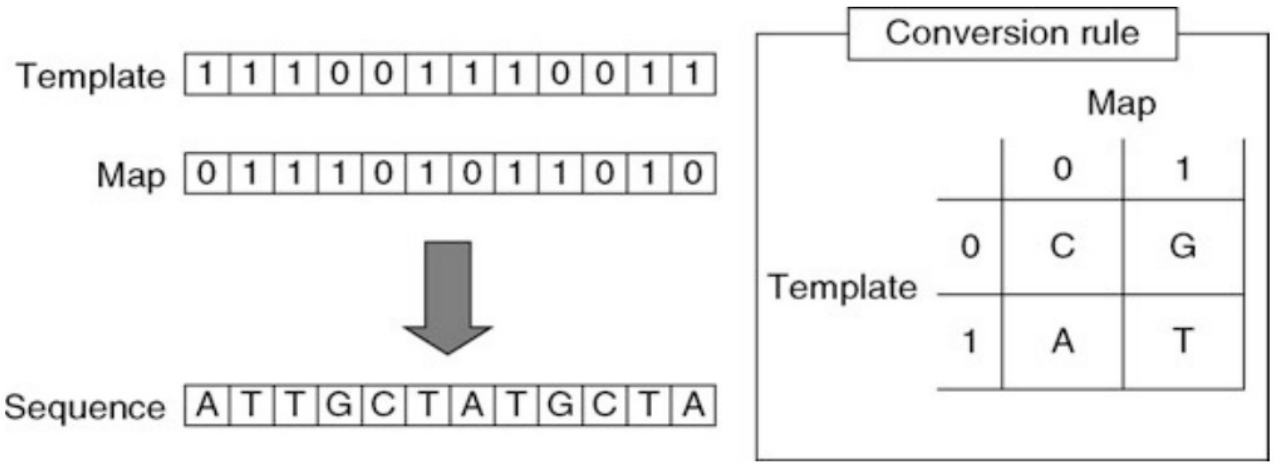
Strategy for Sequence Design

- From coding theory:
  - Maximize the size of a set consisting of equi-length code words so that the Hamming distances between them are above a certain threshold.
  - sequences are produced by crossing two binary codes of length n,
    - namely a template and map.
  - In template for each digit:
    - 0 – position of C or G
    - 1 – position of A or T
  - Adjust correct occurance of 0's in the template to maintain the balance of GC

Template and Map

- For positions t_i=1 using a map $m = m_1m_2...m_n$ (m_i Є0,1).
  - Either A or T is chosen

- For positions t_i=0 using a map $m = m_1m_2...m_n$ (m_i Є0,1).
  - Either G or C is chosen



Library of words

- The problem is to find as many templates and maps as possible
  - such that sequences obtained from them satisfy criteria:
    - uniform GC content,
    - Maximal Hamming distance
    -
- Template method:
  - Map: an error-correcing code
  - Error-correcting code:
    - codewords have at least k mismatches with each other
  - Example:
    - BCH code of length 15 provides at least 128 words with a minimum of five mismatches

Template method

- Template method:
  - Map: an error-correcing code
  - Error-correcting code:
    - codewords have at least k mismatches with each other
  - Constraint:
    - the number of mismatches must be larger than a threschold
  - Features:
    - frameshift hybridization is considered
    - mismatches between sequences and concatenations of two sequences are considered
  - In practice:
    - Sets of templates of length up to 30 can be designed
- Another:
  - Constant weight code – maps instead of templates

De novo design

- generates sequences for which each subsequence of length l appears at most once
- Example:
  - consider a sequence CATGGGAGATGCTTAG, l=6
  - Any subsequence of length 6 such as CATGGG appears only once in this sequence
    - choose a sequence of length l that is not included in a given list of prohibited subsequences and add this sequence to the prohibited list.
    - Prolong to the right the subsequence by concatenating a letter in such a way that the resulting sequence of length l+1 does not have any sequence in the prohibited list as its suffix of length l.
    - Repeat steps above
- consecutive complementary bases are likely to form stable structures
- It still does not guarantee prevention of nonspecific hybridization

Other methods for sequence design

- the sequence design problem can be formulated as a search for sequences that maximize (minimize) the evaluation function or satisfy the constraints.
- Strategies:
  - generate sequences randomly, and then modify them until they satisfy the constraints (software PERMUTE by Faulhammer et al. 2000)
  - stochastic local search (SLS) algorithm:
    - Choose pair of sequences s1 and s2 that violate one of the given constraints;
    - Generate a set S1 of sequences obtained by substituting some bases of s1 and a set S2 by substituting some bases of s2.
    - Select at random s' from S1 or S2 with a given probability, or reduce maximally the number of conflicts
    - If s' from S1, then substitute s1 by s', otherwise substitute s2 by s'
    - Iterate a predefined number of steps

Other methods for sequence design

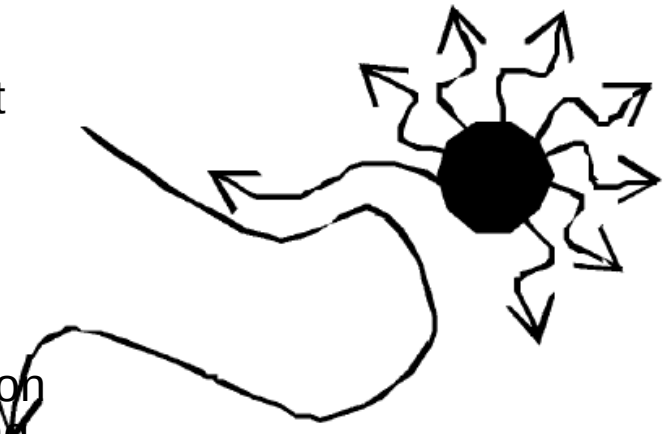
- Genetic algorithms
- Stability estimation based on minimum free energy

