A New Operation Suggested by DNA Biochemistry: Hairpin Lengthening

Victor Mitrana

Faculty of Mathematics and Computer Science
University of Bucharest, Romania
mitrana@fmi.unibuc.ro
and
Department of Information Systems
Polytechnic University of Madrid, Spain
victor.mitrana@upm.es

Contents

Source of inspiration: DNA biochemistry

Hairpin completion

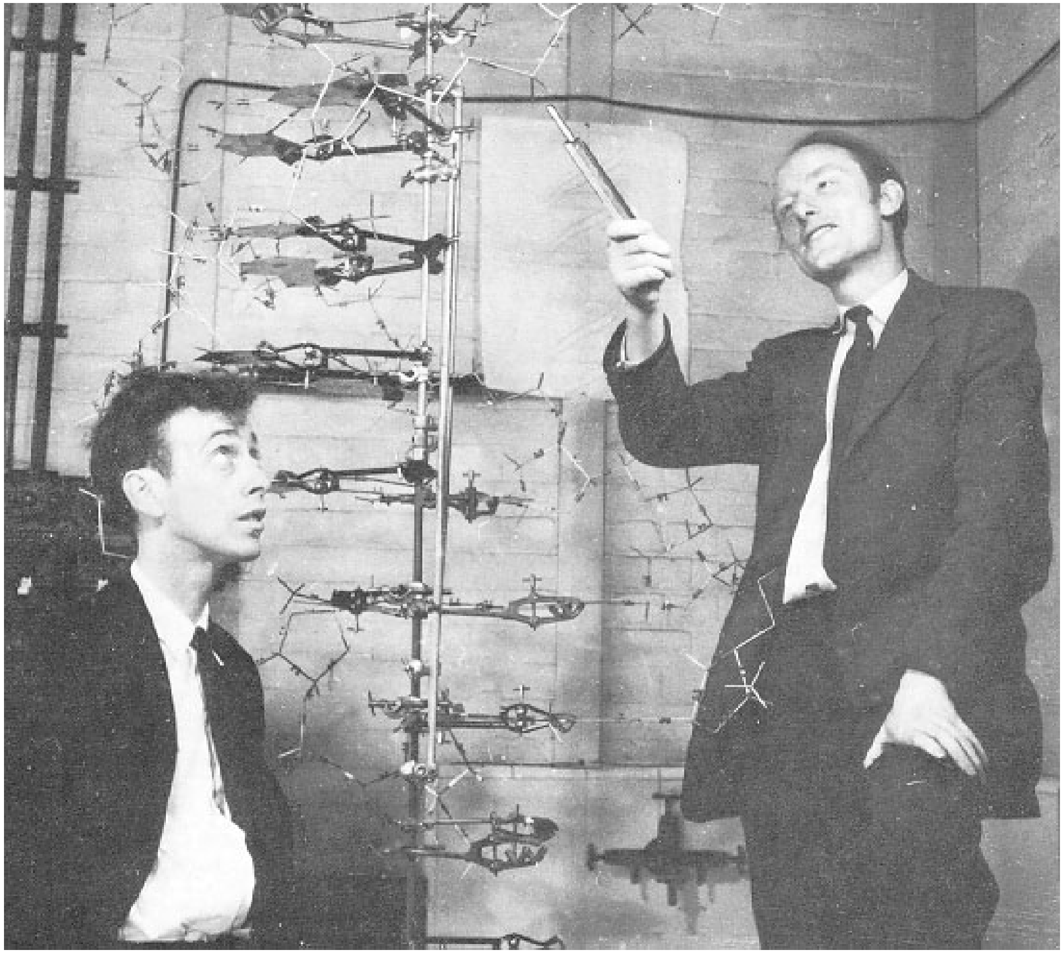
- non-iterated
  - language theoretical properties
  - algorithmic properties
- iterated
  - language theoretical properties
  - algorithmic properties
  - open problems

Variants:

- Bounded hairpin completion
- Hairpin lengthening
- Reductions

DNA (deoxyribonucleic acid)

Watson & Crick (1953): *Nature* 25: 737-738 Molecular Structure of Nucleic Acids: a structure for deoxyribose nucleic acid. Nobel Prize, 1962.



DNA structure (I)

DNA sequences consist of

- A, C, G, T

Nucleotide:

- purine or pyrimidine base
- deoxyribose sugar
- phosphate group

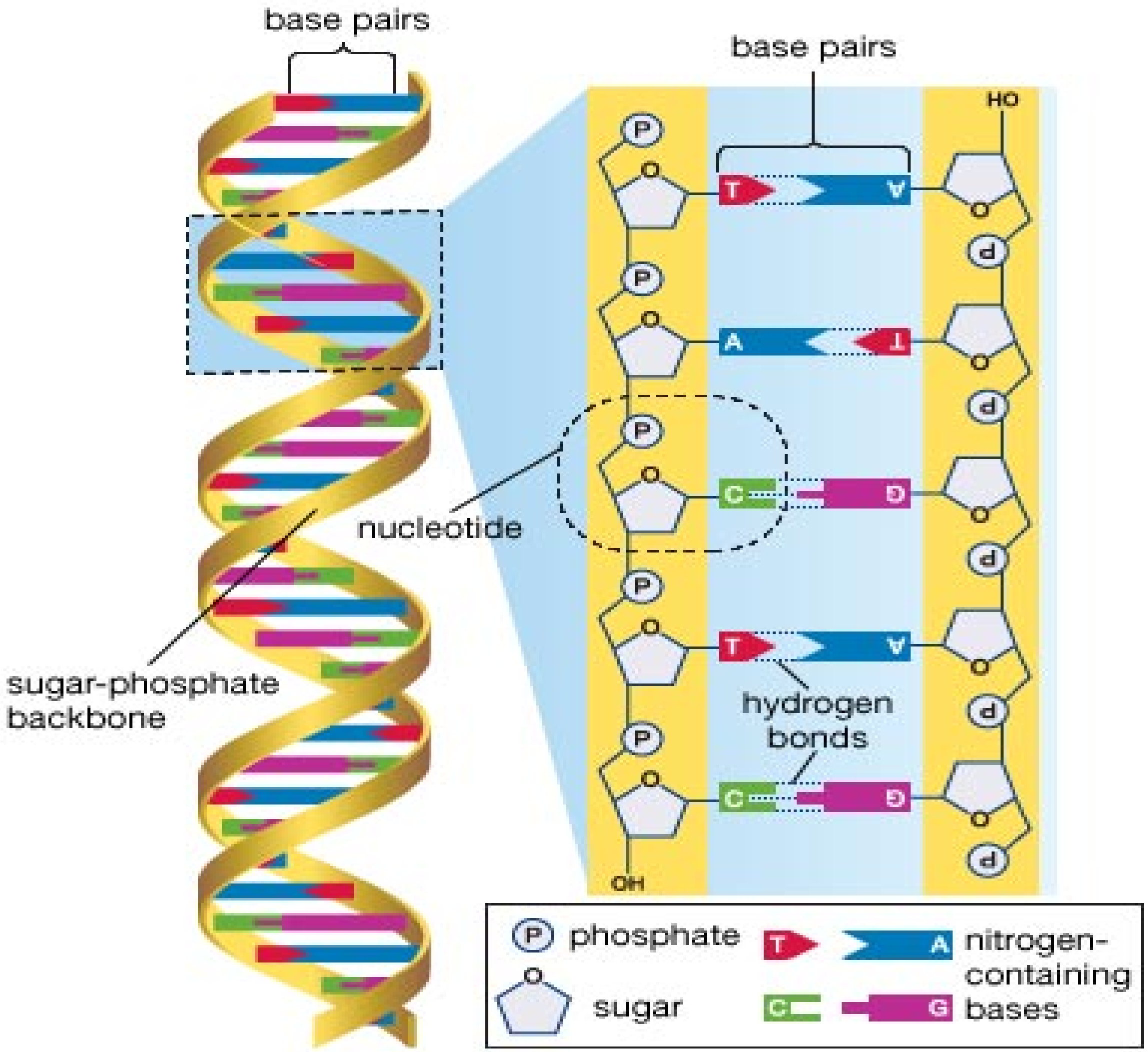
Purine bases

- A(denine), G(uanine)

Pyrimidine bases

- C(ytosine), T(hymine)

DNA structure (II)



