A FEW GREEDY ALGORITHMS FOR COMPUTING UNIFORM TRANSLOCATION DISTANCE

VICTOR MITRANA

Faculty of Mathematics and Computer Science, University of Bucharest, Romania and Department of Information Systems Polytechnic University of Madrid, Spain

mitrana@fmi.unibuc.ro
Translocation operation in genome

Formal definition

Uniform translocation with unique markers

Uniform translocation with multiple markers: singleton target set

Uniform translocation with multiple markers: multiple target set

Open problems
What Is the Human Genome?

- Human cell
- Nucleus
- Chromosomes
A Sample Human Genome

1  2  3  4  5
6  7  8  9  10
11 12 13 14 15
16 17 18 19 20
21 22 X Y

Chromosome 5

p
Centromere
q
Recombination: Crossing Over

Synapsis

Crossing over

Chromosomes after crossing over
Translocation/Crossover - Formal
(x, y) \vdash_{(i,j)} (z_1, z_2) \text{ iff } x = tu, y = vw, z_1 = tw, z_2 = vu, \text{ and } 
|t| = i, |v| = j.

\vdash_{(i,j)} \text{ is said to be } \textit{uniform} \text{ iff } i=j, \text{ so that we shall simply write } \vdash_i

[U]CO(A) = \bigcup \{ z | (x,y) \vdash_{(i,j)} (z,w) \text{ or } (x,y) \vdash_{(i,j)} (w,z) \} 
\{ x, y \in A \}
The Problem  Translocation distance

Given two genomes $G$ and $G'$ what is the minimal number of translocation mutations that transforms $G$ into $G'$?

1. How the translocation is defined: uniform or arbitrar.

2. How the chromosomes in the two genomes are: they are formed by different segments (markers) or not.

3. How large is the target genome: singleton or arbitrary
Uniform translocation and unique markers  
(J. Kececioglu, R. Ravi, 1995)

Assumptions:
1. All chromosomes (words) in both genomes are of the same length $k$.
2. Each marker (symbol) appears at most once in a chromosome and in only one.
3. If $G$ has $n$ chromosomes, then $G'$ must have $n$ chromosomes as well.

Important note: If a symbol appears on the position $i$ in a word in $G$, then it will appears on the same position in a word of $G'$.

**Theorem 1.** The uniform translocation distance between $G$ and $G'$ can be computed in time and memory $O(kn)$.

Ingredients: Greedy strategy  
Cayley (1849): The minimal number of transpositions for sorting $\pi$ is $n-\Psi(\pi)$. 
1. We label the words in $G'$ in some way from 1 to $n$.
2. Associate with each set $G, G'$ a matrix as follows:
   - each column in the matrix represents a word
   - each symbol from a word is represented by the unique word of $G'$ in which it occurs.

Example: $G = \{a_2a_7a_9a_4, a_5a_1a_2a_8, a_10a_3a_6a_{11}\}$
$G' = \{a_10a_1a_9a_8, a_5a_7a_6a_4, a_2a_3a_{12}a_{11}\}$

$$M_G = \begin{pmatrix}
3 & 2 & 1 \\
2 & 1 & 3 \\
1 & 3 & 2 \\
2 & 1 & 3
\end{pmatrix} \quad M_{G'} = \begin{pmatrix}
1 & 2 & 3 \\
1 & 2 & 3 \\
1 & 2 & 3
\end{pmatrix}$$

**Problem:** Select two columns and a natural $l \leq n - 1$ and interchange the elements of the first $l$ rows.
Let $(i, j, l)$: the columns $i$ and $j$ interchange each other the entries of the first $l$ rows. A solution is a sequence

$$(i_1, j_1, l_1), (i_2, j_2, l_2), \ldots (i_p, j_p, l_p)$$

Find the minimal $p$.

A solution $(i_1, j_1, l_1), (i_2, j_2, l_2), \ldots (i_p, j_p, l_p)$ is “bottom-up if there are no $1 \leq s < q \leq n - 1$ such that $l_q > l_s$. 
**Lemma**: Any instance of the problem has a solution which is bottom-up.
A bottom-up sequence is *locally optimal* if the number of transformations applied to the current row in order to transform it into the identical permutation is minimal.