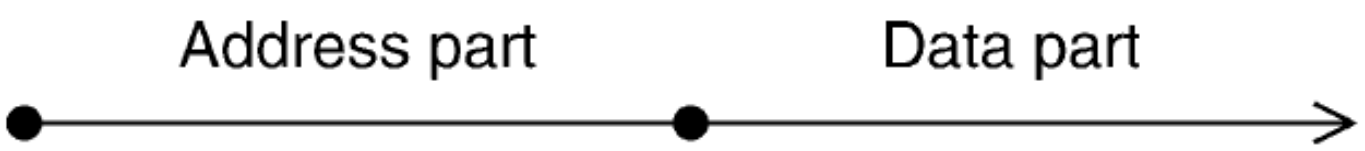
Memory operations

- Issue for construction of molecular memory:
  - how to access a specified word in its address space
  -
- Structure of a memory word:
  - Address:
    - 1 digit
    - Multiple digits (hierarchical address space)
  - Data:
    - Empty
    - 1 bit
    - A word

Access Operations

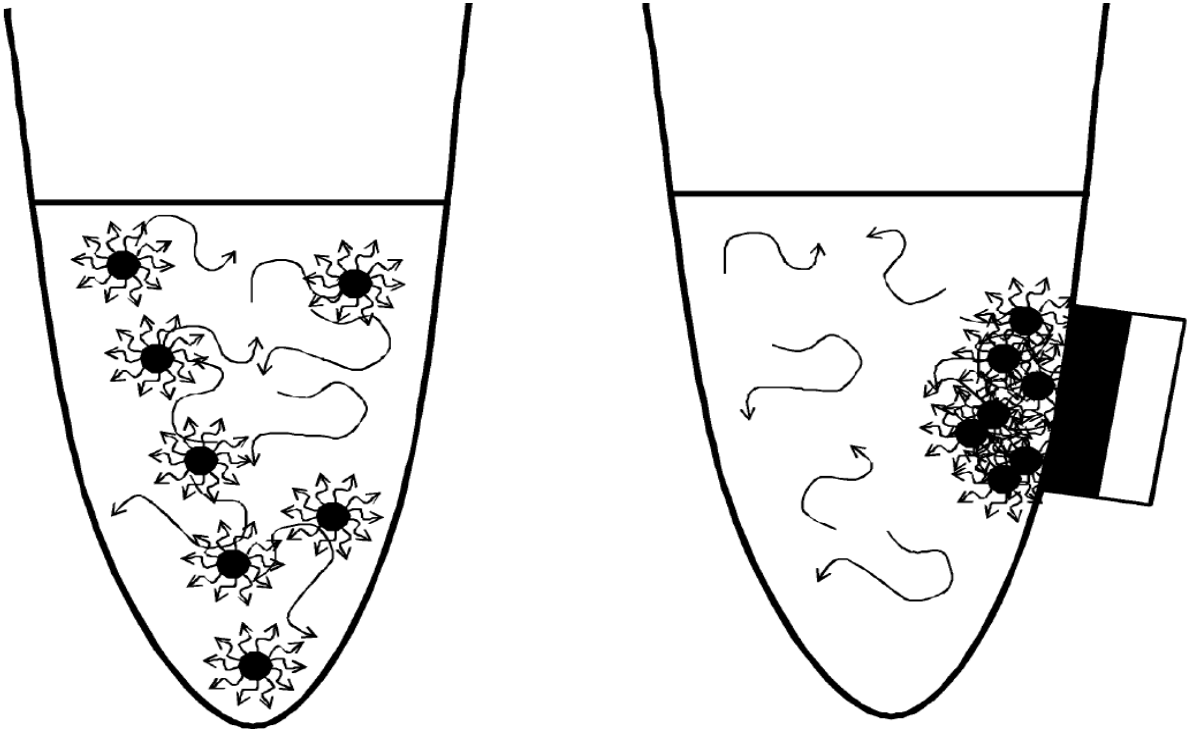
- Extraction by Magnetic Beads
  - any kind of affinity separation based on the address part of a word can be used as an access operation.
  - address part of a memory word - single-stranded segment of a DNA molecule.
  - A strand trapped by a magnetic bead. Probes containing the sequence complementary to the target are attached on the surface of a magnetic bead. A single strand containing the target sequence is trapped by the bead.