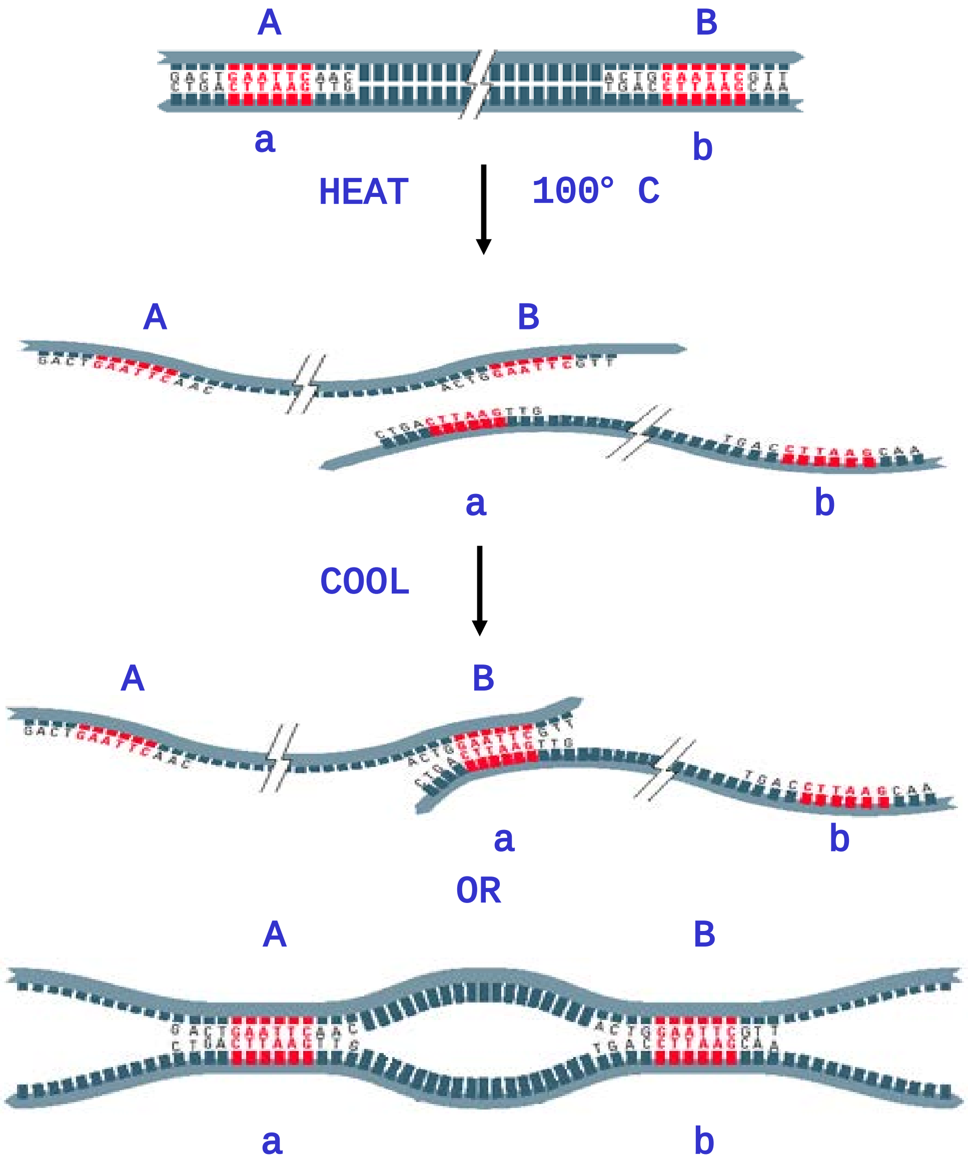
Abstract SS polynucleotide

5′ G→T→A→A→A→G→T→C→C→C→G→T→T→A→G→C 3′

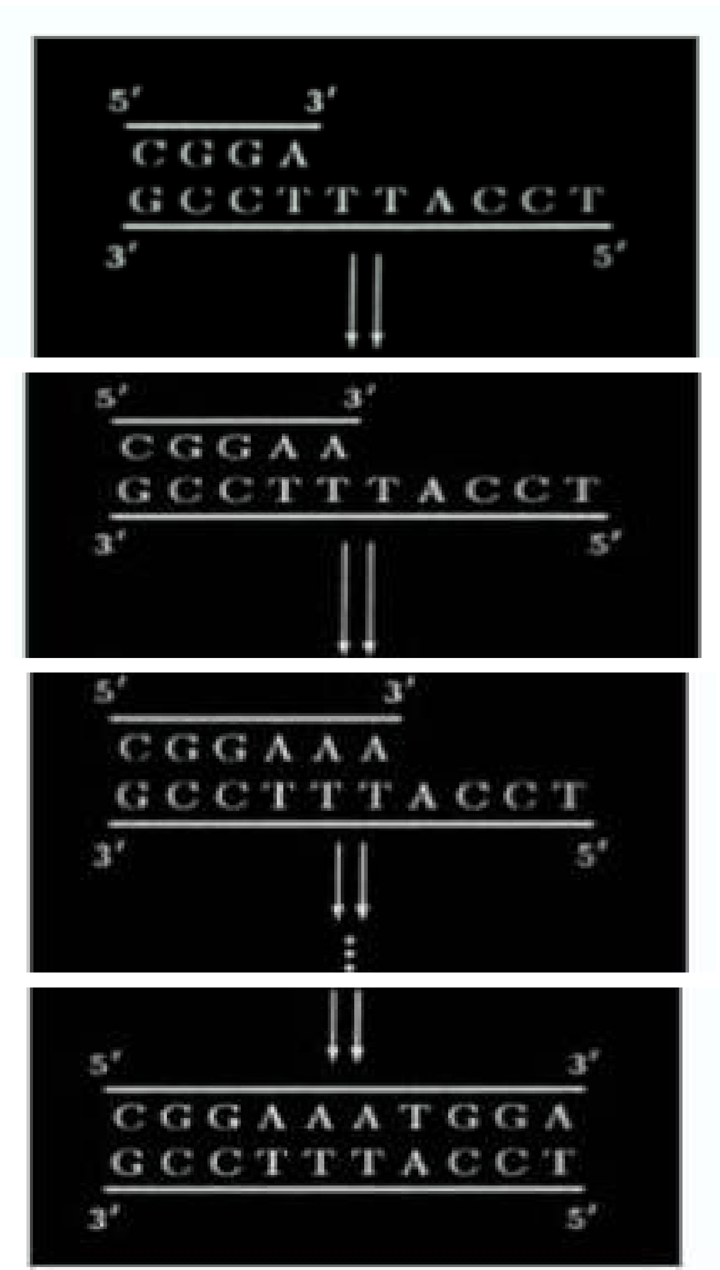
Abstract DS polynucleotide

```
5′ G→T→A→A→A→G→T→C→C→C→G→T→T→A→G→C 3′
   |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
3′ C←A←T←T←T←C←A←G←G←G←C←A←A←T←C←G 5′
```

Melting and annealing



Lengthening DNA

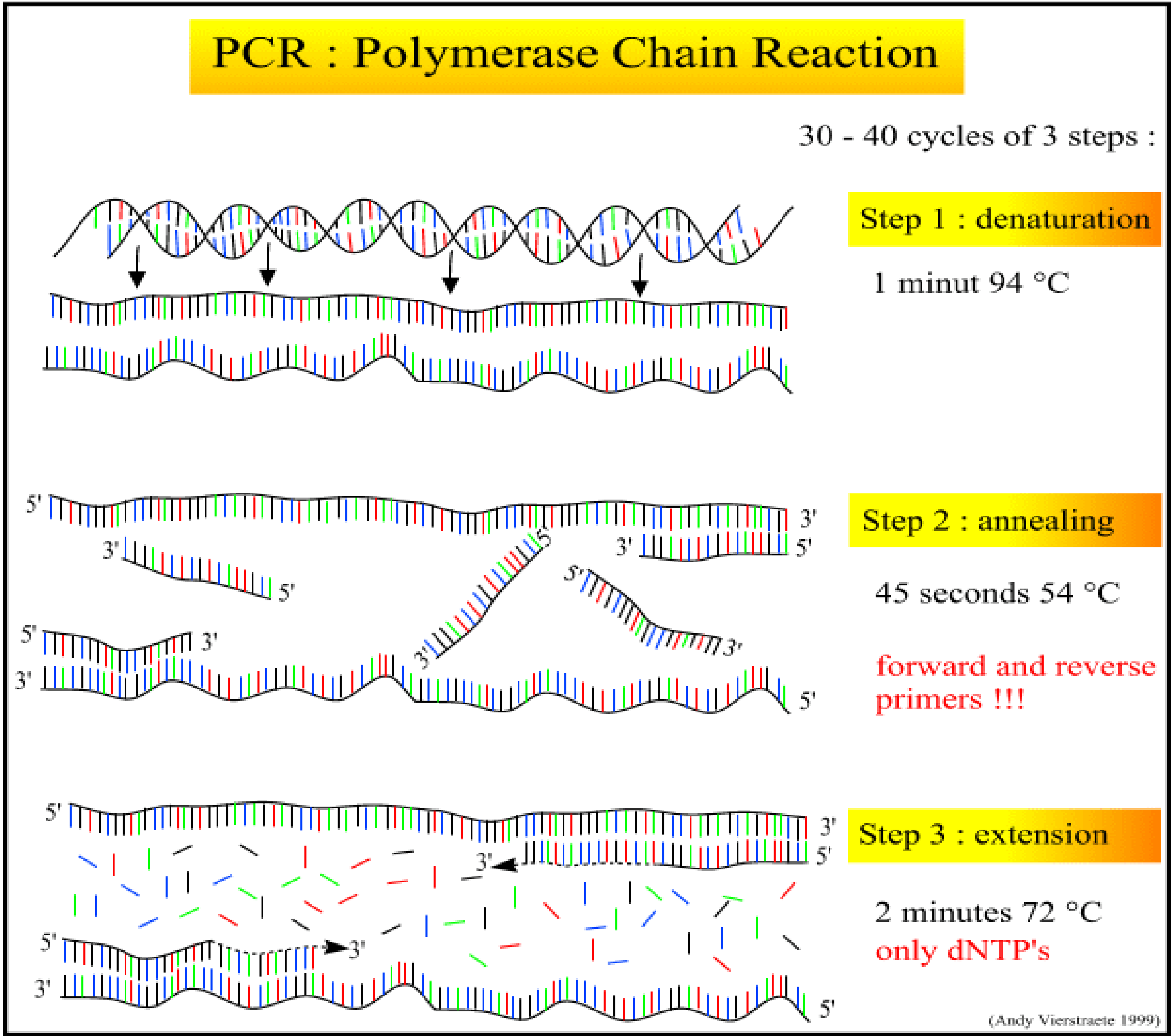


- DNA polymerase enzymes add nucleotides to a DNA molecule

Requirements

- single-stranded template
- primer,
  - bonded to the template
  - 3´-hydroxyl end available for extension