**Lemma 2** A *bottom-up locally optimal* is *totally optimal*.

*Proof.* Let us consider a part of a bottom-up sequence when one starts to "sort the row $i+1$. Let $\pi$ be the current state of the row $i+1$ and $\lambda_i$ the state of the row. After sorting the row $i+1$ the state of the row $i$ is

$$\lambda_i \circ \pi^{-1}.$$
Uniform translocation distance

\[ \sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 5 & 4 & 3 & 1 \end{pmatrix}; \]

\[ PQ = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 4 & 1 & 3 & 5 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 5 & 4 & 3 & 2 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 5 & 3 & 1 & 4 & 2 \end{pmatrix} \neq QP. \]

\[ \pi = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 4 & 3 & 2 \end{pmatrix} \]
Uniform translocation distance

Given a permutation $\pi$, what is the minimal number $m$ of transpositions $\tau_1, \tau_2, \ldots, \tau_m$ such that

$$\pi \circ \tau_1 \circ \tau_2 \circ \ldots \circ \tau_m = \varepsilon_n$$

**Lemma 3 (Cayley)** The minimal number of transpositions for sorting $\pi$ is $n - \Psi(\pi)$.

**procedure** Sort_Crossover_uniform(A,k,n);

Let $\lambda_1, \lambda_2, \ldots, \lambda_k$ the rows of A

$d := 0; \pi := \varepsilon_n;$

for $i := k$ downto 1 do

$\pi := \lambda_i \circ \pi^{-1};$

$d := d + n - \Psi(\pi);$  

endfor;

end.
Translocation distance: Our solution

Assumptions:
1. All chromosomes (words) in both genomes are of the same length $k$.
2. Each marker (symbol) appears may appear more than once in any chromosome and in different chromosomes.
3. If $G$ has $n$ chromosomes, then $G'$ may have as many chromosomes as we want.

A few more definitions:

A translocation sequence: $S = s_1, s_2, \ldots, s_n$, $s_i = (x_i, y_i) \vdash (k(i), p(i)) (u_i, v_i)$

$P_i(S, x) = \text{card}\{j \leq i | x = x_j \text{ or } x = y_j\} + \text{card}\{j \leq i | x_j = y_j = x\}$,
$F_i(S, x) = \text{card}\{j \leq i | u_j = x_j \text{ or } v_j = y_j\} + \text{card}\{j \leq i | u_j = v_j = x\}$, if $x \not\in A$, $\infty$, otherwise

A translocation sequence $S$ is contiguous iff:
(i) $x_1, y_1 \in A$,
(ii) $F_{i-1}(S, x_i) > P_{i-1}(S, x_i)$, and $F_{i-1}(S, y_i) > P_{i-1}(S, y_i)$,
A CTS $S$ is $B$-producing if $F_n(S, z) > P_n(S, z)$ for all $z \in B$.

$$TD(A,B) = \min\{\lg(S) | S \text{ is a } B \text{ – producing CTS}\}.$$
Example: \( A = \{x_1, x_2, x_3, x_4\} \) with

\( x_1 = abcbad, \ x_2 = bbabbd, \ x_3 = accbabbd, \ x_4 = aaab \)

and

\( z_1 = bbcbad, \ z_2 = ababbd, \ z_3 = ababad, \ z_4 = bbcbd, \ z_5 = abababbd \)

\( z_6 = aabad, \ z_7 = abababbd, \ z_8 = bbd, \ z_9 = bbbd, \ z_{10} = bbabad, \)

\( z_{11} = bbbababd, \ z_{12} = bbababbd, \ z_{13} = bababbd, \ z_{14} = accbd, \ z_{15} = bbccbabd \)

\( z_{16} = aababbd, \ z_{17} = abcccbabbd \)

\( z_{18} = abad \)