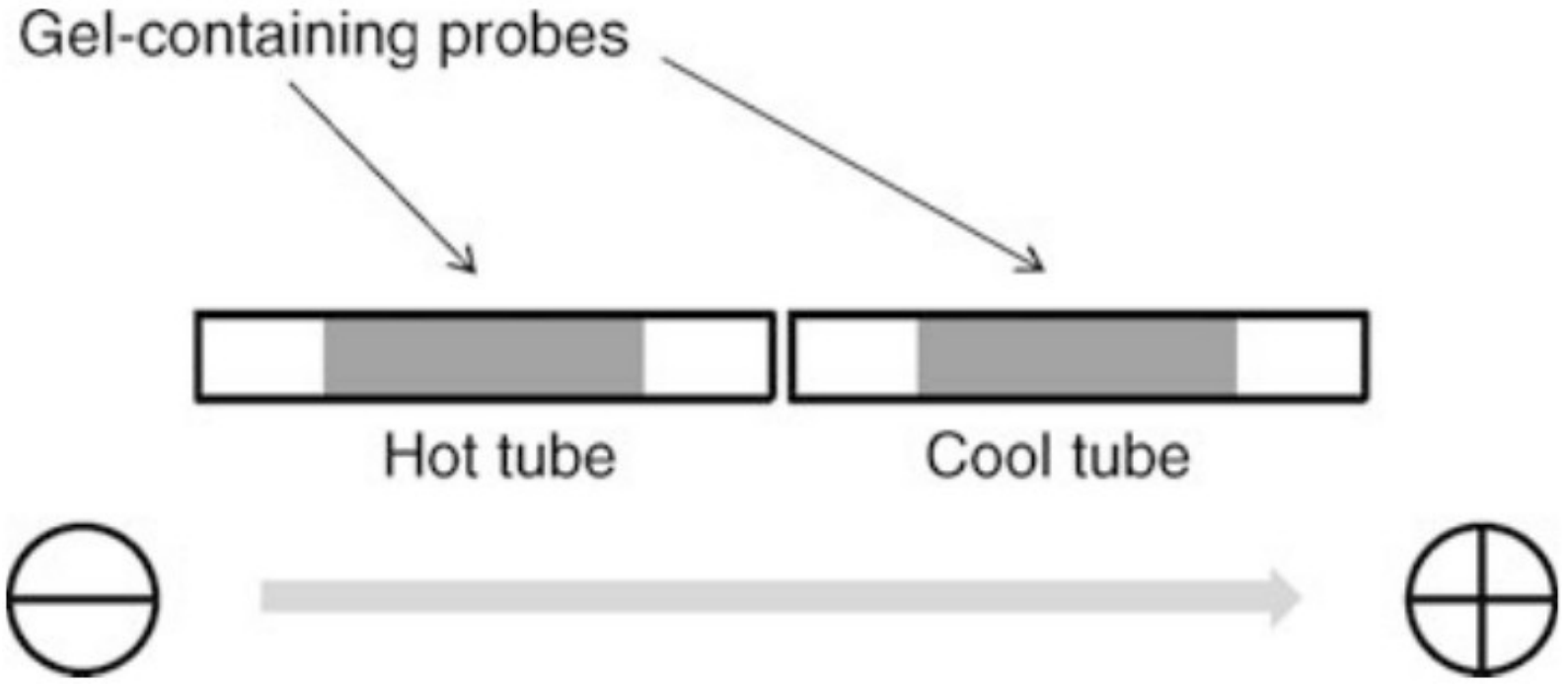
Extraction by Magnetic Beads

- Magnetic beads. In the left panel, magnetic beads are diluted in the tube to trap the target DNA. In the right panel, magnetic beads with the trapped DNA are collected by a magnet.



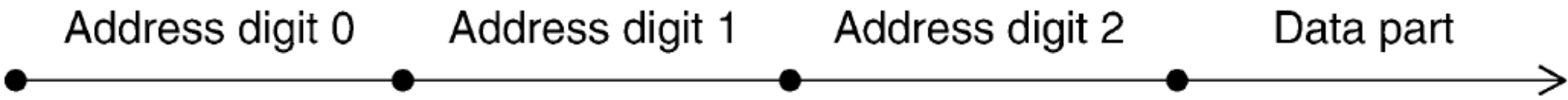
Extraction by Magnetic Beads

- A hot tube connected to a cool tube. DNA molecules released from the hot tube are moved and trapped in the cool tube.



Hierarchical address space

- Each word consists of multiple address digits and a data part.



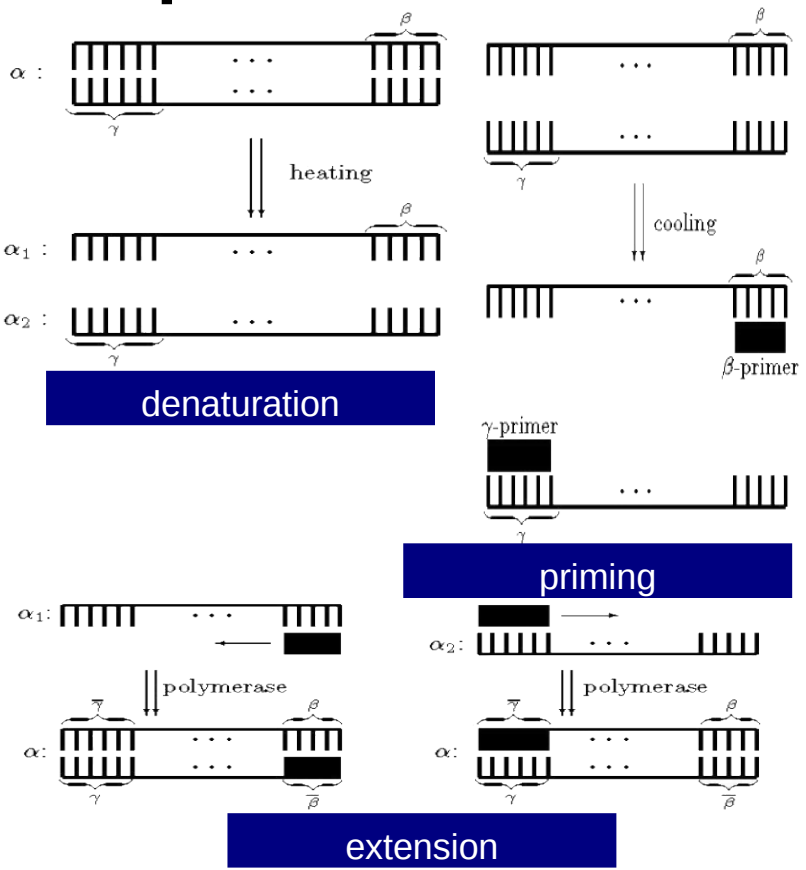
- The access operation is iterated for each digit:
  - affinity separation by the ith digit is applied to the result for (I-1)th digit
  - Each iteration PCR amplification is applied