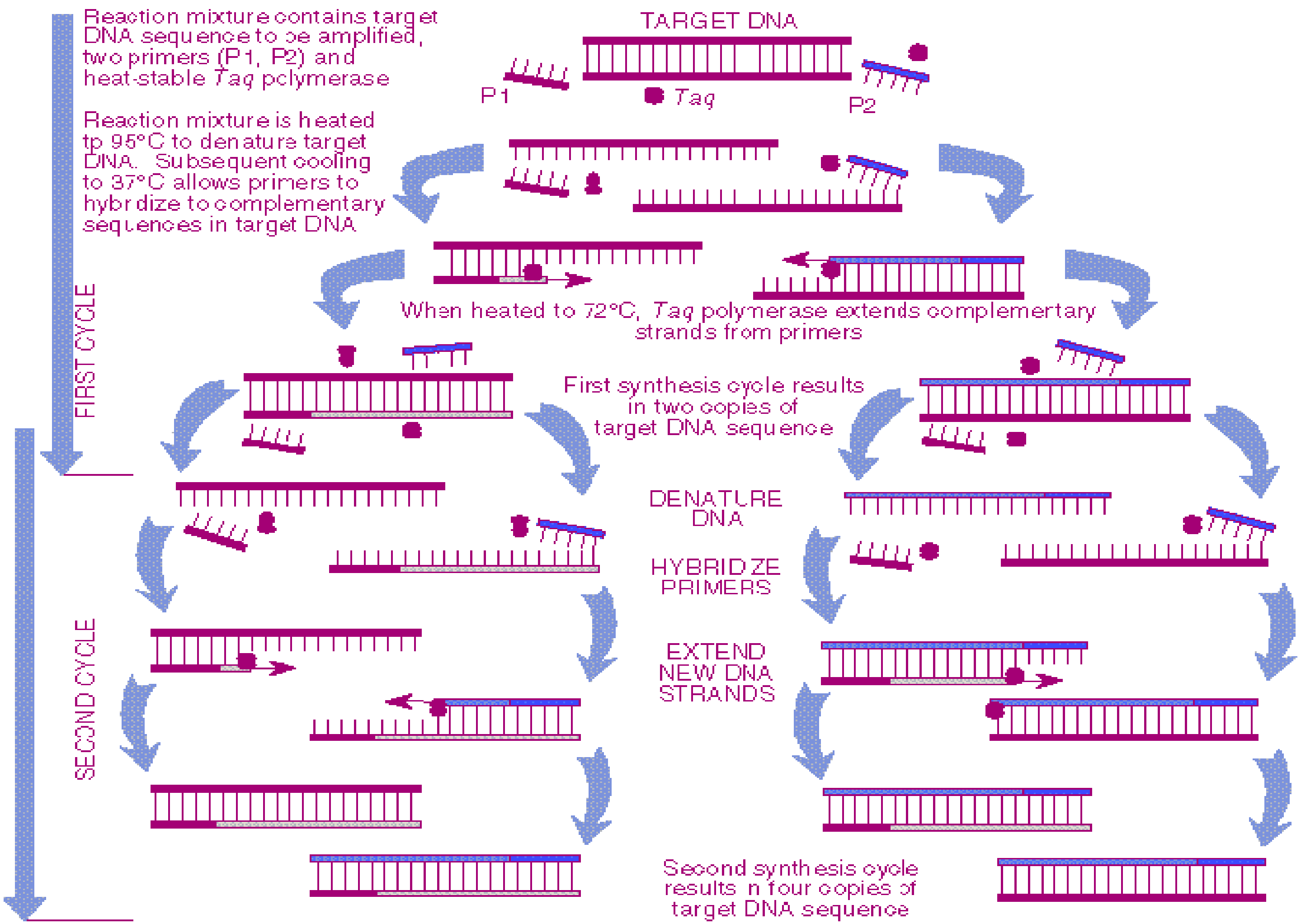
DNA as a computing tool



DNA as a computing tool

ORNL-DWG 91M-17476

DNA Amplification Using Polymerase Chain Reaction



Reaction mixture contains target DNA sequence to be amplified, two primers (P1, P2) and heat-stable *Taq* polymerase

Reaction mixture is heated tp 95°C to denature target DNA. Subsequent cooling to 37°C allows primers to hybridize to complementary sequences in target DNA

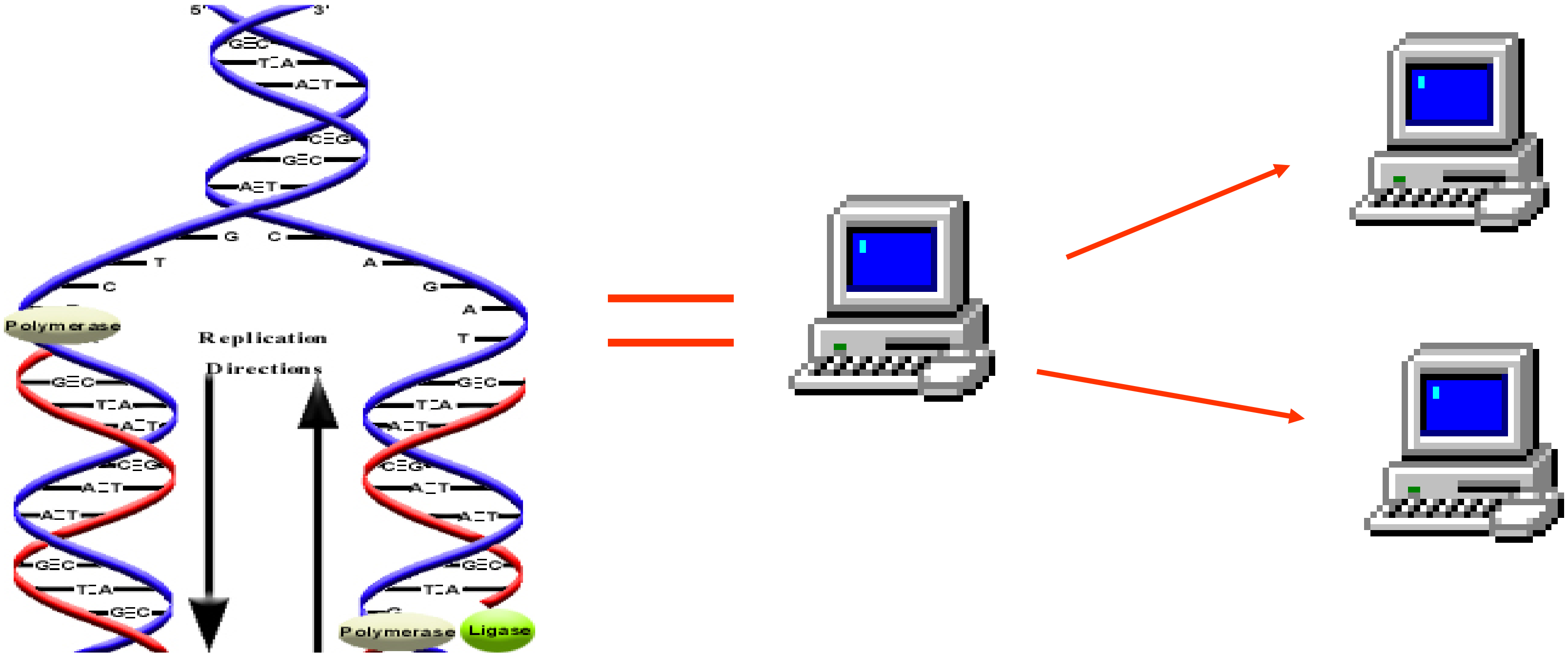
When heated to 72°C, *Taq* polymerase extends complementary strands from primers