A \( B \)-producing CTS, \( B = \{z_4, z_6, z_8, z_{11}, z_{15}, z_{16}, z_{18}\} \).

\[(x_1, x_2) \bowtie (2,2) (z_2, z_1), (z_1, z_2) \bowtie (4,4) (z_4, z_3),\]

\[(z_2, x_2) \bowtie (4,2) (z_7, z_8), (z_3, z_7) \bowtie (2,1) (z_5, z_6), (x_2, x_3) \bowtie (3,3) (z_{12}, z_{14}),\]

\[(z_8, z_{12}) \bowtie (2,5) (z_9, z_{10}), (x_2, x_3) \bowtie (3,3) (z_{12}, z_{14}), (x_2, x_3) \bowtie (3,3) (z_{12}, z_{14}),\]

\[(z_{12}, z_{10}) \bowtie (2,1) (z_{11}, z_{13}), (z_{12}, x_3) \bowtie (2,1) (z_{15}, z_{16}), (x_1, x_3) \bowtie (3,1) (z_{17}, z_{18}).\]
Translocation distance: Our solution

Example: \( A = \{x_1, x_2, x_3, x_4\} \) with
\[ x_1 = \text{abcbad}, \ x_2 = \text{bbabbd}, \ x_3 = \text{acccbabd}, \ x_4 = \text{aaab}, \]
and
\[ z_1 = \text{bbcbad}, \ z_2 = \text{ababd}, \ z_3 = \text{ababad}, \ z_4 = \text{bbbcbd}, \ z_5 = \text{abababbd} \]
\[ z_6 = \text{aabad}, \ z_7 = \text{abababbd}, \ z_8 = \text{bbd}, \ z_9 = \text{bbbd}, \ z_{10} = \text{bbababd}, \]
\[ z_{11} = \text{bbbababd}, \ z_{12} = \text{bbababbd}, \ z_{13} = \text{bababbd}, \ z_{14} = \text{acccbd}, \ z_{15} = \text{bbccbabd} \]
\[ z_{16} = \text{aababd}, \ z_{17} = \text{abcccbabbd}, \ z_{18} = \text{abad} \]

A \( B \)-producing CTS, \( B = \{z_4, z_6, z_8, z_{11}, z_{15}, z_{16}, z_{18}\} \).

\[
(x_1, x_2) \approx_{(2,2)} (z_2, z_1), \ (z_1, z_2) \approx_{(4,4)} (z_4, z_3), \ (x_1, x_2) \approx_{(2,2)} (z_2, z_1),

(x_2, x_2) \approx_{(4,2)} (z_7, z_8), \ (z_3, z_7) \approx_{(2,1)} (z_5, z_6), \ (x_2, x_3) \approx_{(3,3)} (z_{12}, z_{14}),

(z_8, z_{12}) \approx_{(2,5)} (z_9, z_{10}), \ (x_2, x_3) \approx_{(3,3)} (z_{12}, z_{14}), \ (x_2, x_3) \approx_{(3,3)} (z_{12}, z_{14}),

(z_{12}, z_{10}) \approx_{(2,1)} (z_{11}, z_{13}), \ (z_{12}, x_3) \approx_{(2,1)} (z_{15}, z_{16}), \ (x_1, x_3) \approx_{(3,1)} (z_{17}, z_{18}).
\]

\[ TD(A,B) \leq 12 \]
Translocation distance: Our solution

Compute $TD(A,B)$

**B is a singleton:** Let $z$ be a string of length $k$ and $A$ be a set of cardinality $n$. There is an exact algorithm that computes $TD(A,z)$ in $O(kn)$ time and $O(kn)$ space.