Extraction by PCR

- PCR:
  - a method to amplify specified DNA molecules

The PCR technique

- **Denaturation:** heat the solution to 85-95 C: *alpha* denatures into two single strands *alpha1* and *alpha2*
- **Priming:** Cool down the solution to 55C: the primers *beta'* and *gamma'* anneal to their complementary borders
- **Extension:** Heat the solution to 72C: polymerase extend the primers to produce two double stranded DNA molecules *alpha*

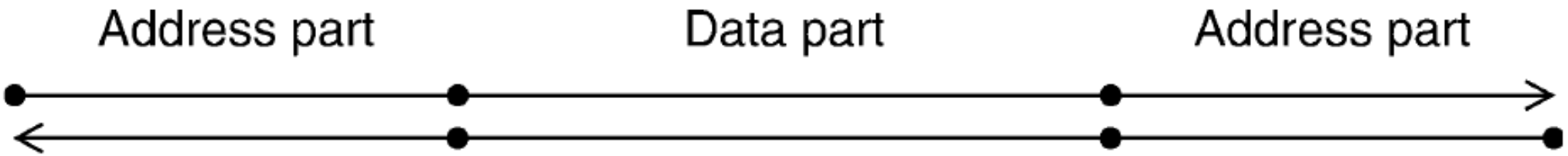


The PCR technique

- Repeat the cycle n times: 2^n copies (in principle): highly efficient bio-copymachine!
- A single cycle takes about 5 minutes: obtain billions of copies in several hours (days for cloning)
- The polymerase must be heat resistant – nature's solution: thermophilic bacteria
- Important observation:
  - one needs to know the borders of the DNA segment to be copied

Accessing by PCR

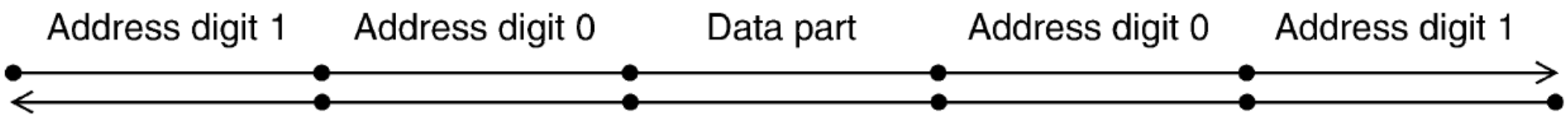
- In order to use PCR for accessing memory words in a DNA memory:
  - each word should have its data part surrounded by two primer sequences at its ends



    - since a memory word can be amplified only if it has both primer sequences.
  - in order to extract a memory word,
    - the solution of the DNA memory is diluted and mixed with the primers in the polymerization buffer.
  - Since PCR only amplifies the target memory word, if the resulting buffer is diluted, the concentration of memory words other than the target becomes less than detectable.

Accessing by PCR, hierarchical address space

- A memory word in a hierarchical molecular memory extractable by PCR. Each word contains nested primer pairs.



Hybridization and Conformational Change

- Instead of extracting a memory word with a specified address, separating it from other words,
  - it is also possible to access a memory word by changing its state,
  - while words with different addresses are unchanged

- In this kind of molecular memory,
  - memory words can be fixed on a surface or in a space

Hybridization of a memory word and a probe

- The address part is changed from a single strand to a double strand

- Simple settings:
  - a memory word does not have a data part
  - This is a write operation
  - Single strand – 0, double strand - 1

A hairpin molecule as a memory word

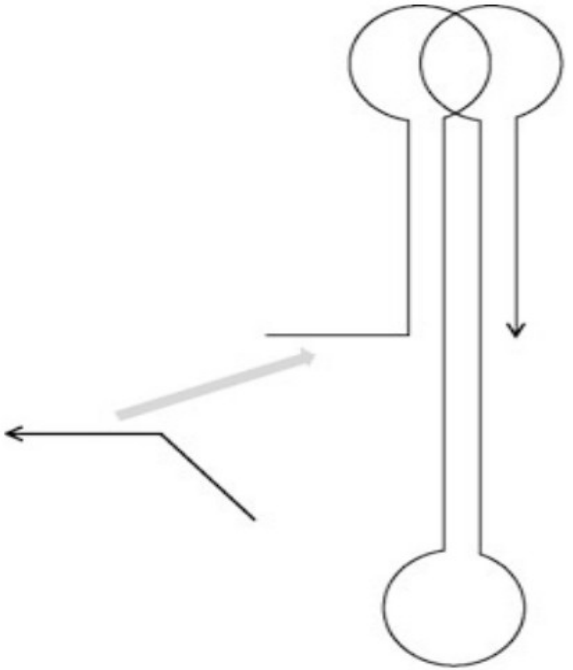
- If the temperature is raised:
  - the hairpin is denatured and
  - it can hybridize with a probe that is complementary to the hairpin loop and one side of the stem