First synthesis cycle results in two copies of target DNA sequence

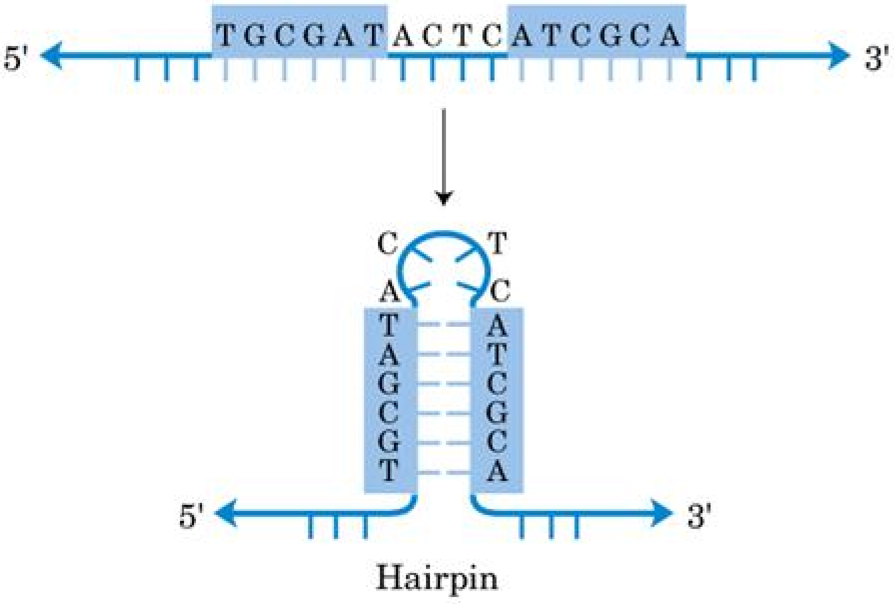
Second synthesis cycle results in four copies of target DNA sequence

Source: *DNA Science*, see Fig. 13.

DNA as a computing tool



Solving problems with hairpin (I)



(a)

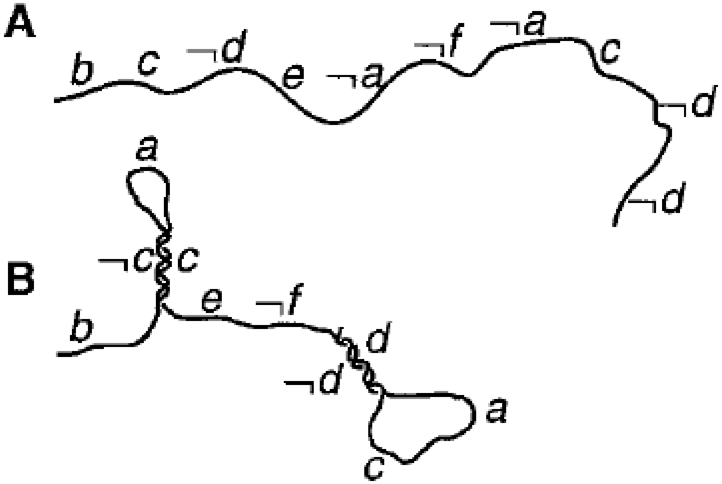
Solving problems with hairpin (II)

The CNF-SAT problem is to find Boolean value assignments that satisfy the given formula in the conjunctive normal form.

$$F=C_1 \wedge C_2 \wedge \ldots C_n$$

$$F=(a \vee b) \wedge (\neg a \vee \neg c) \wedge (\neg b \vee \neg c)$$

b ¬a c, a ¬c, **a ¬a ¬b**



Hairpin completion (I)

A C C T T G C A G A G T C

T G G A A C G T C G T C C T C T A

$$HCS_k\,(w) = \{w\underline{\gamma} \mid w = \gamma\alpha\beta\underline{\alpha},\ |\alpha| = k,\ \alpha,\beta \in V^+,\ \gamma \in V^*\}$$

Hairpin completion (II)

A C C T T G C A G A G T

C

| | | | | | | | |

T G G A A C G T C G C

T

C T

T A

$$HCP_k(w) = \{\underline{\gamma} w \mid w = \alpha\beta\underline{\alpha}\gamma,\ |\alpha| = k,\ \alpha,\beta \in V^+,\ \gamma \in V^*\}$$

Non-iterated hairpin completion (I)