**B is an arbitrary set:** There is a 2-approximation algorithm for computing the translocation distance from two sets of strings.
Let $A = \{x_1, x_2, \ldots, x_n\}$ and $z$ be an arbitrary string of length $k$

$$\text{MaxSubLen}(A, z, p) = \max\{q \mid \exists 1 \leq i \leq n \text{ such that } x_i[p, p + q - 1] = z[p, p + q - 1]\}.$$

Let $z \in TO_*(A)$; define iteratively the set $H(A, z)$ of intervals of natural numbers as follows:

1. $H(A, z) = \{[1, \text{MaxSubLen}(A, z, 1)]\}$;
2. Take the interval $[i, j]$ having the largest $j$; if $j = k$, then stop, otherwise put into $H(A, z)$ the new interval $[j + 1, j + \text{MaxSubLen}(A, z, j + 1)]$.

Note that we allow intervals of the form $[i, i]$ for some $i$ to be in $H(A, z)$; moreover, for each $1 \leq i \leq k$ there are $1 \leq p \leq q \leq k$ (possibly the same) such that $i \in [p, q] \in H(A, z)$.

**Lemma 4** Let $S$ be a $z$-producing CTS in $CO_*(A)$. Then,

$$\lg(S) \geq \text{card}(H(A, z)) - 1.$$
Translocation distance: Our solution

\[ s_i = (x_i, y_i) \vdash_{p_i} (u_i, v_i) \]

\[ A' = \{ x[\text{MaxSubLen}(A, z, 1) + 1, k] | x \in A \}, \]

\[ z' = z[\text{MaxSubLen}(A, z, 1) + 1, k]. \]

For simplicity denote \( r = \text{MaxSubLen}(A, z, 1) \). Clearly, \( H(A', z') = \{ [i-r, j-r] | [i, j] \in H(A, z) \backslash \{ [1, r] \} \} \), hence \( \text{card}(H(A', z')) = \text{card}(H(A, z)) - 1 \). Starting from \( S \) we construct a CTS in \( CO_*(A') \), producing \( z' \) \( S' = s'_1, s'_2, \ldots s'_m \) in the way indicated by the following procedure:
Procedure Construct_CTS(S,r);
begin
  \( m := 0; \)
  for \( i := 1 \) to \( q \) begin
    if \( (p_i > r) \) then
      \( m := m + 1; \ s'_m = (x_{i[r+1,k]}, y_{i[r+1,k]}) \leftarrow p_{i-r} (u_{i[r+1,k]}, v_{i[r+1,k]}); \)
    endif;
  endfor;
end.

Claim 1: \( S' \) is a CTS.

Claim 2: \( S' \) is \( z' \)-producing.
Translocation distance: Our solution

\[ p_{i_1}, p_{i_2}, \ldots, p_{i_m} \] are all integers from \( \{p_1, p_2, \ldots, p_q\} \) bigger than \( r \)

\[
F_{j-1}(S', x_{i_j}[r+1,k]) = \sum_{x[r+1,k] = x_{i_j}[r+1,k]} F_{i_j-1}(S, x) - \text{card}(X) - \text{card}(Y),
\]

\[
P_{j-1}(S', x_{i_j}[r+1,k]) = \sum_{x[r+1,k] = x_{i_j}[r+1,k]} P_{i_j-1}(S, x) - \text{card}(X) - \text{card}(Y),
\]

where

\[ X = \{ t \leq i_j - 1 | p_t \leq r, \ u_t[r+1,k] = v_t[r+1,k] = x_{i_j}[r+1,k] \}, \]

\[ Y = \{ t \leq i_j - 1 | p_t \leq r, \ u_t[r+1,k] = x_{i_j}[r+1,k] \text{ or } v_t[r+1,k] = x_{i_j}[r+1,k] \}. \]
Theorem 2 Let $z$ be a string of length $k$ and $A$ be a set of cardinality $n$. There is an exact algorithm that computes $CD(A, z)$ in $O(kn)$ time and $O(kn)$ space.
**Translocation distance: Our solution**