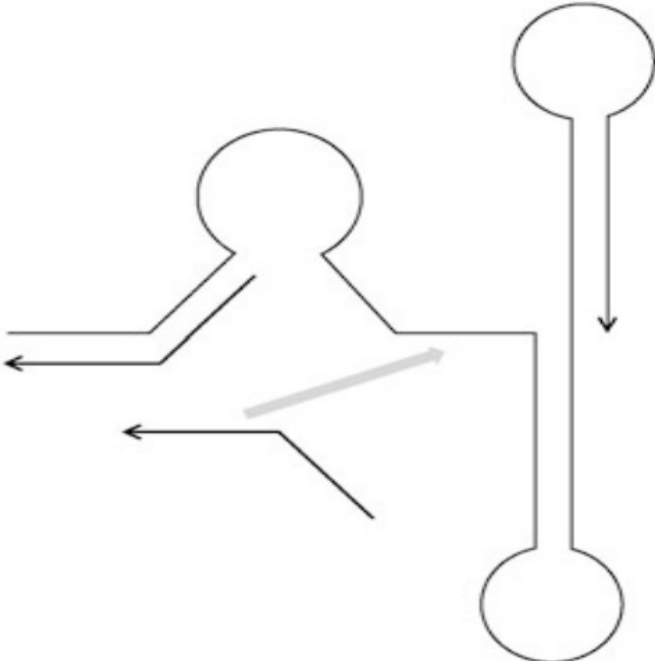
A hairpin molecule as a memory word

- They attach hairpins with different addresses to a gold surface,
- and employ a laser beam to locally raise temperature on the surface.
- Hairpins on the laser spot get denatured
  - In presence of complementary sequences they can hybridize

A hairpin molecule as a memory word with the hierarchical memory

- A memory word consisting of three consecutive hairpins.
- A single strand called an opener hybridizes with the single-stranded part of the memory word and opens the adjacent (first) hairpin.
- As a result, the next opener that opens the second hairpin can hybridize with the exposed stem of the first hairpin.





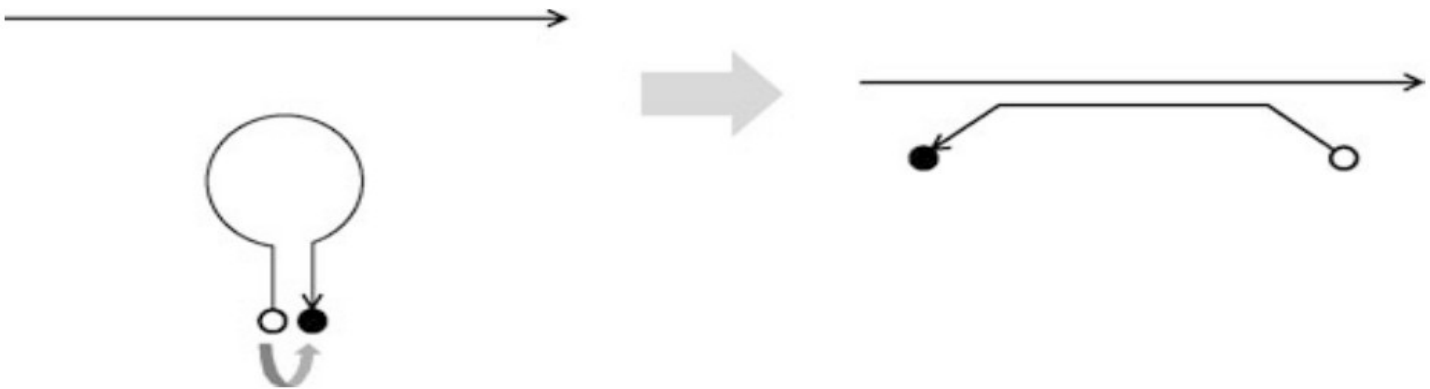
Read Operation

- Case when a memory word consists of only the address part:
  - existence and nonexistence of a word means bit 1 and bit 0, respectively
  - then extracting a memory word is done by reading the bit of the address of the word.