k-hairpin completion

$$\mathbf{HC}_k\,(w) = \mathbf{HCS}_k\,(w) \cup \mathbf{HCP}_k\,(w)$$

hairpin completion

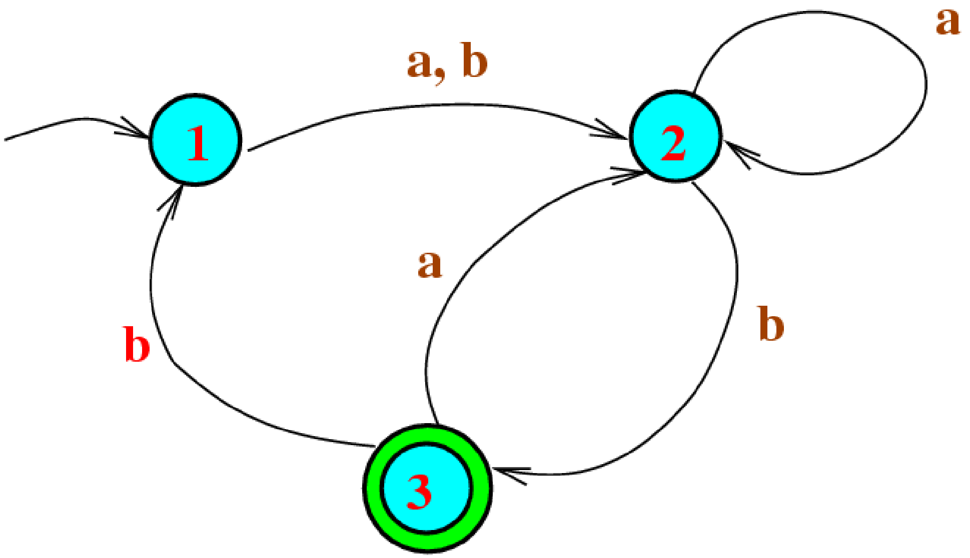
$$\mathbf{HC}\,(w) = \bigcup_{k\geq 1} \mathbf{HC}_k\,(w)$$

$$\mathbf{HC}_k\,(L) = \bigcup_{w\in L} \mathbf{HC}_k\,(w) \qquad \mathbf{HC}\,(L) = \bigcup_{w\in L} \mathbf{HC}\,(w)$$

Non-iterated hairpin completion (II)

Theorem. (Cheptea, Martin-Vide, Mitrana (2006))

For any integer $k \geq 1$, LIN = WCOD(HC$_k$(REG))



	a	b
1	2	2
2	2	3
3	2	1

Corollary. The hairpin completion of a regular language is not necessarily regular but always linear.

$$L_k=\{a^n b^k c \underline{b^k} \mid n \geq 1\}$$

Non-iterated hairpin completion (III)

Given L is it decidable whether or not $HC_k(L)$ is regular?

Theorem. (Diekert, Kopecki, Mitrana (2009))
It is decidable whether or not the hairpin completion of a regular language is still regular.

Remarks:

1. **The problem concern subclasses of linear context-free languages.**
2. **Quite technical proof (approx. 10 pages long)**
3. **A polynomial time algorithm.**
4. **Exponential gap between the size of a DFA for L and a NFA for $HC_k(L)$**

$$L_n = \{bv\underline{a^k}\, ba^k \mid v \in \{a,b\}^n \}$$

Non-iterated hairpin completion (IV)

Theorem. (Diekert, Kopecki, Mitrana (2011))

Let *L* be a regular language accepted by a DFA with *n* states. Then:

1. **The regularity of $\mathrm{HCP}_k(L)$ is decidable in $O(n^2)$ time.**
2. **The regularity of $\mathrm{HC}_k(L)$ is decidable in $O(n^6)$ time.**

Non-iterated hairpin completion (V)

A class of ***mildly context-sensitive of languages*** ***F*****:**

(i) All regular/context-free languages belong to ***F*****.**

(ii) Each language in ***F*** has a constant growth/is semilinear.

(iii) Each language in ***F*** has the membership in **P**.

(iii) ***F*** contains the following three non-context-free languages:

- multiple agreements: $L_1 = \{a^n b^n c^n \mid n \geq 1\}$;
- crossed agreements: $L_2 = \{a^n b^m c^n d^m \mid n,m \geq 1\}$, and
- duplication: $L_3 = \{ww \mid w \in \{a,b\}^+\}$.

Linear indexed grammars, Tree adjoining grammars
Head grammars, Combinatory categorial grammars

Theorem. (Manea, Mitrana, Yokomori (2009))

For any integer *k ≥1*, WCOD(HC_k(LIN)) is a family of mildly context-sensitive languages.

Non-iterated hairpin completion (VI)

A language L over V is called *k-locally testable in the strict sense* (*k-LTSS* for short) if there exists a triple $S_k = (A;B;C)$ such that for any $w \in V^*$ with $|w| \geq k$, $w \in L$ iff $[Pref_k(w) \in A;\ Suff_k(w) \in B;\ Inf_k(w) \in C]$

Proposition. (Manea, Mitrana, Yokomori (2008))

For any given k > 1, REG $\subseteq$ WCOD(HC_k(*k*-LTSS)).

Converse: $L=\{a^n c^k\ b\ \underline{c^k} \mid n \geq 1\}$.

$$HC_k(L) = \{a^n c^k\ b\ \underline{c^k\ a^n} \mid n \geq 1\}.$$

Non-iterated hairpin completion (VII)

A *k-LTSS* language L is *center-disjoint* if there exists a triple $S_k = (A;B;C)$ such that $L = L(S_k)$ *and*

$$((A^{-1} L)B^{-1}) \cap (\underline{A} \cup \underline{B}) = \varnothing.$$

Proposition. (Manea, Mitrana, Yokomori (2008)) **For any $k > 1$ and center-disjoint k-LTSS language L, the morphic image of $\mathrm{HC}_k(L)$ is regular.**

Theorem. **For any $k > 1$, REG is exactly the class of weak-code images of the k-hairpin completion of center-disjoint k-LTSS languages.**

Non-iterated hairpin completion (VIII)

Theorem. (Cheptea, Martin-Vide, Mitrana (2006))
1. NSPACE($f(n)$), where $f(n) \geq \log n$ is a space-constructible function, is closed under hairpin completion.
2. P is closed under hairpin completion.

If L is recognizable in $O(f(n))$ time, then $HC_k(L)$ is recognizable in $O(nf(n))$ time.

Non-iterated hairpin completion (IX)

Pre-processing in $O(f(n))$

$O(1)$

```
$i := 1;$
while $(i + k + 1 \leq n - i - k)$
    if $(w[1..i + k] = \underline{w[n - k - i + 1..n]})$ & $(w[1..n - i] \in L)$
        then $Output : w \in L \searrow_k$ ; halt
        else $i := i + 1$
    endif
endwhile
$Output : w \notin HCS_k(L)$
```

Is the n factor needed?

Non-iterated hairpin completion (X)

Partial answer: (Manea, Martin-Vide, Mitrana (2006))
Hairpin completion of regular languages are recognizable In linear time.

```
Input: $A=(Q,V,\delta, q_0,F)$, $Q=\{0,1,\ldots,p\}$

$m[0]:=0$;
for $t=1$ to $n$
    $m[t]:=\delta(m[t-1],w[t])$;
    $a[t]:=(m[t] \in F)$;
endfor
```

Non-iterated hairpin completion (XI)

Compute a[j]

$O(1)$

```
$i := 1;$
while $(i + k + 1 \leq n - i - k)$
    if $(w[1..i + k] = \underline{w[n - k - i + 1..n]})$ & $(a[n-i])$
        then $Output : w \in L \searrow_k$ ; halt
        else $i := i + 1$
    endif
endwhile
$Output : w \notin HCS_k(L)$
```

Non-iterated hairpin completion (XII)

Partial answer: (Manea, Martin-Vide, Mitrana (2006))
Hairpin completion of context-free languages are recognizable in cubic time.