**Arbitrary Target Sets**

Let $A$ be a finite set of strings and $z \in CO_*(A)$; denote by

$$\begin{align*}
\text{MaxPrefLen}(A, z) &= \begin{cases} 
|z|, \text{ iff } z \in A, \\
\max(\{|q| < |z|, \text{ there exists } x \in A, |x| > q, \\
\text{so that } x[1, q] = z[1, q]\} \cup \{0\}),
\end{cases} \\
\text{MaxSufLen}(A, z) &= \max(\{|q| \text{ there exists } x \in A, |x| \geq |z|, \\
\text{so that } x[|x| - q + 1, |x|] = z[|z| - q + 1, |z|]\} \\
\cup \{0\}),
\end{align*}$$

ArbMaxSubLen(A, z, p) = \max(\{|q| \text{ there exists } x \in A \text{ and } |x| \geq p + q \\
\text{such that } x[p, p + q - 1] = z[p, p + q - 1]\} \\
\cup \{0\}).
We define iteratively the set $ArbH(A, z)$ of intervals of natural numbers as follows, provided that all parameters defined above are nonzero:

1. $ArbH(A, z) = \{[1, MaxPrefLen(A, z)]\}$;

2. Take the interval $[i, j]$ having the largest $j$; if $j = |z|$, then stop. If $j < |z| - MaxSufLen(A, z)$, then put the new interval $[j + 1, j + ArbMaxSubLen(A, z, j + 1)]$ into $ArbH(A, z)$; otherwise put $[j + 1, |z|]$ into $ArbH(A, z)$. 
Theorem 3.1. Let $A$ be a finite set of strings and $B$ be a finite subset of $TO_*(A)$. Then \[
\sum_{z \in B} \frac{\text{card}(\text{ArbH}(A,z)) - 1}{2} \leq TD(A, B) \leq \sum_{z \in B} \text{card}(\text{ArbH}(A,z)) - 1.\]

2. There exist $A$ and $B \subseteq TO_*(A)$ such that $TD(A, B) = \frac{\sum_{z \in B} (\text{card}(\text{ArbH}(A,z)) - 1)}{2}$.

3. There exist $A$ and $B \subseteq TO_*(A)$ such that $TD(A, B) = \sum_{z \in B} (\text{card}(\text{ArbH}(A,z)) - 1)$. 
Proof. 1. We shall prove the first assertion by induction on the length of the longest string in $B$, say $k$. The non-trivial relation is

$$\sum_{z \in B}(\text{card}(\text{ArbH}(A,z))-1) \leq TD(A, B).$$

\[ (*) \]

If $k = 1$, the relation (*) is satisfied. Assume that the relation (*) holds for any two finite sets $X$ and $Y$, $Y \subseteq TO_*(X)$, all strings in $Y$ being shorter than $k$. Assume that $B \setminus A = \{z_1, z_2, \ldots, z_m\}$ and let $S = s_1, s_2, \ldots, s_q$, $s_i = (x_i, y_i) \vdash p_i (u_i, v_i)$, $1 \leq i \leq q$, be a $B \setminus A$-producing $CTS$ in $TO_*(A)$. Note that at least one string in $B \setminus A$ should exist, otherwise the relation (*) being trivially fulfilled.
Consider \( m \) new symbols \( a_1, a_2, \ldots, a_m \) and construct the sets:

\[
A' = \{ x[1, r]a_i x[r + 2, |x|] | x \in A, 1 \leq i \leq m \}, \quad B' = \{ z_i[1, r]a_i z_i[r + 2, |z_i|] | 1 \leq i \leq m \},
\]

where \( r = \min\{p_i | 1 \leq i \leq q\} \). One can construct a \( B' \)-producing \( CTS \) in \( TO_*(A') \) of the same length of \( S \), say \( S' \) by applying a procedure \( \text{Convert} \) illustrated by the next example.
Translocation distance: Our solution

\[ B = \{abacdb, aabcccb, bbaadc\}, A = \{abbcab, aaaadcb, bbccdc\}. \]