- Case when a memory word contains a data part that encodes some additional information
  - Then it is necessary to read the data part by a separate read operation
  - Read operation (applicable to magnetic bead and PCR):
    - Gel electrophorersis (useful for length-encoded info)
    - Sequencing
    - DNA chip detection (oligonucleotides with fluorescent label)
    - Fluorescence resonance energy transfer (FRET) does not require target amplification (also applicable to hybridization-based)
    - Molecular beacon

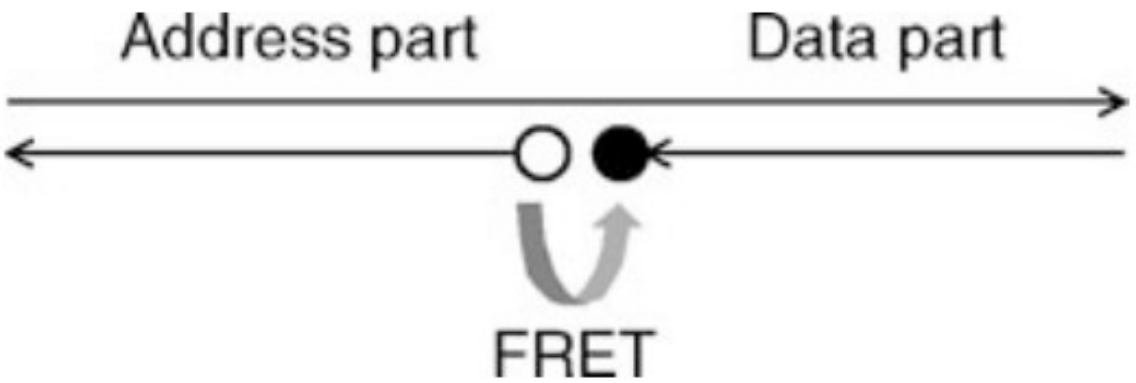
Molecular beacon

- A fluorescent group and a quencher group are attached at the terminals of a small hairpin.
- Since the stem of the hairpin is short, it can hybridize with the strand whose sequence is complementary to the hairpin loop.
- After hybridization, the fluorescent group parts from the quencher and
- fluorescence is observed.



FRET

- The two probes are complementary to the address part and the data part of a memory word.
- If they hybridize with the memory word, they come in close proximity and FRET occurs

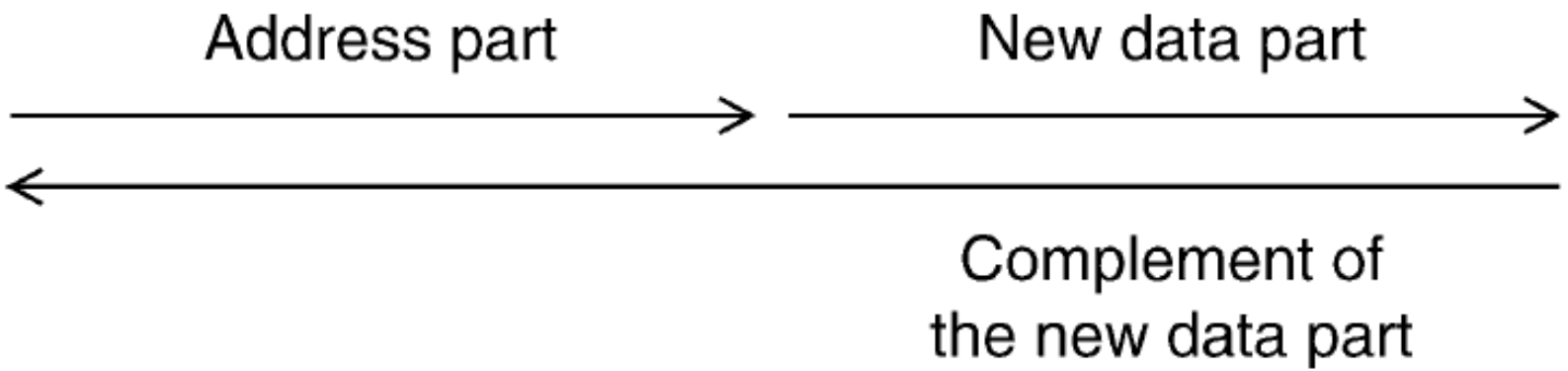


Write Operation

- Case with 0-length of data word:
  - Inserting a memory word – writing 1
  - Deleteng memory word – writing 0
  -
- Otherwise:
  - Access by affinity or PCR separation:
    - Delete old word (difficult for PCR); insert new word
  - Non-hierarchical hybridization

Write operation: hybridization-based memory

- Writing a new data part to a memory word.
- The new data part is put together with the probe complementary to the address part and the new data part.
- After ligation, the new data part is concatenated with the address part.

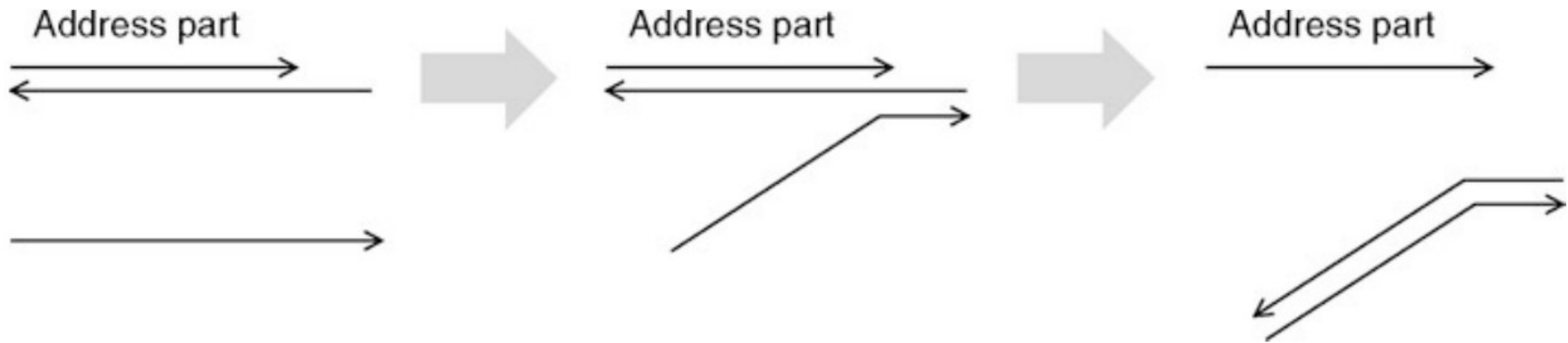


Delete operation: hybridization-based memory

- Deleting the data part of a memory word.
- The probe is complementary to the address part and the adjacent portion of the data part.
- The double strand formed by the word and the probe is cut by a restriction enzyme.