Input: $G=(N,T,S,P)$ in the Chomsky normal form

Cocke-Younger-Kasami Algorithm

$m[i][j]:=\{A\in N \mid A\Rightarrow^* w[i..j]\};$
$a[t]:=(S\in m[1][t]);$

Iterated hairpin completion (I)

$$\mathrm{HC}_k^{0}(w) = \{w\},$$

$$\mathrm{HC}_k^{n+1}(w) = \mathrm{HC}_k(\mathrm{HC}_k^{n}(w))$$

$$\mathrm{HC}_k^{*}(w) = \bigcup_{n\geq 0} \mathrm{HC}_k^{n}(w)$$

$$\mathrm{HC}_k^{*}(L) = \bigcup_{w\in L} \mathrm{HC}_k^{n}(w)$$

Iterated hairpin completion (II)

Theorem. (Cheptea, Martin-Vide, Mitrana (2006)) **For any $k \geq 1$, the iterated k-hairpin completion of a regular language is not necessarily context-free.**

$$L = \{a^k b \underline{a^k} c^n \underline{a^k} \mid n \geq 1\}$$

$$\mathrm{HC}_k^*(L) \cap \{a^k \underline{c^n} a^k \underline{c^m} a^k b \underline{a^k} c^p \underline{a^k} \mid n,m,\, p \geq 1\} =$$

$$\{a^k \underline{c^n} a^k \underline{c^n} a^k b \underline{a^k} c^n \underline{a^k} \mid n \geq 1\}.$$

Iterated hairpin completion (III)

Theorem. (Cheptea, Martin-Vide, Mitrana (2006))

NSPACE($f(n)$), where $f(n) \geq \log n$ is a space-constructible function, is closed under iterated hairpin completion.

```
Function Membership($x$, $HC_k^*(L)$);
Membership:= false;
if $x \in L$ then Membership:= true; endif; halt;
if ($|x| \leq 2k$) and ($x \notin L$) then halt; endif;
choose nondeterministically a decomposition

    $x = \gamma\alpha\beta\underline{\alpha^R\gamma^R}$, with $|\beta\gamma| \geq 1$ and $|\alpha| = k$;

    if (Membership($\gamma\alpha\beta\underline{\alpha^R}$, $HC_k^*(L)$) or Membership($\underline{\alpha\beta\alpha^R}\gamma^R$, $HC_k^*(L)$))
        then Membership:= true; halt; endif;
```

Iterated hairpin completion (IV)

Theorem. (Manea, Martin-Vide, Mitrana (2006))
If L is recognizable in $O(f(n))$ time, then $HC_k^*(L)$ is recognizable in $O(n^2f(n))$ time.

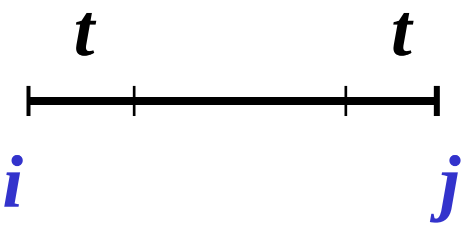
Iterated hairpin completion (V)

```
if $n \le 2k+2$ then if $w \in L$ then Output YES; else Output NO; endif; halt;
for $i=1$ to $n-2k$
  for $j=i+2k$ to $n$
    if $w[i,j] \in L$ then $m[i][j]:=1$; endif;
  endfor;
endfor;
for $l=2k+3$ to $n$
  for $i=1$ to $n-l+1$
    $j:=i+l-1$; $p:=0$;
    for $t=i$ to $i+[(l-1)/2]-1$
      if $w[t]=\overline{w[j-t+i]}$ then $p:=p+1$ else exit; endif;
    endfor;
    if $p \ge k+1$ then for $t=1$ to $p-k$
                          if $(m[i][j-t]=1)$ or $(m[i+t][j]=1)$ then $m[i][j]:=1$; endif;
                        endfor;
    endif;
  endfor;
endfor; if $m[1][n]=1$ then Output YES else Output NO; endif;
```

Iterated hairpin completion (VI)

Can we do it better? **Yes**

1. $P[i][j] = \max(\{t \mid w[i..i+t-1] = \overline{w[j-t+1..j]}\} \cup \{0\}), \; j-i \geq 2k$



```
for $p=2$ to $n-2k+1$
   $i:=p-1; j:=p+2k-1;$
   while $(i \geq 1)$ & $(j \leq n)$
      if $w[i]=\overline{w[j]}$ then $P[i][j]:=P[i+1][j-1]+1$; endif;
      $i:=i-1; j:=j+1;$
   endwhile;
   $i:=p-1; j:=p+2k;$
   while $(i \geq 1)$ & $(j \leq n)$
      if $w[i]=\overline{w[j]}$ then $P[i][j]:=P[i+1][j-1]+1$; endif;
      $i:=i-1; j:=j+1;$
   endwhile;
endfor;
```

Iterated hairpin completion (VII)

2. $right[i]$:= the greatest p, such that $w[i..p] \in (L\rightarrow_k{}^*)$,
$left[j]$:= the least p, such that $w[p..j] \in (L\rightarrow_k{}^*)$.

Initially, $left[j]=i$ and $right[i]=j$, for all i,j such that
$w[i..j] \in L$;

$left[j]=0$ and $right[i]=n+1$, otherwise.