The CTS \( S \) is

\[
(\text{abbcab}, \text{aaaadcb}) \vdash_2 (\text{abaadcb}, \text{aabcccb}), (\text{abbcab}, \text{abaadcb}) \vdash_3 (\text{bbadcb}, \text{abcccb}), \\
(\text{bbccdc}, \text{abcccb}) \vdash_2 (\text{bbccdb}, \text{bbccdc}), (\text{bbcccb}, \text{aaaadcb}) \vdash_3 (\text{bbadcb}, \text{aaaadcb}), \\
(\text{bbadcb}, \text{bbccdc}) \vdash_5 (\text{bbadcb}, \text{bbccdc}), (\text{bbadcb}, \text{aaaadcb}) \vdash_2 (\text{abcccb}, \text{aaaadcb}), \\
(\text{abcccb}, \text{aaaadcb}) \vdash_4 (\text{abacdb}, \text{aaaacdb}).
\]

The procedure \( \text{Convert} \) runs for \( r = 2 \) transforming this sequence into the sequence \( S' \):

\[
(\text{aba}_2\text{cccb}, \text{aaa}_3\text{adcb}) \vdash_2 (\text{aba}_3\text{adcb}, \text{aaa}_2\text{cccb}), (\text{aba}_1\text{cccb}, \text{aba}_3\text{adcb}) \vdash_3 \\
(\text{aba}_1\text{adcb}, \text{aba}_3\text{cccb}), (\text{bbca}_1\text{ccdb}, \text{aba}_3\text{cccb}) \vdash_2 (\text{bbca}_3\text{cccb}, \text{aba}_1\text{ccdb}), \\
(\text{bbca}_3\text{cccb}, \text{aaa}_1\text{adcb}) \vdash_3 (\text{bbca}_3\text{adcb}, \text{aaa}_1\text{ccdb}), (\text{bbca}_3\text{adcb}, \text{bbca}_1\text{ccdb}) \vdash_5 \\
(\text{bbca}_3\text{adcb}, \text{bbca}_1\text{ccdb}), (\text{aba}_1\text{adcb}, \text{aaa}_1\text{ccdb}) \vdash_2 (\text{aba}_1\text{ccdb}, \text{aaa}_1\text{adcb}), \\
(\text{aba}_1\text{ccdb}, \text{aaa}_1\text{adcb}) \vdash_4 (\text{aba}_1\text{ccdb}, \text{aaa}_1\text{acdb}).
\]
Now $S'$ is transformed into $S''$ for $r$ previously defined. $S''$ is a $B''$-producing CTS in $CO_*(A'')$, where

$$A'' = \{ a_i x [r + 2, \| x \|] | x \in A, 1 \leq i \leq m \}, \quad B'' = \{ a_i z_i [r + 2, \| z_i \|] | 1 \leq i \leq m \}$$

For each $1 \leq i \leq m \\text{card}(ArbH(A'', a_i z_i [r + 2, \| z_i \|]))$ is either $\text{card}(ArbH(A, z_i))$ or $\text{card}(ArbH(A, z_i)) - 1$. 
Translocation distance: Our solution

\[
card(\text{ArbH}(A'', a_i z_i[r + 2, |z_i|])) = card(\text{ArbH}(A, z_i)) - 1
\]

there exist at least one step in \(S'\) where the strings exchange prefixes of length at most \(r\). It follows that \(lg(S''') \leq lg(S') - \lceil t/2 \rceil\), where \(t = card(\{i|card(\text{ArbH}(A'', a_i z_i[r + 2, |z_i|])) = card(\text{ArbH}(A, z_i)) - 1\})\).

Consequently,

\[
lg(S) = lg(S') \geq lg(S''') + \lceil t/2 \rceil \geq \sum_1^m (\text{card}(\text{ArbH}(A'', a_i z_i[r + 2, |z_i|])) - 1) + \frac{2}{2} \geq \sum_1^m (\text{Arbcard}(H(A, z_i)) - 1) \frac{2}{2}.
\]
Theorem 4: There is a 2-approximation algorithm for computing the translocation distance from two sets of strings.
1. Is it possible to do it better?

2. Non-uniform translocation?

(i) Non-uniform translocation and unique markers:

2-approximation algorithm

(ii) This definition of translocation distance:
Thank You

READY FOR DISCUSSIONS