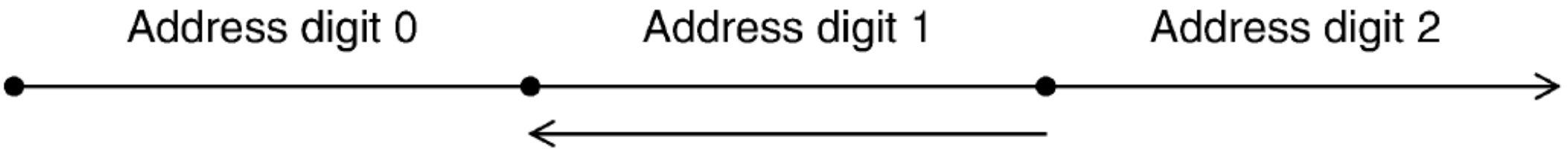
Removing data from a memory word

- The strand complementary to the probe hybridizes with the single-stranded portion of the probe and eventually removes it from the memory word.



Representing address binary digits as single/double strands

- Writing on an address digit.
- A probe hybridizes with the address digit 1.
- This operation is regarded as writing bit 1 to the address digit.

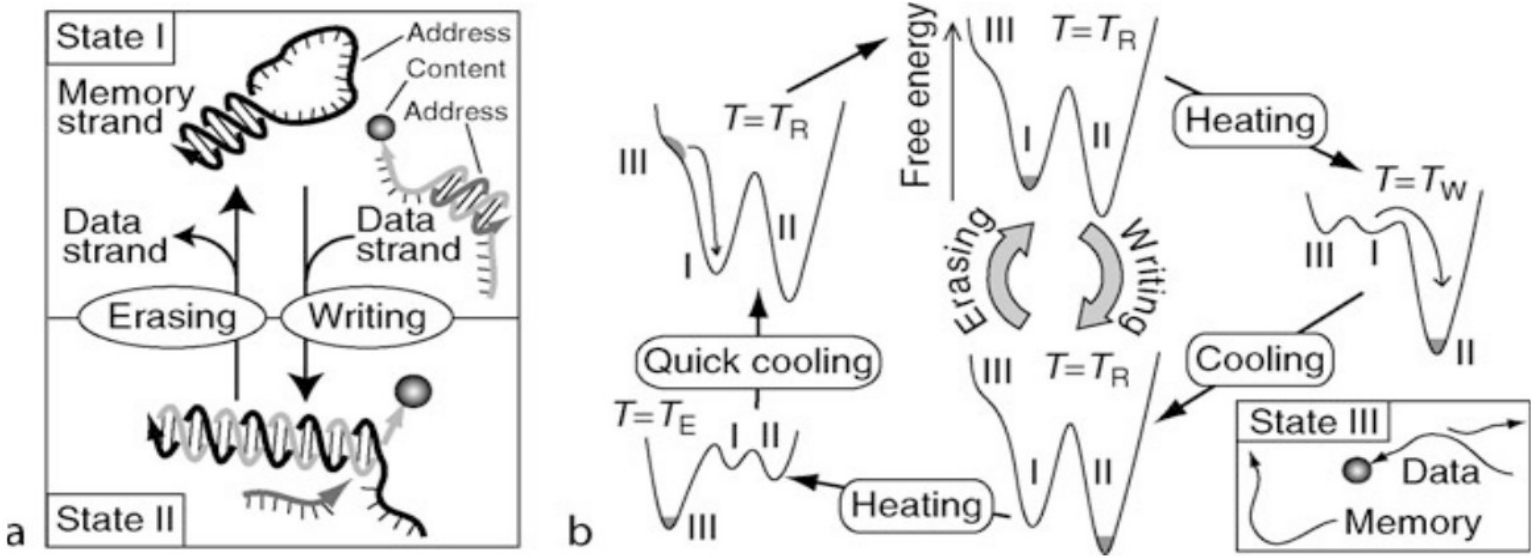


In Vitro DNA Memory

- Solving NP-complete problems:
  - 20-variable SAT by Adleman

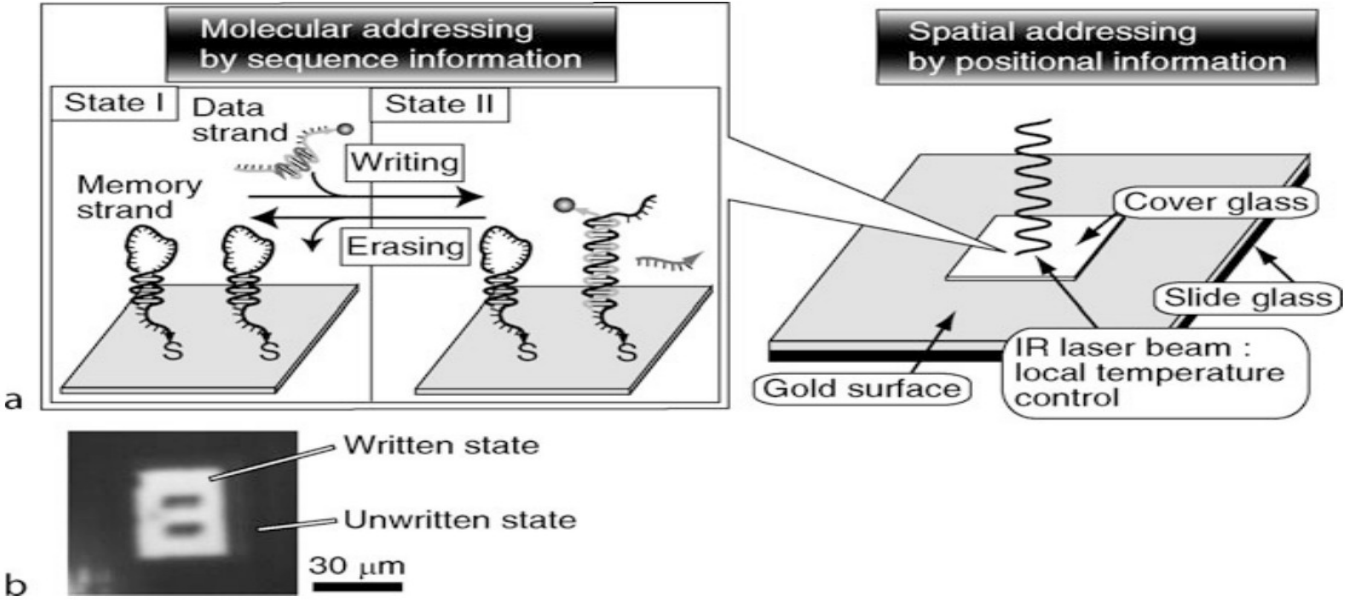
- Nested Primer Molecular Memory (NPMM)
  - Yamamoto et al. constructed a molecular memory consisting of 166 addresses (2008)
  - Each memory word consists of a data part and an address part
  - The data part surrounded by the address part
  - Each digit on the left paired with a digit on the right:
    - CL - BL - AL - Data - AR - BR - CR
    - 16 sequences for each address digit

- Hairpin DNA Memory Dissolved in a Solution
  - (a) Structures of memory and data.
    - State I: unwritten state (hairpin structure).
    - State II: written state (linear structure).
  - (b) Writing and erasing by temperature control.
    - TR, TW, and TE: temperatures of readout, writing, and erasing (TR<TW<TE).
    - State III: TR TW dissociated state.



Hairpin DNA Memory Immobilized on a Surface

- A hairpin DNA memory immobilized on a surface.
  - (a) Hairpin DNA memories are immobilized on a gold surface. Writing and erasing are performed by IR laser beam irradiation.
  - (b) A fluorescence microscope image of the hairpin DNA memory. White area: written state. Black area: unwritten state.



In Vivo DNA Memory

- The perfect design of in vivo memory requires substantial knowledge of cellular and molecular mechanisms.