Iterated hairpin completion (VIII)

```
if $n \leq 2k+2$ then if $w \in L$ then Output YES; else Output NO; endif; halt;
for $i=1$ to $n-2k$
   for $j=i+2k$ to $n$
      if $w[i,j] \in L$ then $m[i][j]:=1$; endif;
   endfor;
endfor;
for $l=2k+3$ to $n$
   for $i=1$ to $n-l+1$
      $j:=i+l-1$; $p:=0$;
      for $t=i$ to $i+[(l-1)/2]-1$
         if $w[t]=\overline{w[j-t+i]}$ then $p:=p+1$ else exit; endif;
      endfor;
      if $p \geq k+1$ then for $t=1$ to $p-k$
                              if $(m[i][j-t]=1)$ or $(m[i+t][j]=1)$ then $m[i][j]:=1$; endif;
                           endfor;
      endif;
   endfor;
endfor; if $m[1][n]=1$ then Output YES else Output NO; endif;
```

Iterated hairpin completion (IX)

```
if  $n \le 2k+2$ then if $w \in L$ then Output YES; else Output NO; endif; halt;
for  $i=1$ to $n-2k$
   for  $j=i+2k$ to  $n$
      if  $w[i,j] \in L$ then  $m[i][j]:=1$; endif;
   endfor;
endfor;
Compute P;
for  $l=2k+3$ to $n$
   for  $i=1$ to  $n-l+1$
      $j:=i+l-1$;

      if  $( j>right[i] \ge j - P[i][j] + 1 + k)$ then $m[i][j]:=1$; $left[j]=i$; $right[i]=j$;
      endif;
      if  $( i<left[j] \le i + P[i][j] - 1 - k)$ then $m[i][j]:=1$; $left[j]=i$; $right[i]=j$;
      endif;

endfor;

endfor; if $m[1][n]=1$ then Output YES else Output NO; endif;
```

Iterated hairpin completion (X)

```
if $n \leq 2k+2$ then if $w \in L$ then Output YES; else Output NO; endif; halt;
```

$O(n^3)$ for context-free languages/ ***$O(n^2)$*** for regular languages

```
Compute P;
for $l=2k+3$ to $n$
  for $i=1$ to $n-l+1$
    $j:=i+l-1;$

    if $(j>right[i] \geq j - P[i][j] + 1 + k)$ then $m[i][j]:=1; left[j]=i; right[i]=j;$
    endif;
    if $(i<left[j] \leq i + P[i][j] - 1 - k)$ then $m[i][j]:=1; left[j]=i; right[i]=j;$
    endif;

endfor;
endfor; if $m[1][n]=1$ then Output YES else Output NO; endif;
```

Iterated hairpin completion (XI)

What kind of language is $HC_k^*(w)$?

(i) It is in NL

(ii) It contains non-context-free languages [Kopecki, 2011]

$$w = a^k b a^k \underline{a^k} a^k c \underline{a^k}$$

Theorem. (Manea, Mitrana, Yokomori (2008))

Let k ≥1. The following problems are decidable for this class:

1. The membership problem is decidable in quadratic time.

2. The inclusion is decidable in quadratic time.

3. The equivalence problem is decidable in linear time.

4. The finiteness is decidable in linear time.

Iterated hairpin completion (XII)

Is the regularity of $\mathrm{HC}_k^*(w)$ decidable?

Theorem. (Kari, Kopecki, Seki (2012))
For every non-crossing word w, it is algorithmically decidable whether $HC_k^*(w)$ is regular.

Theorem. (Shikishima-Tsuji (2015))
For every crossing (2,2)-word w, it is algorithmically decidable whether $HC_k^*(w)$ is regular.

Iterated hairpin completion:
Open problems

1. Is the n^2 factor needed ?
   Other classes for which it is not needed ?
2. Is n^2 optimal for the regular case ?
3. Is it decidable whether or not the iterated k-hairpin completion of a given regular language is still regular?
4. Given two words x and y, can one decide whether $HC_k(x) \cap HC_k(y) \neq \emptyset$?
5. Is the regularity of $HC_k^*(w)$ decidable for every word w?

Hairpin completion distance

$HD_k(x,y) = \min\{p \mid y \in (x \rightarrow_k^p) \text{ or } x \in (y \rightarrow_k^p)\}$

$HD_k(L,y) = \min\{p \mid y \in (L \rightarrow_k^p)\}$

1. **Given x,y,k compute $HD_k(x,y)$**
2. **Given L,y,k compute $HD_k(L,y)$**

Solution: dynamic programming

1. $O(\max(n^2 \log n)$, $n = \max(|x|,|y|)$
2. $O(|y|^2 f(|y|))$

Better ?

Bounded hairpin completion

Ito, Leupold, Mitrana (2009),
Ito, Leupold, Manea, Mitrana (2011).

$$p\mathrm{HCS}_k\,(\mathrm{w}) = \{w\underline{\gamma} \mid w = \gamma\alpha\beta\underline{\alpha}\,,\ |\alpha| = k,\ \alpha,\beta \in V^+\,,\ |\gamma| \leq p\}$$

$$p\mathrm{HCP}_k\,(\mathrm{w}) = \{\underline{\gamma}\mathrm{w} \mid w = \alpha\beta\underline{\alpha}\gamma\,,\ |\alpha| = k,\ \alpha,\beta \in V^+\,,\ |\gamma| \leq p\}$$

***p*-bounded *k*-**hairpin completion

$$p\mathrm{HC}_k\,(w) = p\mathrm{HCS}_k\,(w) \cup p\mathrm{HCP}_k\,(w)$$

Hairpin lengthening

Manea, Martin-Vide, Mitrana (2010, 2012)
Manea, Mercas, Mitrana (2012)

$\mathrm{HLS}_k(\mathrm{w}) = \{w\underline{\delta} \mid w = \gamma\alpha\beta\underline{\alpha}, |\alpha| = k, \alpha,\beta \in V^+, \delta$ *is a suffix of* $\gamma\}$

$\mathrm{HLP}_k(\mathrm{w}) = \{\underline{\delta}w \mid w = \alpha\beta\underline{\alpha}\gamma, |\alpha| = k, \alpha,\beta \in V^+, \delta$ *is a prefix of* $\gamma\}$

k-hairpin lengthening

$$\mathrm{HL}_k(w) = \mathrm{HLS}_k(w) \cup \mathrm{HLP}_k(w)$$

Reductions

$$HRS_k\,(w) = \{\gamma\alpha\beta\underline{\alpha} \mid w = \gamma\alpha\beta\underline{\alpha}\,\underline{\gamma},\ |\alpha| = k,\ \alpha,\beta \in V^+,\ \gamma \in V^*\}$$

$$HRP_k\,(w) = \{\alpha\beta\underline{\alpha}\gamma \mid w = \underline{\gamma}\,\alpha\beta\underline{\alpha}\gamma,\ |\alpha| = k,\ \alpha,\beta \in V^+,\ \gamma \in V^*\}$$

$$HR_k(w) = HRS_k(w) \cup HRP_k(w)$$

- formal operation on languages
- distances
- hairpin root of a word/language

M. Ito, P. Leupoldt, F. Manea, C. Martin-Vide, V. Mitrana

Thank You