- Ideally, the memory function should be isolated from the native physiological functions to avoid unexpected interactions among sequences.
- However, it is difficult to meet this requirement.

- The simplest way to encode information into DNA is to apply a code table translating alphabet
- Encoded:
  - “June 6 Invasion: Normandy” (all in capitals) into a short DNA fragment in vitro
  - encoding a song phrase of “It’s a small world” into the genomes of Escherichia coli and Deinococcus radiodurans

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

Encryption into Coding Regions

- A better way to achieve in vivo memories is to override artificial information on sequences whose biological information is well known.
- If we can superimpose a message in protein coding regions in a way that the message does not affect amino acid sequences to be translated,
- then the message can be used as a secret signature, or DNA watermark.

Codon usage chart of a standard eukaryotic genome. "Stop" indicates stop codons. Bold letters are codons that cannot be used for encoding artificial information (see main text)

UUU Phe	UUA Leu	UAU Tyr	UAA Stop	GAA Glu	GAU Asp	GUG Val	GUU Val
UUC Phe	UUG Leu	UAC Tyr	UAG Stop	GAG Glu	GAC Asp	GUA Val	GUC Val
UCU Ser	UCA Ser	UGU Cys	**UGA** **Stop**	GGA Gly	GGU Gly	GCA Ala	GCU Ala
UCC Ser	UCG Ser	UGC Cys	**UGG** **Trp**	GGG Gly	GGC Gly	GCG Ala	GCC Ala
CCC Pro	CCG Pro	CGC Arg	CGG Arg	AGG Arg	AGC Ser	ACG Thr	ACC Thr
CCU Pro	CCA Pro	CGU Arg	CGA Arg	AGA Arg	AGU Ser	ACA Thr	ACU Thr
CUC Leu	CUA Leu	CAC His	CAG Gln	AAG Lys	AAC Asn	**AUA** **Ile**	AUC Ile
CUU Leu	CUG Leu	CAU His	CAA Gln	AAA Lys	AAU Asn	**AUG** **Met**	AUU Ile

Translation from binaries to DNA bases

00 → T	01 → G	10 → C	11 → A